Die Verwendung von Mischverteilungsmodellen zur Optimierung wiederholter Erhebungen in Patientenorientierter Versorgungsforschung und psychotherapeutischer Praxis

Inauguraldissertation zur Erlangung der Doktorwürde (Dr. rer. nat.)
im Fach Psychologie, Fachbereich I an der Universität Trier

Vorgelegt von:
Jan Rasmus Böhnke

Gutachter:
Prof. Dr. Wolfgang Lutz
Prof. Dr. em. Karl F. Wender

Dezember, 2012

Danksagung

An der Universität Trier möchte ich meinem Doktorvater Prof. Dr. Wolfgang Lutz für Begleitung, Zuspruch und das in mich gesetzte Vertrauen danken. Ein weiterer Dank gilt Prof. Dr. em. Karl F. Wender für die Übernahme der Zweitbegutachtung, die gute kollegiale Zusammenarbeit in der Methodenausbildung und vor allem für viel Geduld. Auch möchte ich mich bei allen Kolleginnen und Kollegen aus der Arbeitsgruppe bedanken und besonders bei Dr. Katharina Köck für jahrelange gute Zusammenarbeit und gemeinsam durchgestandene Höhen und Tiefen.

A special shout-out goes to the team and the visitors of the Psychometrics Centre at the University of Cambridge, 2010 and 2011. Thank you all for being such an inspiration. Anna, Jan, Tim: You know what I mean when I say we should meet again at The Elm Tree for a Dragon Slayer!

Meinen Eltern danke ich für den Anfang: Was immer Ihr Euch dabei dachtet, als Ihr mir das erste Mal vorgelesen habt, es findet hiermit ein Stück Vollendung.

Ganz besonders möchte ich meiner Frau Robina für Ihre Unterstützung danken.
Ich habe nicht mehr losgelassen.

Inhaltsverzeichnis

0. Zusammenfassung

Psychotherapeutische Maßnahmen wirken im Mittel, doch ist unklar, ob eine Therapie bei einem konkreten Patienten auch ihre (maximale) Wirkung zeigt. Befunde der Patientenorientierten Versorgungsforschung zur Wirksamkeit von Feedback zeigen, dass eine Verbesserung des Therapieergebnisses durch Qualitätssicherungsmaßnahmen wie z.B. kontinuierliches Monitoring möglich ist. Diese Forschung und ihre praktische Anwendung machen es nötig, Daten am Einzelfall wiederholt zu erheben. Damit wird es unumgänglich, die Messungen effizienter zu gestalten. Diese Arbeit widmet sich der Frage, wie Mischverteilungsmodelle (Item Response/ Rasch-Modell sowie Latent Profile Analysis) dazu genutzt werden können, Fragebögen (weiter-) zu entwickeln, die mit kürzerem Umfang für diese Zwecke besser eingesetzt werden können.

Gegen die Verwendung von Mischverteilungsmodellen sprach lange, dass spezielle Software und Training erforderlich waren und dies im Praxiskontext nicht machbar war. Mit R (R Development Core Team, 2010) steht eine freie Softwareumgebung ("open source"; Culpepper & Aguinis, 2010) zur Verfügung, die die Schätzung einer ganzen Fülle von Modellen möglich macht, auch von Mischverteilungsmodellen. Da Qualitätssicherung bei frei verfügbarer Software nötig ist, widmet sich Studie I der Frage, ob drei zentrale Pakete zur Schätzung von Rasch-Modellen in der R-Umgebung (eRm, ltm, mixRasch; Details siehe unten) zu akzeptablen Schätzergebnissen führen, d.h. zur Nutzung empfohlen werden können. Hierzu wurden in einer Simulationsstudie die Itemzahl, Stichprobengröße und Verteilung der Stichprobe systematisch variiert und der Effekt auf die Schätzgenauigkeit geprüft. Es zeigte sich, dass alle drei Schätzalgorithmen unter den realisierten Bedingungen zu zufriedenstellenden Genauigkeiten kommen und die Verteilungsform unter den gewählten Bedingungen keinen Einfluss auf die Genauigkeit hatte.

Studie II nutzte das Rasch-Modell um für ein Maß psychischer Belastung (Fragebogen zur Evaluation von Psychotherapieverläufen, FEP; Lutz, Schürch, et al., 2009) Kurzformen für spezifische Erhebungszwecke zu entwickeln: (1) verkürzte Erhebung beim Screening und (2) verkürzte Erfassung im hohen Belastungsbereich. Die Kurzformen wurden mittels Bootstrap und Kreuzvalidierung dahingehend geprüft, ob sie replizierbar eine bessere Messqualität erbrachten als andere

Itemauswahlen aus dem Fragebogen, was sich bestätigte. Durch die Verwendung des Rasch-Modells sind die so erstellten Kurzformen miteinander und auch mit der Vollversion vergleichbar. Dies macht auch ohne die Verwendung spezieller Software (teil-)adaptives Testen möglich.

Studie III untersuchte wie drei Methoden genutzt werden können um festzustellen, welche Items eines Tests sich über den Verlauf einer Therapie als veränderungssensitiv erweisen. Hierzu wurden mittels einer Bevölkerungsstichprobe und den Prä- und Post-Erhebungen einer ambulanten Behandlungsstichprobe Items aus der Skala "Beschwerden" des FEP (Lutz, Schürch, et al., 2009) verwendet. Die drei Methoden waren (1) herkömmliche Mittelwertsvergleiche, (2) Auswahl über Bootstrap-Konfidenzintervalle und (3) Auswahl mittels einer Latent Profile Analysis, die latente Klassen von Varianzmustern um die Itemmittelwerte schätzte. Das Bootstrap-Verfahren erwies sich am Konservativsten (4 Items) während die Auswahl mittels herkömmlicher Mittelwertsvergleiche am liberalsten war (9 Items). Die Effektstärken und Reliabilitäten der Kurzfassungen waren alle im akzeptablen Bereich.

Die Diskussion beginnt mit einer knappen Zusammenfassung der Ergebnisse der drei Studien. Im Anschluss werden die Ergebnisse der Studien auf übergreifende Aspekte bezogen. Dies sind faktorielle Validität, die Angemessenheit von Item Response Modellen zur Repräsentation psychische Belastung und die Anforderungen, die Kurzversionen letztlich erfüllen können. Insgesamt lässt sich festhalten, dass die Methoden nützliche Werkzeuge zur spezifischen Untersuchung von Skalen und zur Erstellung von Kurzformen darstellen. Besonders der in Studie II vorgestellte Bootstrap-Test der Itemauswahl stellt eine relevante Ergänzung der etablierten Vorgehensweise dar, da er empirisch belegt, dass die Auswahl für den jeweiligen Zweck einer Kurzform besser geeignet ist, als andere Items. Klinisch lässt sich festhalten, dass mit statischen Kurzversionen etablierter Messinstrumente auch in Erhebungskontexten ohne computerisierte Erhebungsmethoden hochqualitative Erhebungen durchgeführt werden können.

1. Theoretischer Hintergrund

1.1. Evidenz-basierte Überprüfung psychotherapeutischer Interventionen

Die vorliegende Arbeit beschäftigt sich mit der Frage, wie Erhebungsmethoden in der Psychotherapieforschung und besonders der Patientenorientierten Versorgungsforschung (Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lambert, Hansen, & Finch, 2001; Lutz, 2002) optimiert werden können. Die Bedeutung der Messqualität für die Evaluation von Interventionsmaßnahmen steht außer Frage (Rossi, Lipsey, & Freeman, 2004), doch wird sie in der Psychotherapieforschung dadurch unterstrichen, dass parallel zur Debatte über evidenzbasierte Behandlungen (Chambless & Hollon, 1998) auch eine Debatte über evidenzbasierte Erhebungsmethoden ("evidence-based assessment"; Hunsley & Mash, 2005) geführt wird. Die *Presidential Task Force on Evidence-Based Practice* der Amerikanischen Psychologenvereinigung hielt bei den zukünftigen Entwicklungslinien fest (APA Presidential Task Force on Evidence-Based Practice, 2006, p. 278):

> Some of the most pressing research needs are the following: [...] developing well-normed measures that clinicians can use to quantify their diagnostic judgments, measure therapeutic progress over time, and assess the therapeutic process; [...][1].

In einem Sonderheft von *Psychological Assessment* mit dem Thema "evidence-based assessment" unterstreichen die Autoren die Bedeutung dieser Methoden für die Evidenzbasierung (Hunsley & Mash, 2005, p. 251):

> Indeed, without scientifically sound assessment data, it is impossible to determine whether a treatment, patient characteristic, or therapy relationship variable has any impact on patient functioning.[2]

[1] "Einige der wichtigsten Forschungsfelder sind die folgenden: [...] Entwicklung von angemessen normierten Instrumenten, die Kliniker benutzen können, um ihre diagnostischen Einschätzungen zu quantifizieren, den therapeutischen Fortschritt erfassen helfen und Prozessmerkmale erheben können; [...]", Übers. durch Autor

[2] "Ohne wissenschaftlich angemessene Erhebungsdaten ist es nicht möglich festzustellen, ob eine Intervention, ein Patientencharakteristikum oder Therapiebeziehungsvariable irgendeinen Einfluss auf das Funktionsniveau des Patienten hat." Übers. durch Autor

Evidenz-basierte Erhebungsmethoden ("evidence-based assessment") wird dabei durch zwei Merkmalsgruppen definiert (Hunsley & Mash, 2005):

- Es umfasst die Forderung, dass die verwendeten Instrumente akzeptablen Qualitätsstandards mit besonderer Betonung auf Reliabilität und Validität genügen (in Deutschland vergleichbar mit den klassischen Testgütekriterien; z.B. Fisseni, 2004; Moosbrugger & Kelava, 2007).
- Zusätzlich soll auch die Nützlichkeit der Erhebung für drei Bereiche berücksichtigt werden (Hunsley & Mash, 2005): a) ist die Erhebungsmethode an sich eine Intervention, die zu positiven Ergebnissen führt; b) inwiefern hilft die Methode, eine genaue Diagnose zu stellen; sowie c) inwieweit ist die Methode vor dem Hintergrund weiterer klinischer Kontextfaktoren wie den entstehenden Kosten der Erhebung, der Verbesserung der Steigerung von Sensitivität und Spezifität, etc. eine sinnvolle Ergänzung?

Wie weit es bereits gelungen ist, dem Ziel validierter, inhaltsrelevanter und psychometrisch akzeptabler Messinstrumente nahezukommen, kann als derzeit noch offen bewertet werden (Barlow, 2005; Doucette & Wolf, 2009; Jensen-Doss, 2011; McFall, 2005). Wendet man sich ohne dies der Frage der Lage der Evidenz für die Wirksamkeit psychotherapeutischer Interventionen zu, hat sich seit Eysenck's Feststellung (Eysenck, 1952) das Bild gewandelt[3]. Wenn sich die Einbindung des psychotherapeutischen Angebots zwischen den Nationen unterscheidet, so ist Psychotherapie doch eine feste Größe in der Gesundheitsversorgung der westlichen Industrienationen (Lutz & Grawe, 2007; Schulte, 2007; Strauss & Kaechele, 1998). Viele Arten von psychotherapeutischen Interventionen sind wirksam (Übersichten: American Psychological Association Presidential Task Force on Evidence-Based Practice, 2006; Chambless & Ollendick, 2001; Cuijpers, van Straten, Warmerdam, & Smits, 2008; Grawe, Donati, & Bernauer, 1994; Lambert & Ogles, 2004; Lipsey & Wilson, 1993; M. L. Smith & Glass, 1977; Wampold et al., 1997)[4]. Empirisch zeigte sich aber früh, dass die Unterschiede zwischen den einzelnen Therapieformen vergleichsweise gering sind

[3] Überblick über die unterschiedlichen Phasen der Psychotherapieforschung siehe z.B. Grawe (1997) und Orlinsky, Ronnestad, & Willutzky (2004).

[4] Für eine Analyse des Evidenzkonzepts und die Position, dass keine hinreichende Evidenz für die Wirksamkeit von Psychotherapie vorliegt, die methodischen Anforderungen genügt, siehe Möller (2012).

(Luborsky, Singer, & Luborsky, 1975; Shapiro & Shapiro, 1982; Smith & Glass, 1977). Dieser Befund wird auch durch heutige Forschung immer wieder bestätigt (z.B. Ahn & Wampold, 2001; Cuijpers, van Straten, Andersson, & van Oppen, 2008; Wampold et al., 1997) und auch wenn methodische Punkte kritisch zu sehen sind (Crits-Christoph, 1997; DeRubeis, Brotman, & Gibbons, 2005; Howard, Krause, Saunders, & Kopta, 1997; Krause, Lutz, & Saunders, 2007; Westen, Novotny, & Thompson–Brenner, 2004), bleibt dennoch, dass im Mittel geringe Unterschiede zwischen denjenigen Behandlungen festzustellen sind, die als wirksam evaluiert wurden

Inhaltlich könnte mit Grawe (1997) argumentiert werden, dass nun die Fragen, ob a) Psychotherapie überhaupt wirkt und b) welche Form die Beste ist, in den Hintergrund treten sollten. Stattdessen könnte die Frage, welche Therapie für wen indiziert ist, in den Vordergrund rücken[5]. Diese Frage stellt jede einzelne Therapie in Forschung und Praxis in einen größeren Kontext von Variablen, die alle den Ausgang der Therapie beeinflussen können (z.B. Stiles, Shapiro, & Elliott, 1986: "Matrix-Paradigma"). Aus dieser Sicht spielen insbesondere vorliegende Moderatoren und Mediatoren eine besondere Rolle bei der Beurteilung der Wirksamkeit von Interventionen (Kazdin, 2009; Kraemer & Gibbons, 2009) und wie im Folgenden dargestellt werden wird, haben aktuelle Forschungsparadigmen unterschiedliche Antworten auf diese Herausforderung gefunden.

In der vorliegenden Arbeit geht es um die Bedeutung und Optimierung der Messmethodik im Kontext der Psychotherapieforschung. In allen aktuellen Forschungsparadigmen zur Wirksamkeit und Wirkungsweise psychotherapeutischer Interventionen kommt der Messung der relevanten Variablen eine hohe Bedeutung zu. Doch die Patientenorientierten Versorgungsforschung ist mit ihren kontinuierlichen Rückmeldungen besonders auf eine akkurate Erfassung der Prozess- und Ergebnisvariablen angewiesen. Es soll im Folgenden zunächst ein Überblick über die Forschungstraditionen der Psychotherapieforschung gegeben werden (1.2.1, 1.2.2) und vor diesem Hintergrund eine Einordnung der Patientenorientierten Versorgungsforschung vorgenommen werden (1.2.3). Daran anschließend soll die besondere Rolle des Themas "Messung" an Beispielen aus der Forschung

[5] Dies wiederholt den von Paul (1967, p. 111) artikulierten Punkt: welche durch wen durchgeführte Behandlung ist bei diesem Individuum am effektivsten für dieses spezielle Problem und unter diesen Umständen? ("*What* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of circumstances?" Hervorhebungen im Original; Übers. durch Autor

belegt werden (1.3). Die beiden sich an diese einführenden Kapitel anschließenden Teile der Arbeit stellen die besonderen Herausforderungen der Patientenorientierten Versorgungsforschung und ihre Verbindung zu Diagnostik und evidenzbasierten Erhebungen in den Mittelpunkt (1.4), um mit einer Vorstellung der in dieser Arbeit zusammengefassten empirischen Studien abzuschließen (1.5).

1.2. Hintergrund: Traditionen der Psychotherapieforschung

1.2.1. Efficacy & Effectiveness

Die Wirksamkeitsforschung setzt die Forschung fort, die den Befund der allgemeinen Wirksamkeit therapeutischer Interventionen hervorgebracht hat. Sie befasst sich nomothetisch mit der Frage, welche Veränderungen an Populationen auftreten, wenn sie Interventionen erhalten. Diese Forschungsrichtung lässt sich in zwei Zweige aufteilen. Der eine Zweig beschäftigt sich mit der experimentellen Prüfung der Wirksamkeit (engl. *efficacy*; Howard et al., 1996; Lambert & Ogles, 2004; Lutz & Grawe, 2007). Dieser Begriff wird v.a. auf die Prüfung von Interventionen in klinischen Studien bezogen. Diese Studien prüfen an klar umrissenen Patientengruppen[6] ob eine abgrenzbare Intervention zum erwünschten Effekt führt. Unter Annahme der Geltung der statistischen Annahmen für eine randomisiert-kontrollierte Studie kann mit diesem Design der maximale Effekt einer Intervention unter bestmöglichen Bedingungen geschätzt werden (Everitt & Wessely, 2008; Hsu, 1989; Krause & Howard, 2003; Persons & Silberschatz, 1998).

Der zweite Zweig bezeichnet solche Studien, die die Wirksamkeit von Interventionen unter Praxisbedingungen prüfen (im Englischen abgegrenzt als *effectiveness*; Howard et al., 1996). Dass eine Intervention unter den Bedingungen einer kontrollierten Studie einen gewünschten Effekt erzielt, bedeutet nicht, dass sich dieser (bzw. in dieser Höhe) auch in der Praxis zeigen lässt. Die Unterschiede zwischen Therapien in der Praxis und solchen unter kontrollierten Bedingungen sind vielfältig (Barkham et al., 2008; Chambless & Ollendick, 2001; Lutz & Böhnke, 2010; Lutz & Grawe, 2007; Möller, 2012; Persons & Silberschatz, 1998; Sexton & Kelley, 2010; Stirman, DeRubeis, Crits-Christoph, & Rothman, 2005; Thyer & Pignotti, 2011). Einige Kernunterschiede

[6] In der ganzen Arbeit wird das generische Maskulinum verwendet. Dies dient ausschließlich der Lesbarkeit des Textes und soll keine sprachliche Diskriminierung darstellen.

sind, dass Psychotherapie in der Praxis nicht streng nach einem Manual durchgeführt wird, die Therapiedauer eher von der Gesundheitsversorgung als von Manualen (oder den Belastungsgraden) abhängig ist (Chiles, Lambert, & Hatch, 1999; Hansen, Lambert, & Forman, 2002; Lutz, Böhnke, Köck, & Bittermann, 2011) und die Patienten in der Regel komorbid oder multimorbid belastet sind (Albani, Blaser, Geyer, Schmutzer, & Brähler, 2010, 2011; Wittchen & Jacobi, 2001). Aufgrund dieser Unterschiede werden die Ergebnisse der Efficacy-Forschung von Praktikern oft auch nicht als relevant für ihre alltägliche Arbeit angesehen (J. P. Shapiro, 2009; Stewart & Chambless, 2007, 2010; Stewart, Stirman, & Chambless, 2012).

Die naturalistische Forschung bringt den Vorteil größerer Praxisnähe mit sich. Die Etablierung von Wissenschaftler-Praktiker-Netzwerken (Barkham et al., 1998, 2001; Bickman & Hoagwood, 2010; Borkovec, Echemendia, Ragusea, & Ruiz, 2001; Castonguay, 2011; Kazdin, 2008; Locke et al., 2011) sowie von großen Versorgungsprojekten, konnten aufzeigen, dass Wissenschaft und Praxis gewinnbringend miteinander verbunden werden können (Lutz, Böhnke, Köck, et al., 2011; Puschner & Kordy, 2010; Steffanowski et al., 2011; Wittmann et al., 2011), was zur Jahrhundertwende durchaus noch bezweifelt wurde (Lambert, Whipple, et al., 2001). Durch solche Kooperationen werden große Patientenzahlen erhoben (Gilbody, House, & Sheldon, 2002a; Howard et al., 1996) und ihre Entwicklung unter unveränderten Routinebedingungen dokumentiert ("treatment as usual", TAU; Garland, Bickman, & Chorpita, 2010). Das Dosis-Wirkungs-Modell in der Psychotherapie (Howard, Kopta, Krause, & Orlinsky, 1986; Lambert, Hansen, & Finch, 2001), das Phasenmodell Psychotherapeutischer Veränderung (Howard, Lueger, Maling, & Martinovich, 1993) oder das Modell eines "hinreichend guten Therapieergebnisses" ("good enough level"; Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009; Barkham et al., 2006; Reese, Toland, & Hopkins, 2011) sind Ergebnisse solcher Forschungsbemühungen.

Somit erfüllen beide Zweige wichtige Aspekte der Wirksamkeitsprüfung und schließen sich nicht gegenseitig aus, sondern haben das Potential, sich zu bereichern (Howard et al., 1996; Lambert, Hansen, et al., 2001; Lutz, Böhnke, & Köck, 2011; Lutz, Köck, & Böhnke, 2009; Schindler, Hiller, & Witthöft, 2011; Stewart & Chambless, 2009). Wenn auch für die Feststellung, dass eine Behandlung empirisch fundiert ist ("empirically supported treatment") der replizierte Befund aus

kontrollierten Studien genügt (Chambless & Ollendick, 2001), so besteht doch Konsens, dass Interventionen nur dann als für den Praxiskontext hinreichend validiert gelten können, wenn sie in beiden Paradigmen und möglichst mehreren Kontexten repliziert wurden (Contopoulos-Ioannidis, Alexiou, Gouvias, & Ioannidis, 2008; Jensen-Doss, 2011; Krampen, Schui, & Wiesenhütter, 2008; Lutz & Grawe, 2007; Rounsaville, Carroll, & Onken, 2001).

Innerhalb der Wirksamkeitsforschung können verschiedene weitere Forschungsstrategien unterschieden werden. Die erste Gruppe dieser Ansätze untersucht Behandlungsprogramme in ihrer Wirksamkeit bei bestimmten Populationen oder die Wirksamkeit bestimmter Techniken (Lutz & Bittermann, 2010). Ausgangspunkt dieser Ansätze ist die Annahme, dass es systematische Unterschiede in der Wirksamkeit zwischen Behandlungsprogrammen oder bestimmten Techniken gibt – die Empirie zeige aber bislang nur, dass der Bereich der jeweils maximalen Wirksamkeit schmal sei bzw. bislang nicht entdeckt wurde (siehe z.B. DeRubeis et al., 2005; Stiles et al., 1986). Die Forschung zur Evidenzbasierung der Psychotherapie ("empirically supported treatments"; Sonderheft *Journal of Clinical and Consulting Psychology*, 1998, 66(1); Chambless & Ollendick, 2001; Ollendick & King, 2004) hat das Ziel, definierte Behandlungsprogramme mit spezifischen Zielpopulationen unter kontrollierten Bedingungen auf ihre Wirksamkeit zu prüfen (Chambless & Hollon, 1998). Dieses Paradigma sieht als relevante Variablen a) definierte Populationen (in der Regel durch eine Diagnose charakterisiert, oft qualifiziert mit demographischen Variablen wie Alter oder Geschlecht oder anderen Patientenmerkmalen) und b) die Manualisierungen (der Intervention) an. Dies reduziert die "Matrix" (Stiles et al., 1986) auf zwei definitorische Elemente. In diesem Paradigma wird betont, dass die Behandlungsmanuale unterschiedliche Menschenbilder und theoretische Inhalte umfassen, die nicht als äquivalent angesehen werden können, sowie die Arbeiten, die zeigen, dass Therapien sich doch oft als differentiell wirksam erwiesen haben und sich somit die Orientierung an bestimmten Therapieschulen begründen und an ihnen festhalten ließe (beispielsweise: Berger & Linden, 2009; DeRubeis et al., 2005).

Ansätze des technischen Eklektizismus verlassen die Ebene der Therapieschulen und zielen stattdessen auf ein System von empirisch validierten Interventionstechniken (Beutler & Harwood, 2000; Beutler, Moleiro, & Talebi, 2002; Beutler, 1998, 1999; Lutz & Bittermann, 2010; Wampold,

2001), die unabhängig vom theoretischen Hintergrund kombiniert können werden sollen. Die Effekte der Therapie werden als Resultat spezifischer Elemente (z.B. einzelner Techniken oder des ganzen Programes/Manuals) gesehen. Auch assimilativ-integrative Ansätze (z.B. Castonguay, 2011) versuchen unterschiedliche Schulen zu integrieren, allerdings wird hier ausgehend von einem bestehenden Ansatz versucht, Elemente anderer Therapieformen aufzunehmen, die besonders solchen Patientengruppen helfen, die noch nicht optimal von den ursprünglichen Therapien profitieren (z.B. Kognitive Verhaltenstherapie bei Generalisierter Angststörung, die durch emotional-interpersonelle Elemente angereichert wird; Newman, Castonguay, Borkovec, & Molnar, 2004).

Theoretisch-integrative Ansätze versuchen die Matrix relevanter Therapievariablen unabhängig von den bestehenden Schulen zu entwickeln, indem sie Wirkfaktoren definieren, die therapeutischen Prozessen zugrunde liegen. Demnach gehen die beobachteten Unterschiede zwischen Therapieprogrammen/-schulen auf Unterschiede der in ihnen realisierten Wirkfaktoren zurück. Diese Faktoren müssen von den im Englischen als "common factors" (Wampold, 2001) bezeichneten getrennt werden, denn während diese nur die therapeutischen Elemente bezeichnen, die allen therapeutischen Settings gemeinsam sind (s.u.), bezeichnen die hier gemeinten Wirkfaktoren sowohl Elemente, die Schulen gemeinsam sind als auch spezifische Elemente, die nur bei einer Therapieschule auftreten könnten. Grawe (1998, 2004) versuchte die wesentlichen Elemente des Psychotherapieprozesses durch spezifische Techniken wie auch kontextuellen Faktoren in der Therapie durch vier von ihm postulierte Wirkfaktoren (Problemaktualisierung, Ressourcenaktivierung, aktive Hilfe zur Problembewältigung und motivationale Klärung) zu umschreiben. Diese liegen nach Grawe allen psychotherapeutischen Veränderungsprozessen zugrunde und sind z.B. in den jeweiligen Schulen, den verwendeten Strategien, in jeder Intervention und letztlich jeder Technik (Wampold, 2001) in ihrer Kombination unterschiedlich ausgeprägt. Ein anderes integratives Modell, das "Generic Model" (Howard & Orlinsky, 1972; Orlinsky & Howard, 1986; Orlinsky, 2009) wurde als theorieübergreifender Rahmen entwickelt, der es ermöglichen sollte, unterschiedliche Befunde zu integrieren. Insgesamt entwickelte es sich jedoch auch zu einem Modell, das ausgehend von Prozesselementen der Psychotherapie definierte, welche Variablen bei welchen Patienten und Therapeuten zu einem günstigeren Therapieergebnis führen (Orlinsky, 2009).

Messen die bisher geschilderten Ansätze den Therapieschulen und –techniken großen Wert bei, betont eine andere Tradition besonders die Ähnlichkeiten der therapeutischen Ansätze: Demnach müssen diejenigen Variablen zur Effektivität der Psychotherapie führen, die in allen psychotherapeutischen Kontexten präsent sind (Frank & Frank, 1991; Norcross & Lambert, 2011; Norcross & Wampold, 2011; Wampold, 2001; für Vergleiche der Positionen spezifische vs. nicht-spezifische Wirkfaktoren, siehe z.B.: DeRubeis et al., 2005; Lutz, Ehrlich, & Zaunmüller, 2010).

Die allgemeinste Position des Matrix-Paradigmas ist die Position von Merton S. Krause. Seine Arbeiten betonen, dass das Ausblenden auch nur einer Interaktionen der Variablen, die das Therapieergebnis beeinflussen, zu schwer interpretierbaren (und im Extremfall unverwertbaren) Forschungsergebnissen führt, was deren Bedeutung für die Praxis mindert. Aktuell drängt er darauf, zumindest hinreichende Kausalitätsbedingungen für Veränderungen in der Psychotherapie (und menschlichem Verhalten generell) zu identifizieren. Er greift damit frühere Positionen auf, die eine theoretische geleitete Auswahl von relevanten Kombinationen der Matrix der Veränderungsbedingungen forderten (Stiles et al., 1986). Eine Weiterentwicklung der Forschung sei nur durch eine Weiterentwicklung der Forschung über reine Mittelwertsvergleiche hinaus möglich (Kraemer & Gibbons, 2009; Ruberg, Chen, & Wang, 2010; Thase, Larsen, & Kennedy, 2011). In aktuelleren Arbeiten versucht er einen geeigneten mathematisch-methodischen Konzeptrahmen für diese Forschung zu entwickeln (Krause & Howard, 1999; Krause, Lutz, & Boehnke, 2011; Krause et al., 2007; Krause & Lutz, 2009; Krause, 2010).

1.2.2. Wirkungsweise und Prozessforschung

Eine andere eingeforderte Perspektive auf den therapeutischen Prozess ist die Forschung zur Wirkungsweise von Psychotherapie (Grawe, 1982). Wie oben bereits angerissen, gibt es bislang keine abschließenden empirischen Entscheidungen darüber, wie Veränderung im therapeutischen Setting zustande kommt (Orlinsky, Grawe, & Parks, 1994). Daher stellt die weitere Untersuchung von Faktoren, die zu Veränderung führen, ein zentrales Element der Psychotherapieforschung dar. Für diesen Zweck wurden eine ganze Reihe von Designs entwickelt, die es ermöglichen die vermittelnden Effekte von Behandlungskomponenten (Mediatoren) wie auch augmentierend/ diminuierend auf den Therapieerfolg wirkende Elemente (Moderatoren) zu prüfen. Damit wurden auch die

Designs der in 1.2.1 beschriebenen Forschung verändert (z.B. Kazdin, 1998; 2009). Auch neuere Entwicklungen in der statistischen Methodik zu einer Verbesserung der Auswertung klinischer Studien, die z.T. eher in der Ökonometrie übliche Methoden verwenden, zielen auf die Untersuchung solcher Moderatoren und Mediatoren ab (Emsley, Dunn, & White, 2010).

Die Prozessforschung, die ebenfalls in die Forschung zur Wirkungsweise zu rechnen ist, befasst sich mit der Untersuchung von Prozessmerkmalen der Therapie, d.h. wie der therapeutische Prozess zustande kommt als auch in welcher Beziehung Prozessmerkmale zum Therapieergebnis stehen (für einen Überblick z.B. Elliott, 2010; Lutz & Grawe, 2007; Orlinsky, Grawe, & Parks, 1994). Prozessmerkmale, die mit einem positiven Therapieergebnis in Beziehung stehen sind demnach (Orlinsky et al., 1994): Die Qualität der therapeutischen Beziehung, die Kompetenz des Therapeuten, Kooperation des Patienten, Offenheit des Patienten für Veränderungen und die Behandlungsdauer. Gerade die Forschung zur Qualität der therapeutischen Beziehung erhielt viel Aufmerksamkeit und bislang ist ungeklärt, ob sie zu den kausalen Faktoren für eine gutes Therapieergebnis zu rechnen ist (Flückiger, Del Re, Wampold, Symonds, & Horvath, 2012; Horvath & Symonds, 1991; Lambert & Barley, 2001; Martin, Garske, & Davis, 2000) oder eines unter vielen Prozessmerkmalen ist (Crits-Christoph, Gibbons, & Hearon, 2006; Hentschel, 2005a, 2005b).

Beide Forschungsaspekte können nomothetisch oder ideographisch ausgerichtet sein, quantitativ oder qualitativ verfolgt werden (Kazdin, 2008). Ein gutes Beispiel ist die Forschung zur therapeutischen Allianz, für die nomothetisch-quantitativ etabliert ist, dass sie positiv mit dem Therapieergebnis in Verbindung steht (Horvath, Del Re, Flückiger, & Symonds, 2011; Horvath & Symonds, 1991; Martin et al., 2000) und ergänzend hierzu die qualitative Prozessforschung mit verschiedenen Methoden untersucht, wie diese Verbindung zustande kommt (Hill & Knox, 2009). Die Forschung zu plötzlichen Gewinnen und Verlusten in der Psychotherapie zeigt diese Verbindung in ähnlicher Weise: Nomothetische Ergebnisse zu der Wirkung von Gewinnen und Verlusten in der Psychotherapie (Lutz & Tschitsaz, 2007; Lutz et al., in press; Tang & DeRubeis, 1999a; Tschitsaz-Stucki & Lutz, 2009) parallelisieren Forschung zur Bedeutung einschneidender Ereignisse und therapeutischer Elemente in der Psychotherapie (Elliott, 2010; Gonçalves & Stiles, 2011; Henretty, Levitt, & Mathews, 2008; Knox et al., 2011; Levitt & Piazza-Bonin, 2011; Williams & Levitt, 2008).

1.2.3. Patientenorientierte Versorgungsforschung

Die vorliegende Arbeit ist neben den naturalistischen Wirksamkeitsstudien (s. 1.2.1) der Patientenorientierten Versorgungsforschung verpflichtet. Die Patientenorientierte Versorgungsforschung (Howard et al., 1996; Lambert, Hansen, et al., 2001; Lutz, 2002, 2011) ist eines der zentralen Paradigmen zur Verbesserung der psychotherapeutischen Versorgung, da sie zwischen der akademischen Forschung und Entwicklung auf der einen Seite und der praktischen Anwendung der Psychotherapie am Einzelfall und Qualitätssicherung auf der anderen Seite vermittelt (Lutz & Bittermann, 2010; Lutz, 2011). Sie beschäftigt sich mit der Frage, ob eine bereits laufende Intervention bei einem spezifischen Patienten wirkt und wie die Behandlungsentscheidungen durch Rückmeldungen in den Therapieverlauf unterstützt werden können (Lutz, 2002). Dies ist die für Kliniker zentrale Frage und eine Aufgabe von Forschung sollte es daher sein, Kriterien zu entwickeln, die es ihm ermöglichen, valide zu ermitteln, ob sich ein Patient verbessert oder verschlechtert (Howard et al., 1996; dieser Aspekt wird in 1.4.1 aufgegriffen).

Patientenorientierte Versorgungsforschung versucht statt einer rein nomothetischen Betrachtung (die außer der Prozessforschung allen obigen Forschungsperspektiven in der Regel zugrunde liegt) auch ideographische Elemente zur Beurteilung des Therapieprozesses heranzuziehen. Dies ist trägt zur Praxisorientierung bei, da aus den Veränderungen von Gruppenmittelwerten keine Aussagen über die Entwicklung eines Individuums möglich sind (Kraemer et al., 2003; Krause et al., 2011; Molenaar, 2004; Molenaar & Campbell, 2009; Schmitz, 2000): Aus der Wirksamkeitsforschung folgt lediglich, dass behandelte Patienten im Mittel z.B. eine Erleichterung ihrer depressiven Symptomatik erleben. Das bedeutet aber nicht, dass sich jeder einzelne Patient auch gebessert haben muss. Zu einer aktuellen Debatte dieses Themas sei auf Krause (2011a, 2011b) verwiesen. Wenn "evidenzbasiertes Handeln" im Sinne der in 1.2.1 beschriebenen Efficacy-Forschung verstanden wird als Verwendung von Methoden, die sich in kontrollierten Studien als wirksam erwiesen haben, könnten sich Patienten trotz solcher Interventionen negativ entwickeln. Damit stellt sich die Frage, was "evidenzbasiertes Handeln" im Einzelfall bedeuten kann (Perrez, 2005; Persons & Silberschatz, 1998; Thyer & Pignotti, 2011; Zayas, Drake, & Jonson-Reid, 2011).

Zusätzlich muss festgehalten werden, dass die Übersetzung evidenzbasierter Techniken trotz Wissenschaftler-Praktiker-Netzwerken (siehe 1.2.1), der Sammlung von evidenzbasierten Behandlungen (ebenfalls 1.2.1 und z.B. die Cochrane Collaboration) oder auch praxisorientierte Buchserien nur bedingt geschieht (Stewart & Chambless, 2007, 2010; Stewart et al., 2012). Die Befundlage gerade zu speziellen Techniken und Interventionen bei bestimmten Patientengruppen ändert sich stetig und mit diesen Entwicklungen Schritt zu halten, ist für niedergelassene Psychotherapeuten kaum möglich (z.B. allein die Kosten für die nötigen Fachzeitschriften, Newnham & Page, 2010; für eine Diskussion dieser Problematik bezogen auf Kinder-/ Jugendpsychotherapie sei auf die Sonderausgabe von *Administration and Policy in Mental Health and Mental Health Services Research* verwiesen: Bickman & Hoagwood, 2010).

Die Patientenorientierte Versorgungsforschung ist ein Weg, um einen Dialog zwischen dem Praktiker und der Forschung zu eröffnen. Wie weiter unten vorgestellt wird (1.4.2), ist der Blickwinkel der Patientenorientierten Versorgungsforschung auf den Therapieprozess besonders effektiv bei sich negativ entwickelnden Fällen. Wird ein Fall durch ein solches im Rahmen der Patientenorientierten Versorgungsforschung entwickeltes Rückmeldesystem als sich negativ entwickelnd entdeckt, kann konkret nach empirischer Evidenz zu Behandlungsansätzen und Interventionen zur Behandlung gesucht werden oder konkret in einen klärenden Dialog mit dem Patienten eingestiegen werden (Howard et al., 1996; Lutz & Bittermann, 2010; Lutz, 2002).

Auf der Seite der Forschung kann in der Patientenorientierten Versorgungsforschung jeder Patient als ein Fall der Matrix aus Variablen betrachtet werden, die den Therapieerfolg bedingen (Stiles et al., 1986). Werden Methoden der Patientenorientierten Versorgungsforschung in der Routine etabliert, konsequent fortgeführt und durch eine breite Auswahl relevanter Variablen gestützt, dann können die zunächst für die Qualitätssicherung (s. 1.4.2 und 1.4.4) gesammelten Daten später integriert und im Sinne naturalistischer Studien zur Wirksamkeit ausgewertet werden (Ellwood, 1988; Gilbody et al., 2002a; Howard et al., 1996; Lutz, 2002). So integrieren sich die gesammelten Daten aus dem eher ideografischen Vorgehen dann in die Forschungsdatenbasen der Effectiveness-Forschung, z.B. von Wissenschaftler-Praktiker-Netzwerken (Barkham et al., 2001; Gilbody et al., 2002a; Howard et al., 1996; Lutz, Böhnke, Köck, et al., 2011). Kapitel 1.4 widmet sich dieser For-

schung ausführlich und es werden dort auch Befunde zur Feedbackforschung präsentiert, die unterstreichen, welche Rolle diese Forschung auch praktisch spielen kann und inwiefern Patientenorientierten Versorgungsforschung als Evidenz-basierte Erhebung bezeichnet werden kann (Hunsley & Mash, 2005; Lambert, 2007).

1.3. Die Bedeutung von "Messung" in den behandelten Kontexten

Nach dem Überblick über die Forschungstraditionen der Psychotherapieforschung soll der Blick übergreifend auf das Thema "Messung" gelenkt werden. Die vorliegende Arbeit untersucht wie die Messqualität in Anwendungen der Patientenorientierten Versorgungsforschung verbessert werden kann (Hunsley & Mash, 2005). Dies soll an einigen ausgewählten Beispielen kurz eingeführt werden, bevor sich Kapitel 1.4 der Patientenorientierten Versorgungsforschung im Speziellen zu wendet. Kapitel 1.2.1 bis 1.2.3 sollten verdeutlichen, dass zumindest alle quantitativen Vorgehensweisen Anforderungen wie in anderen typischen quantitativen Forschungsparadigmen stellen: Eine angemessene Operationalisierung, Konstruktvalidität und Messqualität sind nötig. Dies trifft für die Ergebnismaße der Efficacy- und Effectiveness-Forschung genauso zu wie für die Messung der Mediatoren/Moderatoren in der Prozessforschung.

Ein erstes Beispiel sei aus der Efficacy-Forschung herausgegriffen. In dieser Forschungsrichtung ist die Frage, wie viele Patienten in einer Stichprobe nötig sind, um einen Effekt nachzuweisen von zentraler Bedeutung (Maxwell & Kelley, 2011). Dies wird in der Regel darüber gelöst, dass vor einer Studie eine Stichprobenumfangsplanung vorgenommen wird (J. Cohen, 1988; Faul, Erdfelder, Lang, & Buchner, 2007). Das Ergebnis dieser Analyse hängt von der Größe des erwarteten Effektes und z.B. bei standardisierten Mittelwertsunterschieden von der Streuung ab:

$$d = \frac{m_1 - m_2}{SD} \qquad \textbf{Formel 1-1}$$

Die Streuung *SD* ist hier z.B. die gepoolte Standardabweichung aus unterschiedlichen Gruppen (Lutz & Grawe, 2007). Die beobachtete *SD* in einem Instrument hängt von der Messqualität des Instrumentes ab. Der Standardmessfehler (*SE*) beschreibt den Anteil der beobachteten Variation, der auf unsystematische Messungenauigkeit zurückzuführen ist:

$$SE = SD * \sqrt{1 - r_{xx}}$$ **Formel 1-2**

Die für klinische Studien wichtige Feststellung ist, dass bei steigender Reliabilität des Instrumentes die beobachtete SD immer stärker die tatsächliche Variabilität des Merkmals widerspiegelt. Dadurch werden die gesuchten Effekte leichter nachweisbar, da die Variation zunehmend weniger von Zufallsschwankungen überlagert ist. Eine Simulationsstudie zur Untersuchung, wie Item Response Modelle (Kempf, 2008; Rost, 2004) die Messung der primären Endpunkte in klinischen Studien verbessern könnten, zeigte, dass bei bis zu 30 verwendeten dichotomen Items die Power zur Identifikation eines Interventionseffektes (d = .2, .5, .8) deutlich zunahm, danach aber nur noch eine geringe Steigerung zu entdecken war (Holman, Glas, & De Haan, 2003). Diese Unterschiede schlugen sich in der Zahl der benötigten Patienten pro Studienarm nieder: So sank die Zahl der benötigten Patienten pro Studienarm bei einer Effektstärke von d = .2 von N = 950 bei k = 5 auf ein N = 440 bei k = 100 Items (bei d = .8 immerhin noch von N = 70 bei k = 5 auf ein N = 39 bei k = 100; Holman, Glas, et al., 2003). Diese Ergebnisse konnten die Autoren auch an erhobenen Daten mit einem Instrument zur Messung des allgemeinen Gesundheitsstatus (SF-36, SF-12, SF-8; Bullinger & Kirchberger, 1998) replizieren. Dieser oft in der Planung von Studien nur unzureichend berücksichtigte Aspekt (Holman, Glas, et al., 2003) macht deutlich, dass die Messqualität in einer randomisiert kontrollierten Studie von großer Bedeutung ist.

Ein anders Beispiel aus der Prozessforschung, die Bedeutung der Sudden Gains und Sudden Losses (Tang & DeRubeis, 1999a; Tschitsaz-Stucki & Lutz, 2009), unterstreicht diesen Punkt in ähnlicher Weise. Zwei der drei von Tang und DeRubeis (1999) verwendeten Kriterien hängen aufgrund derselben, oben geschilderten Gesetzmäßigkeiten von der Messqualität ab (Formel 1-2). Das erste Kriterium, die beobachtete Differenz von sieben Scorepunkten, ist in Anlehnung an das Kriterium der reliablen Veränderung (Jacobson & Truax, 1991; Kempf, 2008) entwickelt worden und das dritte Kriterium ist ein Signifikanztest, der in ähnlicher Weise vom Messfehler abhängt.

Diese beiden Beispiele zeigen, dass die Messqualität von hoher Bedeutung für die Ergebnisse und Untersuchungsmethoden verschiedener Bereiche der Psychotherapieforschung ist (ein weiteres Beispiel ist die Erstellung von Rückmelderegeln in der Patientenorientierten Versorgungsfor-

schung: siehe 1.4.1). Der Schwerpunkt der folgenden Argumentation liegt auf dem Nutzen, den verbesserte Messinstrumente für die Patientenorientierte Versorgungsforschung haben. Die geschilderten Beispiele unterstreichen, dass Messmethoden einen entscheidenden Beitrag zu jedem Bereich der (quantitativen) Psychotherapieforschung liefern (Margison et al., 2000).

1.4. Forschung und Anwendungsfelder der Patientenorientierten Versorgungsforschung

Im Folgenden wird beschrieben, mit welchen Methoden in der Patientenorientierten Versorgungsforschung Patientenverläufe bewertet werden und wie diese von der Messung der untersuchten Merkmale abhängen (1.4.1). Danach wird die Patientenorientierte Versorgungsforschung in Bezug auf drei praktische Perspektiven eingeordnet: Zunächst die Frage der Evidenzbasierung der Messungen durch den sog. Feedback-Effekt (1.4.2), daran anschließend die Beziehung zur Diagnostik (1.4.3) und abschließend zur Qualitätssicherung (1.4.4).

1.4.1. Praxis der Patientenorientierten Versorgungsforschung: Evaluation von Psychotherapieverläufen

Wie oben dargelegt (1.2.3), können Efficacy- und Effectiveness-Forschung nur beantworten, ob eine Intervention im Mittel wirksam ist. Dies bedeutet, dass die erwartete Besserung des Patienten durch ein bestimmtes therapeutisches Vorgehen lediglich eine Hypothese ist, die es zu überprüfen gilt (Maercker, 2011; Newnham & Page, 2010; Seidenstücker, 1995). Soll ein Therapeut unter dieser Maßgabe in der Psychotherapie begründet handeln (Westmeyer, 1979), müssen ihm Bewertungskriterien für den Einzelfall an die Hand gegeben werden. Zentrale Aufgaben der Patientenorientierten Versorgungsforschung sind somit die Bildung von Maßstäben zur Bewertung von Psychotherapieverläufen und die Ermittlung relevanter Dimensionen der Veränderung[7] (Howard et al., 1996; Krause & Lutz, 2009; Lambert, 2007; Lueger et al., 2001; Lutz, 2002). Dies überschneidet sich mit den Aufgaben der Diagnostik (1.4.3) und Evaluationsforschung (Rossi et al., 2004): Es geht um die Festlegung von Evaluationskriterien, die zu einer informierten Entscheidungsfindung in der Therapieplanung beitragen (Lutz & Böhnke, 2012; Lutz & Grawe, 2007; Lutz, 2011; West-

[7] Dass solche Bewertungsmaßstäbe sinnvoll sein können, weil sie eine Ergänzung zur Perspektive des behandelnden Therapeuten darstellen, wird in 1.4.4 näher erläutert. An dieser Stelle wird nur auf Cuijpers, Li, Hofmann, & Andersson (2010) und Lambert (2007) verwiesen.

meyer, 1979). Da es in der vorliegenden Arbeit um die Optimierung der Messmethoden für die Grundlage der Erstellung solcher Bewertungsregeln (bzw. allgemeiner: Evaluationskriterien) geht, werden daher einige Ansätze zur Erstellung solcher Bewertungsregeln vorgestellt: Rationale Entscheidungsregeln und empirische Entscheidungsregeln.

Die einfachste Möglichkeit für die Festlegung von Evaluationskriterien in der Psychotherapie stellen sog. *rationale Entscheidungsregeln* dar (z.B. Lambert & Ogles, 2009; Lutz, Stulz, Martinovich, Leon, & Saunders, 2009): Rationale Entscheidungsregeln beziehen sich auf Informationen, die mit einem psychometrischen Messinstrument verbunden sind, und anhand derer vor der Therapie festgelegt werden kann, was das gewünschte Ziel der Veränderung eines Patienten ist. Sie stellen damit ein diagnostisches Vorgehen dar, das in den Bereich der kriteriumsorientierten Messung fällt. Ein Beispiel hierfür ist die Vorgehensweise von Jacobson und Truax (1991), die zwei Kriterien definierten, um die Veränderung am Einzelfall zu bewerten. Das erste Kriterium beschreibt, um wie viele Scorepunkte sich ein Patient in einem Instrument verändert haben muss, damit diese Veränderung größer als die erwartete Zufallsvariation aufgrund des Messfehlers ist (in der Regel bezogen auf 5%-Signifikanzniveau und den Standardmessfehler des Tests; Kempf, 2003). Verändert sich ein Patient in stärkerem Maße als dieser Wert, wird von *reliabler Veränderung* gesprochen. Das zweite Kriterium gibt zusätzlich vor, dass eine qualitative Veränderung des Belastungsgrades stattgefunden haben muss. Dies wird in der Regel über die sog. "Cut Off"-Werte operationalisiert, in dem z.B. festgestellt wird, ab welchem Testwert eine Person eher zu einer Normalbevölkerungsstichprobe gehört oder aber zu einer klinisch belasteten (s.a. Kapitel 3).

Diese Kriterien können in verschiedener Weise kombiniert und erweitert werden (Lunnen & Ogles, 1998; Lutz, Stulz, et al., 2009; Lutz, Tholen, Kosfelder, Grawe, & Schulte, 2005; Tingey, Lambert, Burlingame, & Hansen, 1996). Jacobson und Truax (1991) bezeichneten Patienten, die sich vor der Therapie oberhalb des Cut Offs eines Instrumentes befanden und nach der Therapie unterhalb dieses Wertes (wenn hohe Werte hohe Belastung angeben) sowie sich auch reliabel verbessert haben als "klinisch-signifikant" gebessert. Derzeit nicht entschieden, welche der Methoden

am angemessensten ist (Überblick bei: Lambert & Ogles, 2009; Newnham & Page, 2010)[8]. Diese Regeln haben sich insgesamt als wirksam erwiesen, Patienten zu identifizieren, die ein erhöhtes Risiko haben, die Behandlung nicht gebessert zu verlassen (Lambert, Whipple, et al., 2001; Lambert, Hansen, et al., 2001; für eine Diskussion der Wirkung und Nützlichkeit von Feedback im Therapieprozess, s. 1.4.2; speziell Kinder- und Jugendsettings: Bishop et al., 2005; Cannon, Warren, Nelson, & Burlingame, 2010; Warren, Nelson, & Burlingame, 2009).

Diese Regeln hängen wie bereits beschrieben von dem Verhältnis von Messfehler zu Streuung in dem verwendeten Instrument ab (Formel 1.2). Je genauer ein Instrument misst, d.h. je kleiner der Standardmessfehler oder je höher die Reliabilität, desto eher kann mit ihm festgestellt werden, ob eine Person sich verändert hat, da das Intervall für die reliable Veränderung kleiner wird. Auch kann mit größerer Sicherheit entschieden werden, ob eine Person sich von einer qualitativen Kategorie zur nächsten verändert hat (z.B. ambulante Psychotherapiepatienten zu nicht in Behandlung befindlichen Personen), da die Korrelationen mit externen Kriterien stärker werden, wenn das Instrument weniger Fehlervarianz aufweist. Die Einlösung dieser beiden Vorteile bedeutet nicht, dass das Instrument auch sensitiv für die Veränderung zwischen zwei Messzeitpunkten ist. Dies muss empirisch gesondert nachgewiesen werden, doch wird dies plausibler, wenn ein Instrument reliabel ist und damit in seiner Varianz verstärkt Unterschiede zwischen den wahren Werten der Personen abgebildet werden (Burlingame et al., 2006; Lutz, Tholen, Schürch, & Berking, 2006; Meier, 1997; siehe auch Kapitel 4).

Empirische Entscheidungsregeln stellen eine andere Möglichkeit dar, Evaluationskriterien für individuelle Patienten zu finden. Bei diesen wird versucht, die erwartete Veränderung eines Patienten auf Basis bereits behandelter Patienten vorherzusagen. Dieses Vorgehen ist eher normorientiert (Lutz et al., 2009). Für ein solches Vorgehen ist es nötig, dass die relevanten Fortschrittsmaße im

[8] In diese Kategorie rationaler Entscheidungsregeln fallen auch in anderen Forschungszweigen übliche Klassifizierungsmethoden wie die "just noticeable difference", "smallest real differences", "minimally important change" oder "minimally important differences", da sie ebenfalls vordefinierte Kriterien aufgrund psychometrischer Instrumente verwenden (Beaton, 2003; de Vet et al., 2010; Norman, Sloan, & Wyrwich, 2003; Schuck & Zwingmann, 2003; L. J. Stricker, 2000).

Verlauf der Therapie wiederholt erhoben werden und so für jeden vorher behandelten Patienten nicht nur Prä- und Post-Messungen, sondern Verläufe, sog. Trajektorien, vorliegen.

Arbeiten der Howard-Gruppe in diesem Paradigma fußten zunächst auf der von ihnen vorgeschlagenen "patient profiling" Methode (Howard, Moras, et al., 1996; Krause, Howard, & Lutz, 1998). Bei der Profilbildung für neu aufgenommene Patienten geht es zunächst darum, aufgrund vorhandener Daten bereits behandelter Patienten eine Erwartung zu formulieren, wie dieser Patient sich entwickeln wird. Lutz und Kollegen (Lutz, Martinovich, & Howard, 1999) berechneten aufgrund der vorliegenden Verläufe in einer Datenbank mittels konditionaler Wachstumsmodelle ("growth models", Raudenbush & Bryk, 2002), wie sich ein neuer Patient bei einer bestimmten Konstellation von Eingangsmerkmalen entwickeln wird. Lutz und Kollegen (1999) konnten zeigen, dass sich mit den von ihnen verwendeten Prädiktoren substantiell Varianz in den Verläufen aufklären ließ (22% der beobachteten Steigungsvariation). Es zeigte sich darüber hinaus, dass neben den Startwerten auf dem Messinstrument die Therapeuteneinschätzung des Belastungsgrades sowie Chronizität und Zahl früherer Behandlungen relevante Prädiktoren für den Verlauf der Patienten waren. Mit diesen Variablen als Prädiktoren können für zukünftige Patienten Verlaufsvorhersagen gemacht werden ("expected treatment response"; Lutz, 2002) und mittels Konfidenzintervallen um die Kurve bestimmt werden, ob der Patient sich der Vorhersage gemäß entwickelt oder aber positiv/ negativ von der Vorhersage abweicht.

Die Verbindung der nomothetischen und der ideografischen Ebenen ist erkennbar: Bisher behandelte Patienten stellen einen Bezugsrahmen dar, vor dem der individuelle Verlauf eines Patienten bewertet werden kann. Die am Einzelfall vorzunehmende Deutung und Einordnung der Entwicklung des jeweils spezifischen Falles vor diesem Hintergrund verbindet diese Modelle mit der Feedback-Forschung (1.4.3) und der Qualitätssicherung (1.4.4). Dies trifft auch für die rationalen Entscheidungsregeln zu, die ebenfalls auf vorherigen empirischen Untersuchungen fußen (insofern nutzen beide Methoden Normstichproben; Lutz, Stulz, et al., 2009). Rationale Entscheidungsregeln geben jedoch ein Kriterium an (bzw. eine Gruppe von Kriterien), das auf alle

Patienten in derselben Weise angewendet wird, während sich die Ausprägungen der Kriterien bei den empirischen Entscheidungsregeln von Patient zu Patient ändern (Lutz, Stulz, et al., 2009)[9].

Lueger und Kollegen zeigten auf, wie Monitoring Systeme inklusive Rückmeldungen über die Qualität des Verlaufs an die Therapeuten mit dieser Methode entwickelt werden können (Lueger et al., 2001). In einer weiteren Studie zeigte sich, dass die vom Modell erwarteten Veränderungsraten sich mit den empirisch beobachteten deckten und das Vorhersagemodell sowohl für verschiedene Störungsgruppen wie für die Dimensionen des Phasenmodells valide war (Lutz, Lowry, Kopta, Einstein, & Howard, 2001). Eine Erweiterung und Verbesserung stellen sogenannte "adaptive expected treatment response"-Modelle dar, bei denen die Veränderung zwischen Beginn der Therapie und der derzeitigen Zwischenmessung zur Verbesserung der Verlaufsvorhersage verwendet wird (Lutz, Rafaeli, Howard, & Martinovich, 2002). So können Forschungsergebnisse zu frühen Veränderungen in der Therapie (Lambert, 2007; Lutz & Tschitsaz, 2007; Tang & DeRubeis, 1999; Tschitsaz-Stucki & Lutz, 2009) wie auch solche zur Unterschiedlichkeit der Verlaufsformen in frühen Therapiephasen (Lutz, Stulz, Smart, & Lambert, 2007; Stulz, Lutz, Leach, Lucock, & Barkham, 2007; Stulz & Lutz, 2007) berücksichtigt werden.

Schließlich stellen diese Modelle auch eine Basis für die Untersuchung von Therapeuteneffekten dar, da sie es einerseits wie alle Mehrebenenmodelle möglich machen, unterschiedliche Varianzkomponenten voneinander zu trennen (Gelman & Hill, 2007), aber zusätzlich es auch ermöglichen, für relevante Unterschiede im "Case Mix" (Margison et al., 2000) zwischen Therapeuten oder Institutionen zu korrigieren, weil diese konfundierenden Variablen als Prädiktoren verwendet werden können (Crits-Christoph et al., 1991; Lambert & Baldwin, 2009; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Lutz, Martinovich, Howard, & Leon, 2002; Okiishi, Lambert, Nielsen, & Ogles, 2003 und das Sonderheft von *Psychotherapy Research, 16*(2), 2006).

[9] Ein vollständiger Übergang zur ideographischen Perspektive wird immer wieder gefordert (Molenaar & Campbell, 2009; Molenaar, 2004; Schmitz, 2000) und würde ebenfalls eine Anwendung dieser Kriterien ermöglichen (Bergmann-Warnecke, 2011); das Synergetic Navigation System ist allerdings das einzige Therapiemonitoringsystem, das bislang auf diesem Ansatz beruht (Schiepek & Strunk, 2010; Schiepek, Zellweger, Kronberger, Aichhorn, & Leeb, 2011).

Die beschriebene Vorgehensweise beruht stark auf den statistischen Modellannahmen der konditionalen Mehrebenenmodelle. Diese sind zwar flexibler als klassische ANOVAs (Lutz et al., 1999; Quené & van den Bergh, 2004), da sie keinen festen Messplan für die messwiederholte abhängige Variable benötigen, die Varianzen und Kovarianzen der modellierten Variablen nicht konstant sein müssen und durch diese Modelle auch das Problem gelöst wird, dass unterschiedliche Ebenen der Erhebungsmethodik auch unterschiedlichen Einfluss auf die Varianz der AV sowie der verwendeten Prädiktoren nehmen können (Curran & Bauer, 2011). Die Methode benötigt aber die Annahme der multivariaten Normalverteilung (oder andere Verteilungen, wenn die messwiederholte abhängige Variable nicht kontinuierlich ist), die es möglich macht, die genannten Probleme der ANOVA-basierten Ansätze zu lösen. Eine unkritische Übernahme dieser Annahme verdeckt, dass es Kombinationen von Variablenwerten auf Prädiktorseite geben kann (oder Regionen im multidimensionalen Raum), in denen keine Fälle beobachtet werden. Für diese werden aber aufgrund der Modellannahmen trotzdem Schätzer generiert, bzw. aus der erhaltenen Modellgleichung Vorhersagen möglich. Angeregt von Vorhersagemodellen in der Lawinenforschung (Brabec & Meister, 2001) wurde versucht, Vorhersagen durch sog. Nächste-Nachbarn-Methoden ("nearest neighbors"; zur Stichprobenabhängigkeit der Verlaufsmodelle: Krause et al., 1998; Einführung und geschichtliche Ableitung dieses Modells: Lutz et al., 2005) zu verbessern und klinisch valider zu gestalten.

Bei der Vorgehensweise mit "Nächsten-Nachbarn" werden mittels definierter Eingangsvariablen Fälle ausgewählt, die einem neu aufgenommenen Patienten in ihren Charakteristika ähnlich sind. Diese Charakteristika sollen in Bezug auf die Vorhersage des Behandlungsverlaufes die Stichprobe in relevante Subgruppen zerlegen, damit basierend auf den ähnlichsten Fällen eine Vorhersage vorgenommen werden kann. Eine erste, in dieser Weise noch nicht benannte Anwendung stellt eine Studie aus der Lambert Arbeitsgruppe dar (Lambert, Hansen, & Finch, 2001): Aufgrund fehlender Prädiktoren (es wurde nur der Outcome Questionnaire erhoben; Lambert et al., 1996) teilten die Autoren eine große Stichprobe von Patienten aufgrund ihrer erreichten Aufnahmewerte in 50 Scoregruppen (Quantile; jeweils ca. 220 Patienten). Innerhalb dieser Scoregruppen wurden dann unkonditionale Wachstumsmodelle bestimmt und diese als Verlaufsvorhersagen verwendet.

Aufgrund dieser Modellergebnisse wurden dann ebenfalls Konfidenzintervalle berechnet, um Patienten identifizieren zu können, die die vorhergesagte Entwicklungskurve verließen.

In einer ersten Anwendung, die eine multidimensionale Definition von "Nachbarschaft" nötig machte, wurden Instrumente zur Erfassung symptomatischer Belastung, interpersoneller Probleme sowie demographische Information verwendet und es konnte gezeigt werden, dass Nächste Nachbarn Modelle den konditionalen Wachstumsmodelle überlegen waren (höhere Korrelationen zwischen beobachteten und vorhergesagten Verläufen; geringere standardisierte Abweichungen; Lutz et al., 2005). Es konnte ebenfalls repliziert werden, dass die Erweiterung der Vorhersagen um eine adaptive Komponente nach einer gewissen Therapiedauer eine sinnvolle Ergänzung war (Lutz et al., 2005). Lutz und Kollegen (Lutz, Tholen, Kosfelder, Grawe, & Schulte, 2005) zeigten, wie Therapiemodalitäten mit in die Vorhersage einbezogen werden können (siehe auch: Lutz, Saunders, et al., 2006) und nutzten sequenzanalytische Methoden, um nachzuweisen, dass die Modellvorhersagen besser als Zufall waren. In einer weiteren Studie (Lutz et al., 2006) wurde die prädiktive Validität der Einstufungen zu frühen Therapiezeitpunkten geprüft: Es zeigte sich, dass sowohl die Einstufung als positiver als auch diejenige als negativer Verlauf zu einem frühen Zeitpunkt in der Therapie mit dem Therapieergebnis zu einem späteren Zeitpunkt in der Therapie korrelierte.

Es liegen mehrere Vergleiche zwischen rationalen und empirischen Entscheidungsregeln vor, wobei bislang keine der Methoden klar besser abschneidet. Es sind eher Fragen der Umsetzung und der Ziele der jeweiligen Anwendung (Lutz, Stulz, et al., 2009), die über die Angemessenheit des jeweiligen Ansatzes entscheiden. Eine Möglichkeit ist die Verbindung beider Bewertungsansätze (Lueger et al., 2001). In einem Vergleich zwischen empirischen und rationalen Entscheidungsregeln schnitten beide Methoden gut ab (Lambert et al., 2002): ETR war etwas besser in der Prädiktion des tatsächlichen Verlaufes und die rationalen Kriterien waren etwas sensitiver für die Entdeckung von Patienten mit einem Risiko der Verschlechterung. In anderen Studien (Lutz, Lambert, et al., 2006; Lutz, Saunders, et al., 2006) zeigten sich die empirischen Entscheidungsregeln überlegen, bei denen aber berücksichtigt werden muss, dass sie deutlich aufwändiger zu implementieren sind (Lutz, Stulz, et al., 2009). Beide Systeme eigenen sich für die Feedback-Erstellung (s.

1.4.2) und Unterschiede in der Wirkung des generierten Feedbacks sind bislang nicht untersucht worden[10].

Eine weitere Vorgehensweise der empirischen Entscheidungsregeln zur Bewertung von Psychotherapieverläufen ist die Suche nach unterschiedlichen Verlaufsmustern. Bei diesem Vorgehen wird die Annahme der "expected treatment response" und Nächste-Nachbarn Modelle aufgehoben, dass es eine einzige Verlaufskurve in der Stichprobe gibt. Stattdessen werden die Patienten aufgrund ihrer Verlaufskurven in Gruppen mit ähnlichen Veränderungsmustern sortiert (Growth Mixture Models; B. O. Muthén, 2001, 2004; Nagin, 1999). Mittels dieser Modelle lassen sich die Verläufe von Patienten daraufhin untersuchen, ob sie sich in Gruppen einteilen lassen und diese entweder mit Therapieeingangsvariablen oder aber dem Therapieergebnis zusammenhängen.

Lutz und Kollegen (Lutz, Stulz, et al., 2007) untersuchten einen Datensatz aus der Routineversorgung (USA) auf unterschiedliche Verlaufsmuster in den ersten fünf Sitzungen der Therapie. Sie teilten die Patienten aufgrund ihres Belastungsgrades zu Beginn der Therapie in drei Gruppen ein (gemessen im OQ-30; leicht, mittel, hoch) und identifizierten innerhalb dieser drei Gruppen jeweils vier Verlaufsklassen. Während die Klassen nicht konsistent vorhergesagt werden konnten, zeigten sie aber deutliche Zusammenhänge mit dem Therapieergebnis (und leichte mit der Therapiedauer). Stulz und Kollegen (Stulz, Lutz, et al., 2007) konnten in einer anderen Studie (Routineversorgung in Großbritannien, erhoben mit dem CORE-SF, erste sechs Sitzungen; Barkham et al., 1998; Cahill et al., 2006) ebenfalls unterschiedliche Verlaufsklassen identifizieren. Darüber hinaus konnten sie auch noch auf zwei weitere Phänomene hinweisen. Zum einen ließ sich in dem Datensatz deutlich ein differentielles Ansprechen auf die Therapie feststellen, mit einer Gruppe, die bereits in diesen ersten Sitzungen eine deutliche Besserung zeigte ("early improvement") und mehreren Gruppen, die sich in dieser Zeit nicht änderten. Dieser Befund verdeutlichte in den Augen der Autoren die Bedeutung der Forschung zu frühen Veränderung in der Therapie (Haas, Hill, Lambert, & Morrell, 2002; Ilardi & Craighead, 1994; Lambert, 2007). Zum anderen konnten zwei Klassen mit nahezu denselben Verlaufsprofilen identifiziert werden, deren beobachtete Variation um die Verlaufskur-

[10] Für eine kritische Diskussion siehe Percevic, Lambert, & Kordy (2006).

ven jedoch deutlich unterschiedlich war. Die Autoren werteten dies als einen Hinweis auf diskontinuierliche Therapieverläufe (Lutz & Tschitsaz, 2007; Thompson, Thompson, Gallagher-Thompson, & Alto, 1995; Tschitsaz-Stucki & Lutz, 2009). Da diese Forschungsrichtung nicht im Fokus der Arbeit liegt würde eine detaillierte Beschreibung dieser Ansätze zu weit führen. Daher sei an dieser Stelle nur verwiesen auf: Colder, Campbell, Ruel, Richardson, & Flay, 2002; Hunter, Muthén, Cook, & Leuchter, 2010; Lutz, Stulz, & Köck, 2009; B. O. Muthén & Brown, 2009; Stulz, Gallop, Lutz, Wrenn, & Crits-Christoph, 2010; Stulz & Lutz, 2007; Uher et al., 2010.

Die vorangegangene Darstellung unterstreicht, dass an alle Vorgehensweisen, die Informationen für den Therapieprozess bereitstellen (Diagnostik, Qualitätssicherung, Feedback), ein hoher Anspruch an die Qualität der Primärdaten gestellt wird. Zentral ist die Inhaltsvalidität betroffen, da die erhobenen Instrumente über Veränderungen im psychotherapeutischen Prozess informieren sollen. Über die Kriterien, die diesen angemessen widerspiegeln, besteht kein Konsens (Hersen, 2004; Strupp, Horowitz, & Lambert, 1997); siehe auch Kapitel 1.4.3), doch sind die vorgestellten Methoden flexibel genug, um mit Instrumenten unterschiedlichster therapeutischer Orientierungen verwendet zu werden (Bram, 2010; Fowler et al., 2004; Hersen, 2004; Howard et al., 1996; Lueger et al., 2001; Waldron et al., 2011). Aus psychometrischer Sicht sind niedrige Standardmessfehler wünschenswert, da sie die Anwendung von Entscheidungsregeln trennschärfer machen. Bei der Anwendung von rationalen wie empirischen Entscheidungsregeln lassen sich die Veränderungen im Instrument und die tatsächliche Veränderung in der Therapie besser aufeinander beziehen. Je messgenauer die Instrumente sind, desto eher sind mit den beschriebenen Methoden bedeutungsvolle Verläufe identifizierbar. Bei den empirischen Entscheidungsregeln spielt die Messgenauigkeit auch eine entscheidende Rolle auf der Seite der Prädiktoren: Je geringer der Messfehler ist, desto eher beziehen sich die Vorhersagen auf klar definierte und replizierbare Subpopulationen.

1.4.2. Der Nutzen von Rückmeldesystemen: Anwendungen der Patientenorientierten Versorgungsforschung

Zeitintensive, wiederholte Erhebungen von diagnostischen Instrumenten (wie z.B. Fragebögen) müssen mehr als reiner Selbstzweck sein, da z.B. die Erhebungen Patienten belasten und Daten von ihnen gespeichert werden. Psychologen sollen im besten Interesse der ihnen Anvertrauten arbeiten

(American Psychological Association, 2002; C. Miller & Evans, 2004), daher muss also gerechtfertigt werden, warum ein bestimmtes Erhebungsvorgehen notwendig ist (Meyer et al., 2001). Ein Ansatzpunkt für diese Rechtfertigung ist die Perspektivdivergenz innerhalb des therapeutischen Settings, die notwendigerweise zwischen Patient und Therapeut bestehen muss. In einer Meta-Analyse wurde untersucht, ob es systematische Unterschiede zwischen von Patienten und von Therapeuten bewerteten Symptombelastungen gibt (Cuijpers, Li, Hofmann, & Andersson, 2010). In $N = 48$ Studien zur Therapie von Depression zeigten sie, dass die durch Therapeuten bewertete Veränderung in der indirekten Veränderungsmessung etwa ein $g = .88$ betrug und die aus dem Selbstreport $g = .67$. Die Autoren nennen zwei Interpretationen des Befundes: Entweder sind vom Patienten bewertete Therapieergebnisse konservativer oder aber die Kliniker sind besser darin, Veränderungen festzustellen. Eine andere Interpretation ist, dass Kliniker über-optimistisch sind, was zu den Befunden passt, dass Kliniker nicht so gut darin sind, Verschlechterungen bei ihren Patienten zu erkennen (Lambert, 2007; und 1.4.4).

Der Befund unterstreicht, dass die Wahrnehmung von Veränderungen in der Therapie unterschiedlich sein kann und dies ein Grund für die Verwendung von wiederholten Erhebungen in der Therapie ist. Aus der Sicht der evidenzbasierten Erhebungen kann dies bereits als ein Beleg gesehen werden, dass die Hinzunahme weiterer Perspektiven auf den Behandlungsprozess eine relevante Information für die Diagnostik darstellt (Hunsley & Mash, 2005). Meyer und Kollegen (2001) demonstrieren in ihrer Übersicht, dass der Psychologie viele valide Erhebungsinstrumente zur Verfügung stehen, doch merken sie an, dass ein Nachweis in der Breite fehlt, dass diese Erhebungen einen zusätzlichen Nutzen über die Diagnostik durch den Kliniker hinaus haben. Das Forschungsfeld zur Wirkung von systematischen Rückmeldungen in der Psychotherapie versucht genau diese Lücke in der Patientenorientierten Versorgungsforschung zu schließen.

Die Forschung zur Wirkung von Rückmeldungen (und damit der inkrementellen Validität zusätzlicher Erhebungen über das Therapeutenurteil hinaus) in der Psychotherapie wurde insbesondere durch die Gruppe von Michael J. Lambert vorangetrieben. In diesen Studien wurden die beschriebenen Methoden (s. 1.4.1) benutzt, um Patienten zu identifizieren, die drohten, sich negativ zu entwickeln. Vorher wurden bereits solche Maßnahmen zur Qualitätssicherung eingesetzt

(Barkham et al., 2001; Beutler, 2001; Kordy, Hannöver, & Richard, 2001; Lutz, 1997), aber bis zu dem Beginn ihrer Forschung war noch nicht systematisch aufgezeigt worden, ob solche Maßnahmen überhaupt einen Effekt auf das Therapieergebnis haben (Lambert et al., 2001)[11]. In ihrer ersten kontrollierten Studie zu der Fragestellung, ob sich die Therapieeffekte verbessern lassen, wenn Feedbacksysteme genutzt werden, nutzten die Autoren ein System mit rationalen Entscheidungsregeln (Lambert et al., 2001). Als Maß für die Entwicklung in der Psychotherapie wurde der Outcome Questionnaire (OQ)-45 verwendet (Lambert et al., 1996): Basierend auf einer Normierungsstichprobe wurde in Anlehnung an die Methode von Jacobson und Truax (1991) die nötige Differenz errechnet, die eine reliable Veränderung bedeutete (in diesem Fall 14 Scorepunkte). Außerdem wurde bestimmt, ab welchem Scorewert es für eine Person wahrscheinlicher ist aus der dysfunktionalen (hohe OQ-Werte) als der funktionalen Stichprobe (niedrigere OQ-Werte) zu stammen. Dieser Wert ("Cut Off") lag bei 64 Scorepunkten, d.h. ab einem Wert von 63 oder darunter war es für einen Patienten wahrscheinlicher aus der funktionalen Stichprobe zu stammen.

Das Spektrum der Anfangswerte im OQ wurde nun zusammen mit den Veränderungswerten zu der aktuellen Sitzung kodiert. Für einen Patienten, der beispielsweise mit einem OQ-Wert von 90 im stark belasteten Bereich oberhalb des Cut Offs die Therapie begann, gab es ein Warnsignal (rot auf dem Feedback-Report), wenn er sich um mehr als 5 Punkte verschlechterte; ein abgeschwächtes Warnsignal (gelb) gab es, wenn er sich um weniger als 5 Punkte verschlechterte, sich aber auch nicht um mehr als 7 Punkte verbesserte. Wenn er sich zwischen 7 und 25 Punkten verbesserte, gab es ein Signal für guten Fortschritt (grün) und bei einer noch stärkeren Verbesserung wurde er als im funktionalen Bereich (weiß) markiert. Das Signal wurde auch in verbale Rückmeldungen übersetzt, z.B. für das grüne Feedback (Lambert et al., 2001, p. 55): "The rate of change the client is making is in the adequate range. No change in the treatment plan is recommended."[12]

[11] Zu Theorien der Nutzung von Feedback und einem Überblick unterschiedlicher Feedback Systeme sei auf *Journal of Clinical Psychology/In Session 61*(2), 2005 und dort auf Claiborn & Goodyear (2005), Sapyta, Riemer, & Bickman (2005) sowie zusätzlich die Publikation Lambert, Whipple, et al. (2001) verwiesen.

[12] "Die Veränderungsrate des Patienten ist im angemessenen Bereich. Es wird keine Veränderung des Behandlungsplanes empfohlen." Übers. durch Autor

Die Forscher untersuchten nun, ob die Warnmeldungen (gelb und rot) tatsächlich Patienten mit negativen Aussichten identifizierten und sich durch die Rückmeldung an die Therapeuten eine Veränderung des Therapieeffektes einstellten. Dazu wurden die Patienten in zwei Gruppen aufgeteilt: In beiden Gruppen wurde der OQ erhoben, in der einen erhielten die Therapeuten das Feedback und in der anderen nicht. Die Ergebnisse zeigten einen deutlichen Unterschied zwischen den Entwicklungen der Patienten in der Feedbackgruppe und der Nicht-Feedback-Gruppe: Patienten, für die es zu einem Zeitpunkt der Therapie eine Warnmeldung gab und diese als Report auch an den Therapeuten ging, schnitten deutlich besser ab als diejenigen, für die es eine Warnmeldung gegeben hätte, deren Therapeuten aber kein Feedback erhielten (Effektstärke: $d = .44$).

Dieser Effekt ist mittlerweile in verschiedenen Studien belegt. Die Arbeiten der Gruppe von Michael J. Lambert finden in ihrem Setting den Effekt zuverlässig (Lambert, 2005, 2007; Lambert et al., 2003; Sapyta, Riemer, & Bickman, 2005; Shimokawa, Lambert, & Smart, 2010), mit dem System der Gruppe um Duncan und Miller werden ähnliche Befunde erzielt ("The Partners for Change Outcome Management System"; Lambert & Shimokawa, 2011; Miller, Duncan, Sorrell, & Brown, 2005). Und eine Meta-Analyse über Feedback als therapeutisches Mittel im Allgemeinen belegt diesen Effekt darüber hinaus (Hanson & Poston, 2011; Lilienfeld, Garb, & Wood, 2011; Poston & Hanson, 2010). Durch die verwendeten Designs (randomisiert kontrollierte Studien mit Messwiederholung und ohne Rückmeldungen in einer Bedingung) kann weitestgehend ausgeschlossen werden, dass der gefundene Effekt nur auf die wiederholte Erhebung zurückzuführen ist, und auch eine systematische Untersuchung von Wiederholungseffekten zeigte, dass dieser Effekt zwar vorhanden ist, aber nur sehr gering (Durham et al., 2002). Rückmeldungen zeigen ebenfalls Effekte in der Paartherapie (Anker, Duncan, & Sparks, 2009; Reese, Toland, Slone, & Norsworthy, 2010), in Kinder- und Jugendsettings (Bickman, Kelley, Breda, De Andrade, & Riemer, 2011) und mehrere Studien konnten zeigen, dass die gleichzeitige Bereitstellung von klinischen Entscheidungsbäumen und Handlungshinweisen ("clinical support tools") für die sich negativ entwickelnden Patienten noch stärkere positive Therapieeffekte erzielte (Harmon et al., 2007; Lambert & Cattani, 2012; Whipple et al., 2003), aber nicht für die Patienten, die sich gemäß der Vorhersagen entwickelten (Washington, 2010). Insgesamt liegt so viel Evidenz zu diesem Thema vor, dass die

Presidential Task Force on Evidence-Based Practice der APA den Bereich der Feedbackforschung auf die Liste der relevanten Zukunftsentwicklungen aufnahm (APA Presidential Task Force on Evidence-Based Practice, 2006, S. 275, S. 278).

Zwei aktuelle Studien weisen den Weg für die Untersuchung eines weiteren Faktors, der auch bei dem Feedback-Effekt eine Rolle spielt: Der Therapeut. Simon und Kollegen (Simon, Lambert, Harris, Busath, & Vazquez, 2012) konnten zeigen, dass es auch hier Variabilität zwischen den Therapeuten gibt, wie stark die Patienten von dem Feedback profitieren. De Jong und Kollegen (De Jong, Van Sluis, Nugter, Heiser, & Spinhoven, 2012) zeigen, dass die Stärke des Effektes von der Nutzung des Feedbacks wie auch mehreren Therapeutenvariablen abhängt (Selbstwirksamkeit und Wille, das Feedback zu nutzen).

Im deutschen Sprach- und Versorgungsraum stehen größere Studien zu diesem Thema noch aus. Es liegt lediglich eine Replikation aus dem stationären Setting vor (Berking, Orth, & Lutz, 2006). Puschner und Kollegen (Puschner, Schöfer, Knaup, & Becker, 2009) konnten im stationären Setting dagegen keine veränderte Behandlungspraxis oder veränderte Effekte feststellen. Im Rahmen des Modellvorhabens der Techniker Krankenkasse (Lutz, Böhnke, Köck, et al., 2011; Wittmann et al., 2011) konnte zwar Evidenz gesammelt werden, dass die dort eingesetzten Entscheidungsregeln sich negativ entwickelnde Fälle identifizieren, doch es fehlte eine Kontrollgruppe für den Beleg eines Feedback Effektes für die sich negativ entwickelnden Patienten.

In medizinischen Bereichen wird eine ähnliche Debatte geführt. Sie bezieht sich hier auf die Erhebung sog. "Patient reported outcomes" (PRO), weit definiert als vom Patienten selbst berichtete Aspekte der Gesundheit (Food and Drug Administration, 2006; Valderas et al., 2008a). Die Wichtigkeit der Erhebung von Patientenangaben und Informationen zur Prozessteuerung in der Medizin wurde von Ellwood formuliert (Ellwood, 1988; s.a. Knaup, Koesters, Schoefer, Becker, & Puschner, 2009) und wurde spätestens mit der Stellungnahme der Food and Drug Administration zu diesem Thema zu einem festen Bestandteil, da in ihr PROs auch als valide Ergebnismessungen in klinischen Studien festgehalten werden, die zur Dokumentation des Effektes einer Intervention genutzt werden können (Food and Drug Administration, 2006; Revicki, Gnanasakthy, & Weinfurt,

2007; Willke, Burke, & Erickson, 2004). Die Messung der PROs wird oft als Mittel der Prozesssteuerung aufgefasst und daher wird in der Regel untersucht, ob die Erhebung und Rückmeldung dieser Ergebnisse insgesamt über alle Patienten hinweg einen Effekt auch auf andere Merkmale als das Therapieergebnis hat. Dies ist ein anderer Fokus als der der Feedbackforschung, bei der es besonders um die Förderung einer spezifischen Patientengruppe geht.

Die Befundlage zu diesem Aspekt kann eher als durchmischt bezeichnet werden und steht damit nicht im Widerspruch zu den Befunden zur Steigerung der Therapieeffektivität für sich negativ entwickelnde Patienten. In einem Literaturüberblick, der sich mit den relevanten Dimensionen von PROs beschäftigt, wird über drei systematische Arbeiten hinweg zusammengefasst, dass die Behandlungsentscheidungen von Medizinern durch systematische Rückmeldungen der Lebenszufriedenheit und –qualität ("Quality of Life") nicht beeinflusst werden (Fung & Hays, 2008). Valderas und Kollegen (Valderas, Kotzeva, Espallargues, Guyatt, Ferrans, Halyard, Revicki, Symonds, Parada, & Alonso, 2008a, 2008b) berichten zusammenfassend über 28 randomisiert kontrollierte Studien, dass es Effekte von generellem Feedback auf Prozessmerkmale gibt, aber seltener auf tatsächlich mit dem Gesundheitsstatus verbundene Indikatoren. Dabei ist einschränkend zu berücksichtigen, dass 1.) über die Studien keine konsistenten Effekte über verschiedene Prozessmaße gefunden wurden und auch nicht für die Kumulierung des α-Fehlers auf Studienebene kontrolliert wurde; und 2.) in nur einem Teil der Studien tatsächlich auch gesundheitsrelevante PROs untersucht wurden. Knaup und Kollegen (2009) führten eine Meta-Analyse mit 12 Studien durch, die den Effekt von Feedback für die psychische Gesundheitsversorgung ("mental health") untersuchten. Sie fanden einen kleinen Effekt für Feedback an alle Patienten von Hedges $g = .1$, doch die Tatsache, dass lediglich 11 Studien ohne Effekt nötig wären, diesen Effekt statistisch nichtsignifikant werden zu lassen, zeigt, dass dieser positive Befund nicht als robust angesehen werden kann. Im bislang umfangreichsten Literaturüberblick (Carlier et al., 2012) fanden die Autoren in der Mehrheit von 52 kontrolliert randomisierten Studien hypothesenkonforme Effekte, integrierten diese aber nicht zu einer Gesamteffektstärke.

Es gibt also zwei Gründe, die Verwendung von Fragebögen als "evidenzbasiert" zu bezeichnen: Den Feedback-Effekt für sich negativ entwickelnde Patienten sowie eine mögliche Erhöhung

der Prozessqualität. Die Instrumente haben das Potential Informationen über die reine Beobachtung durch den Kliniker hinaus bereitzustellen (Meyer et al., 2001). Die Verwendung der Fragebögen vor dem Hintergrund der oben beschriebenen Methoden, die die Grundlagen für solche Rückmeldungen darstellen, zeigt, wie wichtig eine hohe Qualität der verwendeten Instrumente ist. In manchen Anwendungen wird lediglich ein Instrument verwendet, um das Feedback zu erstellen (bei Duncan und Miller zwei Skalen, aber lediglich acht Items; Miller, Duncan, Brown, Sorrell, & Chalk, 2006; Miller, Duncan, Sorrell, & Brown, 2005). Solche Instrumente müssen eine hohe Inhalts- und Konstruktvalidität aufweisen, ohne die sie nicht verlässlich genutzt werden können. Zusätzlich ist aus Sicht der evidenzbasierten Erhebungen anzumerken, dass letztlich für jedes spezifische Instrument bzw. jede Feedbackanwendung gezeigt werden muss, dass es die dargestellten Vorteile erreicht (Hunsley & Mash, 2005).

Damit ein Instrument als evidenzbasiert gelten kann, muss auch der Mechanismus geklärt werden, wie die Fragebogenergebnisse auf die Prozess- und Ergebnisqualität wirken. Die Frage, was Therapeuten mit Rückmeldungen machen, ist nur kursorisch beforscht worden (Lambert, 2007). De Jong und Kollegen (De Jong et al., 2012) fanden in ihrer Untersuchung mit einer Versorgungsstichprobe in den Niederlanden keinen allgemeinen Effekt für das Feedback und konnten auch den Befund zu sich negativ entwickelnden Patienten nicht replizieren. Ihre Studie zeigte stattdessen, dass nur bei Therapeuten, die angaben, das Feedback zu nutzen, dieses auch eine positive Wirkung auf den Verlauf der Patienten hatte. Dies belegt, dass nicht die Gabe, sondern die Rezeption des Feedbacks ein wesentlicher Teil der Behandlung ist. Hatfield und Kollegen (D. Hatfield, McCullough, Frantz, & Krieger, 2010) zeigten an einer kleinen Stichprobe über die Adressbasis der American Psychological Association gezogener Therapeuten ($N = 40$), dass diese Verschlechterung bei ihren Patienten nicht über die Verwendung von Erhebungsinstrumenten feststellten (N = 4 Nennungen; am Häufigsten: verbaler Selbstbericht des Patienten, N = 17). Wenn eine Verschlechterung festgestellt wurde, war die am Häufigsten genannte Kategorie die Übersendung an medizinischen Kollegen ($N = 20$) gefolgt von mehr Sitzungen ($N = 12$) und der Suche nach weiterer Information ($N = 12$). Lutz (1997) stellte in einer Untersuchung im stationären Setting mit Dokumentation und Feedback fest, dass über die Hälfte der befragten Therapeuten zwar bejahte, Verän-

derungen eingeleitet oder geplant zu haben (aufgrund der Rückmeldungen; 55.9%), welcher Art diese Veränderungen waren, wurde aber nicht genauer erfasst. Patel und Riley (2007) berichten in einer qualitativen Folgestudie (N = 37) zu der Entwicklung eines internetbasierten Monitoring- und Rückmeldesystems für die Versorgung von Kindern und Jugendlichen, dass nur wenige Mitarbeiter das System regelmäßig zur Unterstützung klinischer oder Management-Entscheidungen nutzten. In der Evaluation des Modellvorhabens Psychotherapie der Techniker Krankenkasse ergab sich, dass nahezu die Hälfte der Therapeuten die Ergebnisse der Erhebungen mit den Patienten besprach; alle anderen abgefragten Verhaltensweisen traten in weniger als einem Drittel der Fälle auf (z.B. Ressourcenförderung, Anpassung der Intervention; Lutz, Böhnke, Köck, et al., 2011). Die Anpassungen unterschieden sich kaum zwischen den Patientenverläufen: Nur von der Tendenz her ergab sich, dass Therapeuten die Verläufe mit ihren Patienten eher besprachen oder eher ihre Intervention anpassten, wenn sich Stagnation oder Verschlechterung zeigten und eher die Allianz förderten, wenn die Patienten im Verlauf stagnierten.

Die Gründe von Therapeuten, Fragebogeninstrumente in der Therapie (nicht) zu nutzen, sind etwas besser beforscht als die Frage, wie Therapeuten das Feedback verwenden. Positive Gründe fanden sich beispielsweise in einer Befragung amerikanischer Psychotherapeuten (D. Hatfield & Ogles, 2007): Für die Nutzung bei Therapeuten sprach hauptsächlich die Möglichkeit, den therapeutischen Fortschritt festzuhalten oder aber zu beurteilen, ob Effekte der Therapie aufrecht erhalten wurden. Als Gründe für die Nicht-Nutzung von psychometrischen Instrumenten findet sich in verschiedenen Studien immer wieder, dass die Nutzung dieser Instrumente zu zeitintensiv sei; dass die zusätzliche Belastung der Patienten zu hoch sei und auch das generelle Gefühl, dass diese Instrumente nicht hilfreich seien (Hagemeister, Lang, & Kersting, 2010; D. Hatfield & Ogles, 2007; D. R. Hatfield & Ogles, 2004; Steck, 1997). Diese grundsätzlich skeptische Sicht könnte etwas gemildert werden, wenn Patientenorientierte Versorgungsforschung und Feedback als Teile eines umfangreicheren diagnostischen Vorgehens verstanden werden und dort bereits etablierte Konzepte genutzt werden, um die Methoden zu erweitern (Krampen & Hank, 2008; Lutz, Böhnke, Köck, et al., 2011; Lutz, 2002).

1.4.3. Patientenorientierte Versorgungsforschung als Diagnostik

Diagnostik dient dazu, ein möglichst breites Bild über den Patienten und seine Lebensumstände zu erstellen, um mit diesen Informationen Entscheidungen darüber zu treffen, wie unerwünschte Ausgangszustände mit Hilfe (psychologischer) Interventionen auf erwünschte Zielzustände hin verändert werden können (Maercker, 2011; Seidenstücker, 1995). Dabei lassen sich die folgenden diagnostischen Phasen unterscheiden (z.B. nach Stieglitz, 2003): Diagnostik zu Therapiebeginn, im Verlauf der Therapie, zum Ende der Therapie und katamnestische Erhebungen. Diese Phasen haben unterschiedliche Ziele und weisen bestimmte Funktionen im Rahmen des Therapieprozesses auf (für eine detailliertere Diskussion diagnostischer Vorgehensweisen und Konzepte: Grosse Holtforth, Lutz, & Egenolf, 2010; Lutz, Mocanu, & Weinmann-Lutz, 2010).

Die Diagnostik zu *Therapiebeginn* dient der Indikationsstellung, der Therapiezielbestimmung und der diagnostischen Einordnung. Zu Therapiebeginn steht die Erfassung der Symptome im Vordergrund (im Sinne störungs-, lebens-, therapierelevanter Merkmale; Grosse Holtforth et al., 2010), während in der therapiebegleitenden Diagnostik die Veränderung dieser Symptome und Prozessmerkmale im Fokus stehen. Erhebungen am Ende der Therapie dienen zur Evaluation des Therapieerfolgs und katamnestische Erhebungen zur Feststellung der Stabilität der erreichten Erfolge. Für die Patientenorientierte Versorgungsforschung ist die Feststellung des Status zu Beginn der Therapie ein relevanter Faktor (s.a. 1.4.1), da alle Bewertungsmethoden darauf beruhen, dass Veränderungen zum Eingangsstatus festgestellt werden.

Die therapie*begleitende* Diagnostik kann in zwei Unterkategorien aufgeteilt werden, die Prozessdiagnostik und die Verlaufsdiagnostik (Grosse Holtforth et al., 2010). Die Prozessdiagnostik erhebt Merkmale des therapeutischen Prozesses wie beispielsweise die Qualität der Interaktion zwischen Patient und Therapeut oder die Realisation von Wirkfaktoren. Da sich diese Merkmale im Verlauf ändern können oder auch in einzelnen Sitzungen unterschiedlich realisiert sein können, werden sie üblicherweise zu jeder Sitzung erhoben. Hierzu eigenen sich speziell entwickelte Stundenbögen (Flückiger, Regli, Zwahlen, Hostettler, & Caspar, 2010; Krampen, 2002).

Die Verlaufsdiagnostik erfasst Veränderungen in den Problem- und Störungsbereichen. Diese Informationen sollen aus zur Erstellung von Verlaufs- und Ergebnisprognosen herangezogen werden und den Therapeuten auf problematische Entwicklungen hinweisen (Grosse Holtforth et al., 2010; Lutz et al., 2010). Dies setzt ein Verständnis der Intervention voraus, bei dem sowohl diagnostische Bewertungen (inkl. der Diagnostik zu Therapiebeginn) als auch der daraus resultierende Behandlungsplan immer nur als Hypothesen angesehen werden. Mit dem Durchlaufen des therapeutischen Prozesses treten Veränderungen und neue Informationen auf, die ggf. eine Revision des bisherigen nötig machen (Maercker, 2011; Seidenstücker, 1995). Hier fügt sich die Patientenorientierte Versorgungsforschung in die diagnostische Praxis ein, indem sie durch die methodischen Mittel der Entscheidungsregeln und Verlaufsdokumentationen einen konzeptuellen Rahmen schafft, der diesen Teil der Diagnostik erfüllen kann (Lueger et al., 2001; Lutz, 2002; Lutz & Böhnke, 2010; Lutz et al., 2010). Andere konzeptuelle Rahmen, in die sich diese Auffassung der Patientenorientierten Versorgungsforschung auf eine ähnliche Weise einfügt, sind das Konzept der "Kontrollierten Praxis" (Petermann & Müller, 2001) und der "Kliniker als Wissenschaftler" (G. Stricker, 2006). Newnham und Page (2010) schlagen ein Modell therapeutisch-diagnostischen Vorgehens vor, das sich aus den Elementen a) wiederholter Erhebungen, b) Testung diagnostischer Hypothesen und c) Reformulierungen des Geschehens der konkreten Therapie enthält.

Diese Festlegung, den Prozess der Therapie entweder aus diagnostischen Gründen oder aus Sicht der der Patientenorientierten Versorgungsforschung heraus um wiederholte Erhebungen zu erweitern, legt nicht fest, welche Elemente im Verlauf erhoben werden sollen. Die Festlegung von Evaluationskriterien ist aber wichtiger Bestandteil jedes Erhebungssystems, da es sonst nicht möglich ist, informierte Entscheidungen zu treffen (Westmeyer, 1979; s.a. Abschnitt zu Entscheidungsregeln 1.4.1). Bislang ist nicht geklärt, welche Kriterien des Therapieerfolges relevant sind (De Los Reyes & Kazdin, 2006; De Los Reyes, Kundey, & Wang, 2011; Hersen, 2004; Lambert & Ogles, 2004; Sonnanburg, 1996; Strupp et al., 1997; Strupp, 1963). Auch die Patientenorientierte Versorgungsforschung gibt hier keine Lösungen vor, außer dass veränderbare Merkmale, die in der jeweiligen Therapie als relevante Ergebnisse angesehen werden, dokumentiert werden sollen (Howard et al., 1996; Lueger et al., 2001; Lutz, 2002).

Dennoch gibt es verschiedene Konzeptionen, die aufgegriffen werden können. In den Anwendungen der Arbeitsgruppe von Michael Lambert steht lediglich ein einziges Instrument im Mittelpunkt der Systeme, das ein breites Spektrum an Symptomen und Beschwerden abfragen soll (zur Diskussion z.B. Lambert, Whipple, et al., 2001). In einer solchen Anwendung wird der Fragebogen als ein Indikator des allgemeinen psychischen Zustandes gesehen und soll einen Eindruck vermitteln, wie es dem Patienten insgesamt geht. Außerdem trägt eine solche Herangehensweise der Tatsache Rechnung, dass in den meisten Fragebogeninstrumenten keine klare Trennung zwischen verschiedenen Dimensionen vorgenommen werden kann, da sie sich zumindest in der empirischen Bewertung mit faktoranalytischen Techniken oftmals nicht findet (Böhnke & Lutz, 2011; Connell et al., 2007; Halstead, Leach, & Rust, 2007; Lambert et al., 1996; Lutz & Böhnke, 2008; Lutz et al., 2009; Schürch, Lutz, & Böhnke, 2009)[13].

Diagnostik sollte aber als multimodaler Prozess (Seidenstücker & Baumann, 1987) aufgefasst werden, der unterschiedliche Datenebenen (z.B. biologisch/ somatisch, psychisch, sozial, ökologisch), Datenquellen (Selbst-/ Fremdberichte), Untersuchungsmethoden (z.B. Fragebogen, Interview, Verhaltensprobe) und Konstrukte bzw. Funktionsbereiche (z.B. Symptome, Ressourcen) erfasst (s.a.: Grosse Holtforth et al., 2010). Zwei bestehende Konzeptionen machen zumindest Vorschläge, welche Funktionsbereichen im Rahmen der therapiebegleitenden Diagnostik erfasst werden sollten. Eine der Positionen ist das aus dem Dosis-Wirkungs-Modell der Psychotherapie (Howard et al., 1986) hervorgegangene Phasenmodell der psychotherapeutischen Veränderung (Howard et al., 1993). Ein anderer Entwurf findet sich bei Schulte (1993).

Im Phasenmodell werden drei Phasen des Therapiefortschrittes identifiziert, in denen jeweils andere Dimensionen der Veränderung im Vordergrund stehen. Zu Anfang der Therapie steht der Aspekt der Remoralisierung der Patienten im Vordergrund (Frank & Frank, 1991). Die Erfahrung der Patienten, durch ihre Lebenssituation stark belastet zu sein und u.U. die Wünsche ihrer Umwelt an sie nicht befriedigen können, lässt sie in einem Stadium der Hilf- und Hoffnungslosigkeit zu-

[13] Stiles, Shapiro und Elliott (1986) beschreiben als Grund dafür, dass keine Unterschiede in der Effektivität verschiedener Therapien gefunden wurden, dass die Ergebnismaße nicht spezifisch genug waren. Dieses Argument kann auch hier genutzt werden: Die Ergebnismaße sind immer noch zu global, um unterschiedliche Dimensionen der therapeutischen Veränderung feststellen zu können.

rück, das am Anfang beseitigt wird; ein Prozess, der schon vor der Therapie beginnen kann (Lueger, 1995). Der Effekt der Remoralisierung zeigt sich besonders auf der Dimension allgemeinen Wohlbefindens ("subjective well-being"; Howard et al., 1993). Danach folgt die Phase, die die Autoren Remediation nennen, in der innerhalb der Therapie an den Symptomen gearbeitet wird. Neben störungs-/ problemspezifischen Interventionen stehen hier auch solche Interventionen im Mittelpunkt, die zu einer Verstärkung der Ressourcen der Patienten sowie ihrer bereits vorhandenen positiven Coping-Strategien beitragen. In dieser Phase sollte sich besonders eine Veränderung im Bereich der allgemeinen Beschwerden und Störungssymptome abzeichnen (Howard et al., 1993). In der letzten Phase, der der Rehabilitation, hat der Patient dann Kraft, nachdem an konkreten Belastungen gearbeitet wurde, sich allgemeineren Fragestellungen und Veränderungen in Bezug auf sein Lebensumfeld zu stellen. Das "psychosoziale Funktionieren" ("life functioning"; Howard et al., 1993) ist in dem mit der Theorie verbundenen Erhebungssystem (COMPASS; Howard et al., 1993; Grissom, Lyons, & Lutz, 2002) besonders auf die Rollenerfüllung in sozialen Kontexten abgezielt, doch gibt es auch z.B. Umsetzungen mit den interpersonellen Problembelastungen der Patienten (Grosse Holtforth et al., 2010; Lutz, 2010; Lutz et al., 2009).

Diese drei Dimensionen stellen eine Möglichkeit dar, ein Monitoring System auf eine breitere Basis zu stellen als lediglich die allgemeine psychologische Belastung. Auch diese drei Dimensionen geben noch keinen schulenspezifischen Rahmen vor, was zu erheben ist. Die Dimensionen können immer noch unterschiedlich operationalisiert werden (Fowler et al., 2004; Howard et al., 1996; Lueger et al., 2001). Doch gehen mit diesem Modell sowohl eine Theorie über die Verlaufserwartung einher (Verbesserungen in den drei Dimensionen sind stochastisch von einander abhängig) und es werden drei Bereiche definiert, die sowohl klinisch relevante Informationen liefern (theoretische Einbettung: Howard et al., 1993) als auch in ihrer Sensitivität für Veränderung bestätigt sind (zur empirischen Bestätigung des Modells siehe Kopta, Howard, Lowry, & Beutler, 1994; S. E. Stevens, Hynan, & Allen, 2000; Stulz & Lutz, 2007)

Schulte (1993) schlug drei andere Bereiche als relevant für die Beurteilung des Therapieerfolgs vor. Der erste ist die Messung des Symptom- oder Beschwerderückganges; der zweite die Erhebung der der jeweiligen Therapietheorie angemessenen störungsspezifischen Ursachen; und

schließlich die Erhebung von Störungsfolgen, die z.B. durch die Übernahme der Krankheitsrolle entstehen oder andere aus der Störung resultierenden Einschränkungen. Auch diese drei Dimensionen können kontextspezifisch operationalisiert werden und stellen drei Bereiche dar, die eine psychotherapeutische Intervention im Verlauf sinnvoll informieren können (detailliertere Diskussion unterschiedlicher Dimensionen und Operationalisierungen bei Grosse Holtforth et al., 2010; Hersen, 2004; Lutz & Böhnke, 2010; Lutz, 2010; Newnham & Page, 2010; Stieglitz, 2003).

Im Sinne der multimodalen Diagnostik machen beide Positionen keine Einschränkungen, aus welcher Datenquelle oder mit welchen Erhebungsmethoden die therapiebegleitende Diagnostik vorgenommen wird. Die Anwendungen beruhen in der Regel auf Selbstberichten der Patienten, da diese wichtiges systematisches Feedback über die Entwicklung aus der subjektiven Sicht des Patienten enthalten (Ausnahme z.B. bei Kinder-/ Jugendpsychotherapie, z.B.: Bishop et al., 2005). Doch gerade im Rahmen der Prozessdiagnostik ist die systematische Erhebung von Patienten- und Therapeutenwahrnehmungen durchaus üblich (Flückiger et al., 2010; Grosse Holtforth et al., 2010). Die Aufnahme dieser unterschiedlichen Perspektiven ist auch durchaus sinnvoll, wie im Abschnitt zur Qualitätssicherung aufgezeigt wird (s. 1.4.4).

Andere Vorschläge nehmen ähnliche Grunddimensionen an, legen aber noch zusätzliche Betonungen auf andere Elemente. Ey und Hersen (2004) heben die Erhebung der nicht-spezifischen Wirkfaktoren hervor, dabei besonders die Therapeutischen Allianz, die Therapiemotivation und die Bereitschaft zur Veränderung auf Seiten der Patienten (siehe z.B. Basisverhalten bei Schulte & Eifert, 2002). Vissers (2010) vertritt den Standpunkt, dass Symptomreduktion nur bedingt als zentrale Ergebnisdimension gesehen werden kann, da Patienten dies nicht als zentral bewerteten. Auch kann dieses Ziel bei Patientengruppen als zu ambitioniert angesehen werden (Lutz & Grawe, 2007; Tingey et al., 1996). Diese von Vissers vorgetragenen Gründe werden im breiteren Kontext der Gesundheitsversorgung bereits länger in der Debatte um Patient-Reported Outcomes betont (s. Kapitel 1.4.2). Dem Konzept der "Lebenszufriedenheit/ -qualität" ("Quality of Life") kommt hier eine zentrale Bedeutung zu, da aus der Sicht der Patienten oft weniger zentral sei, ob sich die Symptome verändern, sondern, dass sich die wahrgenommene Lebenssituation und in die in ihr erlebte Qualität verbessern (Fung & Hays, 2008; Huppert & Whittington, 2003; Lauer, 1998).

Zusätzlich ist das Konstrukt der Lebensqualität über viele Gesundheitsbelastungen hinweg vergleichbar, auch wenn die Zielkriterien sich zwischen Störungen bzw. Erkrankungen sowie soziodemographischen Charakteristika unterscheiden (Huppert & Whittington, 2003; Nuevo et al., 2010; Walker, Böhnke, Cerny, & Strasser, 2010). Aus der Forschungsperspektive wird betont, dass es günstiger wäre, viele Variablen zu erheben, damit eine spätere differenzierte Beschreibung und der Vergleich von Patienten ermöglicht wird (Gilbody et al., 2002a). Dies ist sowohl aus der Sicht des eingangs beschriebenen Matrix-Paradigmas sinnvoll (Stiles et al., 1986), wie auch bei der gezielten Anpassung und Vorhersage für neue Patienten (Lutz et al., 1999; Lutz, Leach, et al., 2005). Allerdings stellt eine ungezielte Erhebung von Patientenmerkmalen eine Zusatzbelastung dar, die im Widerspruch zu ethischen Grundsätzen steht, wenn die inkrementelle Validität jedes einzelnen Instrumentes nicht klar nachgewiesen ist (Hunsley & Mash, 2005; Meyer et al., 2001). Als eine zusätzliche Dimension, die besonders vom Patienten zu erheben und erfassen ist, wird auch die Zufriedenheit mit der Qualität des Angebotenen Services als zentrales Maß angesprochen (Fung & Hays, 2008), die auch regelmäßig in der psychotherapeutischen Versorgung als eine relevante Dimension in der Beschreibung der Versorgung thematisiert wird (Gmür & Straus, 1998; Jacob & Bengel, 2003; Lutz, 1997; Wittmann et al., 2011).

Im Modellvorhaben der Techniker Krankenkasse (TK) wurde versucht, mehrere dieser Aspekte zu berücksichtigen (Fydrich, Nagel, Lutz, & Richter, 2003; Lutz, Tholen, Kosfelder, Tschitsaz, et al., 2005; Wittmann et al., 2011). Die Diagnostik zu Beginn der Therapie schloss sowohl die strukturierte Erhebung der Psychopathologie durch den Therapeuten ein (Internationale Diagnose Checklisten für DSM IV; Hiller, Zaudig, & Mombour, 2004) wie auch eine dreidimensionale Erhebung der Belastung zu Therapiebeginn im Selbstbericht durch die Patienten: Die allgemeine Belastung (Brief Symptom Inventory; Franke, 2000), die interpersonale Belastung (Inventory of Interpersonal Problems; Horowitz, Strauss, & Kordy, 2000) und die störungsspezifische Belastung durch einen für die Diagnosen festgelegten Selbstberichtsbogen (Köck, 2012; Lutz, Böhnke, Köck, et al., 2011; Wittmann et al., 2011). Eine direkte Bewilligung der Therapie fand dann statt, wenn aufgrund der Entscheidungsregeln (vgl. 1.4.1; Lutz, Tholen, Kosfelder, Grawe, et al., 2005) eine

relevante Belastung in mindestens einer der psychometrischen Erhebungsdimensionen festgestellt wurde, wie auch zusätzlich eine Diagnose durch den Therapeuten vergeben wurde.

Auch die Rückmeldungen im Verlauf der Therapien des TK Modellvorhabens waren multidimensional. Neben Rückmeldungen in Bezug auf die drei Selbstberichtsdimensionen erhielt der Therapeut zusätzlich standardmäßig Rückmeldung über die Therapeutische Arbeitsbeziehung (Helping Alliance Questionnaire; Bassler, Potratz, & Krauthauser, 1995) sowie die gesundheitsbezogene Lebensqualität (SF-12; Bullinger & Kirchberger, 1998). Die zusammenfassenden Einschätzungen des Therapiefortschrittes mit Referenz zum Eingangswert basierten auf Aggregationen der drei Selbstberichtsdimensionen, die eine insgesamt positive Einschätzung des bislang erreichten Fortschrittes widergaben, wenn mindestens zwei der drei Dimensionen Besserung zeigten, eine neutrale Einschätzung, wenn eine oder keine als gebessert eingestuft wurde, und schließlich eine Warnung, wenn mindestens eine der Dimensionen sich verschlechtert hatte (Details siehe: Lutz et al., 2009; Lutz, Tholen, Kosfelder, Tschitsaz, et al., 2005; Wittmann et al., 2011). Dies zeigt, wie Rückmeldungssysteme auch mit den rationalen Entscheidungsregeln zur multidimensionalen therapiebegleitenden Diagnostik erweitert werden können. Die Verbindung der einfachen Diagnostik mit psychometrischen Instrumenten mit den Modellen der Patientenorientierten Versorgungsforschung (im Beispiel speziell die rationalen Entscheidungsregeln) stellt eine sinnvolle Kombination dessen, was aus diagnostischer Perspektive sowieso gefordert ist (Grosse Holtforth et al., 2010; Lutz & Böhnke, 2010) und einem Element, das als evidenzbasierte Praxis bezeichnet werden kann, dar (Lutz, 2010; Newnham & Page, 2010).

Im Rahmen dieser Arbeit ist festzuhalten, dass immer dann, wenn Fragebögen zur Evaluation der verwendeten Dimensionen verwendet werden, die bereits mehrfach erwähnte hohe Messqualität eine Rolle spielt (s. 1.3). Therapiebegleitende Diagnostik mit Fragebögen (sei es aus Perspektive der Patientenorientierte Versorgungsforschung oder der Diagnostik; im Selbstbericht des Patienten oder einer anderen Perspektive) ist nur dann informativ, wenn die Instrumente hinreichend (konstrukt-)validiert sind. In dieser Arbeit geht es um die Verwendung von standardisierten Fragebögen insbesondere für die Erhebung von Selbstberichten. Standardisierte Fragebögen haben sich in der klinischen Praxis in verschiedenen Kontexten bewährt. Sie geben die Möglichkeit mit relativ ge-

ringem Aufwand auch umfangreichere Selbstberichte zu erheben und dem Kliniker zur Verfügung zu stellen (Fragebogen vorher ausfüllen statt Gespräch über alle abgefragten Symptome; schafft Raum in der Therapie). Sie ermöglichen ein Monitoring der Entwicklung des Patienten, lassen einordnen, wie der Beschwerdegrad im Verhältnis zu anderen Patienten-/ Normstichproben ist, und lassen sich für verschiedene Screening-Anwendungen einsetzen (Ey & Hersen, 2004; Lutz & Grawe, 2007; Lutz, Martinovich, et al., 2002).

1.4.4. Patientenorientierte Versorgungsforschung als Qualitätssicherung

Ein dritter Rahmen, in dem die Patientenorientierte Versorgungsforschung gesehen werden kann, ist das Gebiet der Qualitätssicherung bzw. des Qualitätsmanagements (Howard et al., 1996; Johnson & Shaha, 1996). Dies ist im deutschen Kontext wichtig, da laut Sozialgesetzbuch V eine Verpflichtung zur Qualitätssicherung und –entwicklung in allen Teilen des Gesundheitssystems besteht (z.B. Härter, Linster, & Stieglitz, 2003); Zitat aus SGB V, §135a, Abs. 1):

> Die Leistungserbringer sind zur Sicherung und Weiterentwicklung der Qualität der von ihnen erbrachten Leistungen verpflichtet. Die Leistungen müssen dem jeweiligen Stand der wissenschaftlichen Erkenntnisse entsprechen und in der fachlich gebotenen Qualität erbracht werden.[14]

Es existieren eine ganze Reihe von Versuchen, "Qualität" im Rahmen der Psychotherapie zu definieren (Laireiter & Vogel, 1998; Lutz, 1997). Das Deutsche Institut für Normung definiert Qualität danach, wie stark die Merkmale eines Produktes vordefinierte Anforderungen erfüllt (DIN EN ISO 9000; Piechotta, 2008). Diese Definition weist auf die Ursprünge der Sicherung der Qualität von Produkten und Dienstleistungen in der industriellen Fließbandarbeit in den USA hin. Die dort angestrebte Vermeidung von Fehlern führte zur genauen Prüfung von Produkten am Ende eines Herstellungsprozesses, der so genannten Qualitätskontrolle (Härter et al., 2003). Qualität im Gesundheitswesen ist jedoch weniger klar umrissen als in der Industrie. Die Joint Comission on Accreditation of Health Care Organizations (1996) hat den Fokus in ihrer Definition von Qualität auf das Therapieergebnis gelegt, indem sie Qualität als die Wahrscheinlichkeit unter Einbezug des

[14] Heute wird in der Regel von Qualitätsmanagement gesprochen, da es nicht nur um die Aufrechterhaltung eines Qualitätsstandards geht ("Sicherung"), sondern auch um die Verbesserung und Weiterentwicklung der Standards (Farin & Bengel, 2003; Härter, Linster, & Stieglitz, 2003).

derzeit vorherrschendem Wissens, ein für den Patienten erstrebenswertes Therapieresultat zu erzeugen, definierte (Härter et al., 2003).

Ein Vorteil dieser Definition ist, dass sie den Begriff "Gesundheit" außen vorlässt und ein stärker operational orientiertes Kriterium vorschlägt. Dass der Begriff "Gesundheit" schwierig ist, liegt an der Problematik psychische Gesundheit und Störung zu trennen (Krause & Lutz, 2009; Schwartz, Siegrist, von Troschke, & Schlaud, 2003; Wulff, 1988/2004) und daran, dass Therapeuten und Patienten eine aktive Rolle am Erreichen dieses Zielzustandes zukommt (Härter et al., 2003). Wie bereits aufgezeigt (1.1) ist diese Abgrenzung über das zu erzielende Ergebnis nicht sehr scharf: Vor dem Beginn einer Therapie können zur Erreichung des für den Patienten erstrebenswerten Therapieresultates nach dem derzeitigen Wissen therapeutische Interventionen im Sinne gesamter Therapieschulen gelten, wie auch bestimmte Techniken, wie auch unspezifische Wirkfaktoren. Um es mit Perrez (2005) zu fassen: Die therapeutische Handlungsregel, die evidenzbasiert erstellt werden soll, ist sehr unspezifisch. So wichtig diese Evidenz auch ist und sicher einen Rahmen von Strukturqualität schafft (z.B. durch die Prüfung, ob zugelassene Therapeuten diese Techniken beherrschen und anwenden können; zur Unterscheidung der Qualitätsebenen z.B. Böhnke & Lutz, 2011b; Donabedian, 2005), so wenig ist sie jedoch im konkreten Einzelfall hilfreich, da sich diese Art von Evidenz mit der Untersuchung von Gruppenmittelwerten beschäftigt (Krause et al., 2011; Lutz & Grawe, 2007; Persons & Silberschatz, 1998).

Diese Definition von Qualität verdeutlicht, dass in der Psychotherapie eine Kombination aus existierendem (Forschungs-)Wissen wie auch einer Orientierung am Einzelfall vorherrschen sollte (Howard et al., 1996; Lutz & Grawe, 2007; Schmitz, 2000). Wie beschrieben (1.2.3 und 1.4.1), stellt die Patientenorientierte Versorgungsforschung Elemente bereit, die diese Kombination ermöglichen. Das existierende Forschungswissen wird genutzt, um die Bewertungsregeln zu erstellen. Die Bewertungsregeln können in Verbindung mit jeder Art von Therapieprogramm verwendet werden, solange diese das Kriterium erfüllen, dass sie die Wahrscheinlichkeit auf ein gewünschtes Therapieergebnis erhöhen. Außerdem stellen sie gleichzeitig eine Überprüfung dar, ob dieses Qualitätskriterium gemäß Joint Comission on Accreditation of Health Care Organizations (1996) in der laufenden Therapie erfüllt wird. Und die Rückmeldung nicht erreichter Ziele kann die

Hypothesenhaftigkeit des diagnostischen Vorgehens unterstützen (Grosse Holtforth et al., 2010; Howard et al., 1996; Lambert, 2001; Lutz, 2002; Newnham & Page, 2010; Seidenstücker, 1995).

Die Patientenorientierte Versorgungsforschung als Element der Qualitätssicherung ermöglicht sich negativ entwickelnde Patienten (oder solche, die sich zumindest nicht auf das Therapieziel zubewegen) zu identifizieren. Schätzungen, wie viele Patienten dies betrifft, gehen auseinander, und liegen zwischen 5% und 25% (Lambert & Shimokawa, 2011; Lambert, 2007; Mohr, 1995). Daten für das deutsche Versorgungssystem sind z.B. 6.8% bis 8.4% Fälle mit einer Verstärkung der Problematik im TK Modellvorhaben (indirekte Verlaufsmessung mit drei-dimensionalem Ergebniskriterium; Wittmann et al., 2011: 135), 1.4% bis 4.3% für allgemeine Problembeschreibungen im Rückblick auf die Therapie (direkte Veränderungsmessung, allerdings nicht gleich im Anschluss an die Therapie; Albani et al., 2010, 2011). Wie bereits dargelegt, eigenen sich Feedback-Systeme, diese Raten zu reduzieren und können damit sowohl als qualitätssichernd wie auch evidenzbasiert gelten (Lambert, 2007; Lutz, 2002; Newnham & Page, 2010)[15].

Bleibt die Frage, ob die Qualitätssicherung eine relevante Ergänzung des therapeutischen Prozesses darstellt, da ein Kliniker, der eine Behandlung durchführt, doch erkennen sollte, wenn sich ein Patient negativ entwickelt. Bereits in der Vergangenheit wurde dies allerdings kritisch hinterfragt (Katsikopoulos, Pachur, Machery, & Wallin, 2008; Meehl, 1954) und aktuell wiederholt es die Frage von Hunsley und Mash (2005) nach dem zusätzlichen Nutzen diagnostischer Mittel. Forschungsbefunde wurden dazu bereits in den Abschnitten zu Feedback-Effekten (1.4.2) und zu Patientenorientierter Versorgungsforschung als Diagnostik (1.4.3) vorgestellt, daher wird hier nur noch auf zwei Forschungsbefunde eingegangen, die sich auf einen Zusatzgewinn beziehen, der eher der Qualitätssicherung zuzurechnen ist. In einer Untersuchung wurden das OQ-Feedback-System und Einschätzungen der Therapeuten verglichen (Hannan et al., 2005). Die Therapeuten sollten Ein-

[15] Ein weiterer Aspekt von Qualitätssicherung, Evaluation klinisch-psychologischer Interventionen und evidenzbasierter Forschung muss hier abgegrenzt werden: Das Forschungsfeld zu nicht wirkungsvollen oder schädlichen Behandlungsmethoden. Diese Debatte ist für den allgemeinen Wissensstand der Psychotherapieforschung wichtig, da sie ebenfalls Hinweise auf Moderatoren/ Mediatoren der Intervention-Ergebnis-Beziehung bringt (Kazdin, 2009). Zur Debatte hierüber sei auf die folgende Literatur verwiesen: Barlow, 2010; Bootzin & Bailey, 2005; Castonguay, Boswell, Constantino, Goldfried, & Hill, 2010; Dimidjian & Hollon, 2010; Lilienfeld, 2007.

schätzungen abgeben, ob ihre Patienten die Therapie gebessert, unverändert oder verschlechtert verlassen werden. Sie wurden über die Kriterien des Feedback-Systems aufgeklärt und auch darüber, dass etwa 10% der Therapien als verschlechtert klassifiziert werden. Nach dem Feedback-System ergab sich eine Verschlechterung in 7.3% der Stichprobe auf – die Therapeuten sagten dagegen lediglich für .01% der Patienten vorher, dass sie sich verschlechtern würden und trafen nur einen einzigen der Fälle korrekt. Eine andere Untersuchung stützte sich auf Fallnotizen und psychometrische Daten abgeschlossener Therapien (D. Hatfield et al., 2010). Sie untersuchten Fälle, die sich im OQ-45 (Lambert et al., 1996) zu Beginn der Therapie im klinisch belasteten Bereich befanden und sich im Verlauf reliabel verschlechtert hatten. In knapp 60% der Fälle befand sich in den Notizen kein Hinweis auf eine Einschätzung der Statusveränderung und in lediglich 21% wurde von einer Verschlechterung des Status geschrieben. Bei einer Auswahl der Fälle über noch stärkeres Verschlechterungskriterium (30 OQ-Punkte, entspricht 2 x reliable Veränderung) blieb die Quote der Nicht-Benennungen von Veränderung bei knapp 60%, aber die eingestuften Verschlechterungen stiegen auf knapp 32%.

Beide Studien bauten wieder auf dem OQ auf, der nur ein allgemeines Bild der Belastung zeichnet und ein Kliniker (sowie auch die Dyade Kliniker-Patient) mögen insgesamt andere Zielkriterien im Auge haben. Dennoch verweisen diese Ergebnisse darauf, dass diese Systeme wertvolle zusätzliche Informationen bereitstellen (Lambert et al., 2001; Lutz, 2002). Die Ergebnisse zur Entdeckung sich negativ entwickelnder Patienten sowie zum Effekt von Rückmeldungen auf Prozess- und Ergebnisqualität zeigen, dass Qualitätssicherung nicht nur eine durch Anforderungen des Marktes bestimmte Verwertung der Ergebnisse von Psychotherapieforschung ist. Hielten Eisen und Dickey (1996) noch fest, dass das Interesse an der praktischen Verwertung der Ergebnisse zur Wirksamkeit von Psychotherapie v.a. aus Akkreditierungsprozessen und damit immer noch der Rechtfertigung der Wirksamkeit der Psychotherapie gespeist sei, und konstatierten Johnson und Shaha (1996), dass Monitoring Systeme statt als Kostenersparnis nur über Plausibilität als Instrumente der Qualitätssicherung verkauft werden konnten, zeigen die vorliegenden empirischen Ergebnisse, dass der Einsatz dieser Systeme sinnvoll ist (Lambert, 2007; Lutz, 2011).

1.5. Hintergrund der vorliegenden Arbeiten

Die vorliegenden Arbeiten gehen davon aus, dass Patientenorientierte Versorgungsforschung ein Weg zur verbesserten psychotherapeutischen Versorgung ist (Howard et al., 1996; Lutz & Bittermann, 2010; Lutz, 2011). Wie bis hierhin dargestellt, gründet dieses Argument darauf, dass diese Forschung Modelle entwickelt, die eine Bewertung von Psychotherapieverläufen ermöglichen (1.4.1) und somit als wichtiger Bestandteil der Diagnostik gesehen werden können (1.4.3). Zusätzlich wurde gezeigt, dass die Verwendung von Rückmeldesystemen positive Effekte für die Behandlung von Patienten hat, die drohen, die Therapie nicht gebessert zu verlassen (1.4.2). Patientenorientierte Versorgungsforschung kann Qualitätssicherung und –entwicklung in der Praxis unterstützen (1.4.4) und durch die erhobenen Daten im Rahmen von Wissenschaftler-Praktiker-Netzwerken den Wissensstand über Psychotherapie und Veränderungsprozesse erweitern. Somit können wiederholte Erhebungen als sinnvolle Ergänzung des Therapieprozesses gelten (Hunsley & Mash, 2005; Meyer et al., 2001). Die stattgefundene Weiterentwicklung von Praxis und Theorie der Intervention sollte sich in einer Weiterentwicklung der zugrunde liegenden Methoden widerspiegeln, und dies ist in diesem Fall die Messtheorie, die den Instrumenten zugrunde liegt, die für die laufenden Erhebungen verwendet werden (Doucette & Wolf, 2009; Margison et al., 2000).

Auf diesen Aspekt zielen die drei vorgestellten Arbeiten. Sie bewegen sich im Bereich der Grundlagenforschung zu den verwendeten Methoden in der Patientenorientierten Versorgungsforschung. Eine der Fragen in der Pateientenorientierten Versorgungsforschung ist, wie die Erhebungen der Verlaufsdiagnostik (und ggf. Prozessdiagnostik) effizienter gestaltet werden können (siehe 1.4.3). Die Länge der Erhebungen kann entweder durch die Anwendung in spezifischen Populationen ein Problem werden (Gilbody, House, & Sheldon, 2002b; Lutz, Tholen, et al., 2006; Meyer et al., 2001; Walker et al., 2010), oder aber durch die Bandbreite der erhobenen Maße, die von einem einzigen Instrument (S. D. Miller et al., 2005; Shimokawa et al., 2010) bis hin zu umfangreichen Erhebungsplänen reicht. Ein Beispiel für einen solchen umfassenden Plan findet sich bei Lutz, Mocanu, et al. (2010): In dem vorgestellten Fallbeispiel werden aus der Selbstberichtsperspektive zur Indikationsstellung 15 Fragebogeninstrumente erhoben und im Verlauf zehn (hier bereits zum Teil Kurzformen; Lutz, Tholen, et al., 2006), die im Rahmen einer differentiellen Indikation sowie

der Fortschrittsbegleitung im Sinne der Patientenorientierten Versorgungsforschung als gerechtfertigt erscheinen. Die im TK Modellvorhaben eingesetzte Batterie war kleiner (fünf Instrumente für die Zwischenerhebungen), aber dennoch wird im Fazit an mehreren Stellen betont, dass der durch die psychometrischen Erhebungen verursachte Mehraufwand kompensiert werden müsse (Wittmann et al., 2011). Im Sinne der diskutierten multimodalen oder zumindest multidimensionalen Diagnostik (s. Seidenstücker & Baumann, 1987; und 1.4.3) erscheinen umfangreichere Erhebungspläne sinnvoll, doch müssen für ihren Einsatz sowohl die empirische Evidenz zur Wirkung verbreitert und die Effizienz vor dem Hintergrund psychometrischen Wissens gesteigert werden (Lutz, Tholen, et al., 2006; Meier, 1997; s.a. Kapitel 3).

Die Grundlage für die drei Studien legen mathematische Modelle, die zur Klasse der Mischverteilungsmodelle gehören (Nussbeck, Eid, & Geiser, 2010). Da mathematische Modellbildung ein wesentlicher Aspekt der Strukturierung quantitativer Daten ist, wird zunächst dargestellt, welchen Nutzen mathematische Modelle haben (1.5.1). Daran schließt sich die Frage an, weshalb eindimensionale Modelle für die Darstellung von Testdaten wünschenswerte Eigenschaften haben und wie diese die Patientenorientierte Versorgungsforschung unterstützen können (1.5.2). Als Abschluss werden die in der Arbeit verwendeten Mischverteilungsmodelle vorgestellt (1.5.3 und 1.5.4).

1.5.1. Der Nutzen mathematischer Modellbildung

Immer wenn Daten erhoben werden, stellt sich die Frage, wie sie sinnvoll zusammengefasst oder zur Beantwortung einer Frage aufbereitet werden können. In der quantitativen Forschung werden dazu mathematische Modelle verwendet, die Annahmen darüber treffen, auf welche Weise die Daten zustande gekommen sind (nicht "sein könnten": Bei der Verwendung eines Modells gibt es genau einen datengenerierenden Prozess[16]) und daraus ableiten, welche Statistiken bezogen auf diesen Prozess relevante Zusammenfassungen darstellen. Dieser Aspekt ist schon bei sehr einfachen "Modellen" erkennbar. Werden an Individuen Daten erhoben (z.B. ein Ergebnismaß in der Psychotherapie), stellt sich beispielsweise die Frage, wie die Verteilung der Werte sinnvoll ge-

[16] In dieser Arbeit wird konsequent aus einer frequentistischen Perspektive argumentiert, d.h. die durch statistische Modelle geschätzten Parameter stellen unbekannte Größen dar, die geschätzt werden um den Prozess quantifizierbar zu machen. Für die Argumentation aus Bayesianischer Sicht sei auf Jackman (2009), Kapitel 1 + 2 verwiesen.

kennzeichnet werden kann. Mittelwert und Standardabweichung sind angemessene Zusammenfassungen für die Lage einer Stichprobenverteilung, wenn in der dazugehörenden Population die Annahme einer Normalverteilung getroffen werden kann (Eid, Gollwitzer, & Schmitt, 2010; Rodgers, 2010): Diese Parameter helfen dabei, die Informationen vieler einzelner Respondenten auf zwei Kennziffern zu reduzieren und so die Verteilung der einzelnen Individuen zusammenzufassen. Auch wenn die Normalverteilung nur ein sehr einfaches mathematisches Modell ist, dient sie unter den genannten Bedingungen angemessen dem Ziel, Information (und damit Komplexität) zu reduzieren. Ein "Modell" ist also ein Objekt, das die Außenwelt in einer bestimmten Hinsicht annähern oder widerspiegeln soll, in allen anderen Aspekten aber deutlich einfacher ist (Rodgers, 2010).

In dem sehr einfachen Beispiel der Verteilungsbeschreibung kommen bereits die wesentlichen Elemente eines mathematischen Modells zum Vorschein (Hennig, 2010; Rodgers, 2010): Als "mathematisches Modell" wird eine Sammlung mathematischer Strukturen bezeichnet (in dem vorangegangenen Beispiel Mittelwert, Standardabweichung und die Formel der Dichtefunktion der Normalverteilung), die zwei Bedingungen erfüllt (Hennig, 2010). Diese sind:

a) Die mathematischen Objekte haben eine Interpretation in Form von Objekten[17] außerhalb des Modells: Im Beispiel der Verteilungsbeschreibung geht es nicht um die Normalverteilung an sich, sondern um die Beschreibung oder Zusammenfassung der Verteilung der erreichten Werte einer Stichprobe in dem Ergebnismaß.

b) Die mathematischen Operationen, die ausgeführt werden, haben ebenfalls eine Widerspiegelung in der Welt: Im Beispiel der Verteilungsbeschreibung spiegelt die berechnete mathematische Dichte der Verteilung die vorhergesagte Dichte eines bestimmten Wertes im Ergebnismaß in der Population wider.

Ein mathematisches Modell überführt Hypothesen, Annahmen o.ä. eines Gegenstandsbereiches in ein Gerüst aus Formeln (detaillierte Diskussion bei Luce, 1995). Durch ihre Formalisierung sind

[17] Der Begriff "Realität" und Umschreibungen dieses Begriffes werden in dieser Arbeit lediglich für die Kennzeichnung intersubjektiv nachvollziehbarer Sachverhalte verwendet. Eine ontologische Forderung geht damit nicht einher. Für Diskussionen des Realitätsbegriffes sei verwiesen auf: Borsboom, Mellenbergh, & Van Heerden, 2003; Hennig, 2010; Longino, 1990; van Fraassen, 1980.

mathematische Modelle präzisier formuliert als sprachliche Modelle und daher besser falsifizierbar. Damit enthält die Falsifikation mehr Information, da deutlicher erkennbar ist, in welcher Hinsicht das gewählte Modell nicht passt. Modellierung dient auch dazu, eine klare und präzise Sprache zu finden, die es ermöglicht Ideen zwischen Personen mit möglichst geringem Informationsverlust auszutauschen. Durch ein Modell wird kommuniziert, welche Elemente als relevant angesehen werden (welche ausgelassen werden), und damit wird bewertbar, ob das Modell für den ausgewählten Gegenstandsbereich (und den Zweck des Rezipienten) angemessen ist.

Dies verdeutlicht, dass es bei der mathematisch-statistischen Modellbildung (im Folgenden "Modellierung") nicht um einen Primat der Mathematik geht, sondern darum, Beschreibungen der Welt mit formalen Modellen zu ermöglichen, die sparsamer als die Gesamtdaten sind, und bezogen auf den relevanten Untersuchungsaspekt hinreichend gut sind. Der Anspruch ist ein explorativ erkenntnistheoretischer: "...mathematical modelling is about the investigation of the implications of ways of thinking about reality"[18] (Hennig, 2010, p. 43). Mathematische Modelle ermöglichen, Daten in Hinblick auf eine spezifische Frage zu vereinfachen und durch die Kenntnis der mathematischen Eigenschaften des gewählten Modells, zu neuen testbaren Implikationen führen[19]. Durch a) neue testbare Implikationen und b) den Vergleich verschiedener Modelle wird es möglich, unterschiedliche Szenarien zu vergleichen. So können mathematische Modelle nicht nur die Kreativität fördern, sondern ermöglichen ein strukturiertes Untersuchen und damit Lernen über mögliche Verhaltensweisen der Welt. Wichtig ist dabei zu verstehen, dass auch mathematisches Modellieren immer einen explorativen Charakter hat, da die Passung des Modells immer nur eine Hypothese ist. Der größte erkenntnistheoretische Gewinn entsteht nach Hennig (2010) in der Regel da, wo die Unterschiede zwischen den Modellen und Beobachtungen untersucht werden können.

[18] "...mathematische Modellierung beschäftigt sich mit der Untersuchung der Konsequenzen unterschiedlicher Arten über die Realität zu denken." Übers. durch Autor

[19] Diese Perspektive unterstreicht, warum die Verwendung von Standardmodellen wie dem Nullhypothesen-Test aus der Sicht der Modellierung zu kritisieren ist: a) durch die unüberlegte Anwendung des Modells ist kein Gewinn an Erkenntnis aus den mathematischen Eigenschaften zu erwarten; b) ist damit nicht geprüft, ob dieser Test überhaupt ein sinnvolles Modell z.B. zur Beschreibung des Unterschiedes zwischen zwei Populationen ist; und c) fehlt der Vergleich mit einem relevanten Alternativmodell, was die Nullhypothese ja in der Regel nicht ist (Harlow, Mulaik, & Steiger, 1997)

Schon allein aus dem Grund, dass Modelle das Ziel der Vereinfachung haben, und die Realität auf eine gewisse Anzahl von Parametern reduzieren, wird bereits deutlich, dass vor der Modellierung die Wertentscheidung getroffen werden muss, was interessante Komponenten der Realität sind, auf die das Modell hin vereinfachen soll. Dieser Aspekt mathematischer Modelle wird in dem mittlerweile fast sprichwörtlichen Zitat von Box auf den Punkt gebracht (Box, 1979, p. 2): "Models, of course, are never true, but fortunately it is only necessary that they be useful".[20] Dies stellt die Frage der Nützlichkeit als zentralstes Bewertungselement von Modellen in den Mittelpunkt. Und diese lässt sich nur aus den Verwendungszwecken des Modells heraus beantworten.

Die Konsequenzen aus der Nicht-Passung eines Modells können unterschiedlich sein (Rodgers, 2010). Zum Einen könnte das Modell komplexer gemacht werden, um weitere Aspekte zu berücksichtigen. So machen Modelle aufmerksam auf das, was in der verwendeten Theorie zur Erstellung des Modells noch gar nicht berücksichtigt wurde. Zum Anderen liegt ein Wert eines Modells darin dass es eben einfacher ist als die Realität, da es ein Modell für bestimmte Aspekte der Realität ist. Bei der Verwendung mathematischer Modelle muss kenntlich gemacht werden, welche Aspekte der Realität modelliert und welche ausgelassen werden sollen. Und ein Modell, das bestimmte Aspekte auslässt, kann immer noch sehr brauchbar für den Zweck sein, den der Modellierer im Sinn hatte und/oder für den es der Anwender benötigt.

In der vorliegenden Arbeit werden Mischverteilungsmodelle zur Modellierung von Fragebogendaten verwendet. Mischverteilungsmodelle sind mathematische Modelle, die nicht davon ausgehen, dass alle Untersuchungseinheiten aus einer Population stammen. Sie nehmen stattdessen an, dass die Verteilungswerte, die als Population beobachtet werden, durch verschiedene Subpopulationen zustande kommen (Nussbeck et al., 2010; Rost, 2004). Ein Beispiel für manifeste Mischverteilungen sind Daten in einem Instrument zur Messung der symptomatischen Belastung, die aus Stichproben inner- und außerhalb von Behandlungskontexten erhoben werden. Die gemeinsame Verteilung dieser Daten ist oft multimodal und die jeweilige Population, aus der sie stammen, klärt diese Multimodalität durch jeweils zwei eigene Verteilungen auf (s. Kapitel 3; Connell et al.,

[20] "Modelle sind natürlich niemals wahr, aber glücklicherweise müssen sie lediglich nützlich sein." Übers. durch Autor.

2007). Die Variablen, die die "Mischung" herstellen, sind aber oft nicht bekannt, und so ist es das Ziel bei der Anwendung von Mischverteilungsmodellen, aufgrund mathematischer Prinzipien die beobachtete Stichprobe zu "entmischen" – sie zumindest deskriptiv in unterschiedliche Gruppen zu zerlegen. Dies geschieht durch die Annahme mindestens einer latenten Variable, die die Zugehörigkeit zu einer latenten Population repräsentiert (Gollwitzer, 2007; Kempf, 2003, 2008; Nussbeck et al., 2010; Rost, 2004). Für diese Arbeit werden eindimensionale Modelle verwendet und diese Wertentscheidung soll im Folgenden vor der Darstellung der Modelle mit dem besonderen Nutzen eindimensionaler Modelle begründet werden.

1.5.2. Der Nutzen eindimensionaler Modelle

Wie dargestellt (1.5.1) muss bei der Verwendung von Modellen entschieden werden, welche Aspekte von Realität berücksichtigt werden sollen. Die Auswahl eines mathematischen Modells hängt dabei

a) von den untersuchten Daten ab (hier: Daten eines Instrumentes zur Messung psychischer Belastung, d.h. ein Modell für das Zustandekommen von Antworten auf Items muss verwendet werden),

b) und von Praktikabilitätseinschätzungen des Gegenstandsbereiches (hier v.a. die Frage, wie die Ergebnisse in der therapeutischen Praxis eingesetzt werden können).

In dieser Arbeit werden eindimensionale Modelle verwendet: Das Rasch-(bzw. Partial-Credit-) Modell sowie eine eindimensionale Latent Profile Analysis (mit latenten Varianzen als Merkmal zur Klassendefinition). Ein eindimensionales Modell hat den Vorteil, dass es annimmt, alle Unterschiede in der multivariaten Verteilung der Testitems würden sich auf quantitative Unterschiede zwischen den Personen zurückführen lassen (Kempf, 2003, 2008; Rost, 2004). Dies ist wünschenswert, da sich so in der Anwendungspraxis durch die Verwendung einer Kennziffer (der Summe der Itemantworten) beurteilen lässt, ob sich der entsprechende Patient verbessert oder verschlechtert hat, welcher Patient im Vergleich stärker belastet ist usf. Dieser Praktikabilitätsaspekt wird oft so hoch bewertet, dass trotz gegenteiliger Evidenz Auswertungsmethoden für Fragebogen-

daten verwendet werden, die eine Eindimensionalität voraussetzen (Doucette & Wolf, 2009; M. C. Edwards, Cheavens, Heiy, & Cukrowicz, 2010)

Neben diesem praktischen Vorteil eindimensionaler Erhebungen, gibt es auch Vorteile für die Forschung. Für die Verwendung eindimensionaler Modelle spricht, dass eine Theorie oder Hypothese, die wiederholt getestet wird, plausibler wird, je nachdem, wie gut sie sich bestätigen lässt. Dieser Prozess hängt von vielen Faktoren ab und es ist in der Regel nicht entscheidbar, woran die Bestätigung oder Verwerfung scheitern (Duhem, 1998; Quine, 1951). Im Idealfall eindimensionaler Konstrukte ist bei der Testung zumindest gewährleistet, dass die Beziehungen zwischen den Konstrukten, über die durch die Operationalisierungen in den verwendeten Messinstrumenten gesprochen wird, nicht durch weitere Variablen kontaminiert waren (G. T. Smith, McCarthy, & Zapolski, 2009): Soll die Wirkung bestimmter Interventionsmethoden auf die Dimension "Depression" untersucht werden, wird aber durch das Erhebungsinstrument depressive wie auch angstbezogene Belastung miteinander vermischt, ist nicht mehr klar nachvollziehbar, welche der Dimensionen sich geändert hat, oder ob sich nicht Veränderungen auf beiden Dimensionen vielleicht gegenseitig aufgehoben haben (Hunsley & Mash, 2005; McFall, 2005).

Eine Konsequenz ist, dass mit eindimensionalen Maßen deutlicher feststellbar ist, welche Komponente mit welchen externen Variablen variiert (z.B. Art der Intervention). Ist das Maß für diese Belastung (zu) heterogen, kann dies vor allem zwei Konsequenzen haben. Zum einen fällt die Beantwortung der Frage des Zusammenhanges mit anderen Konstrukten schwer, da zwar Korrelationen gefunden werden, aber nicht beurteilt werden kann, durch was diese Zusammenhänge entstehen (z.B. kognitive vs. affektive Symptomatik im Beck Depressionsinventar: Keller, Hautzinger, & Kühner, 2008; Keller & Kempf, 1997; Quilty, Zhang, & Bagby, 2010). Zum anderen wird die Veränderungsbeobachtung erschwert, weil bspw. Verbesserungen in einem Teil des heterogenen Konstruktes stattfinden, aber Verschlechterungen im anderen Teil und sich diese Veränderungen gegenseitig aufheben. Dieser Punkt wurde in ähnlicher Form bereits von Stiles und Kollegen (1986) als einer der Gründe dafür festgehalten, dass die Forschung im Matrix-Paradigma keine systematischen Unterschiede zwischen den Therapieschulen hervorbringen konnte.

So lange das Konstrukt zwar in heterogen scheinende Subkonstrukte zerlegbar ist, die heterogenen Anteile aber auf dieselbe Art mit den externen Variablen zusammenhängen, "…then there is no evidence that the two measures reflect meaningfully different psychological processes. The use of two terms and two measures would be both unnecessary and potentially misleading" [21] (G. T. Smith et al., 2009, p. 274). Mit anderen Worten: Wenn etabliert werden kann, dass die Subkonstrukte in ähnlicher Weise auf die relevanten Kriterien zusammenhängen, dann ist eine konzeptuelle Trennung in der Theorie angemessen, aber in der Praxis ist sie zumindest vor dem Hintergrund der vorhandenen Messinstrumente nicht sinnvoll (Böhnke & Lutz, 2011a; Reininghaus, McCabe, Burns, Croudace, & Priebe, 2011; Vissers, 2010).

In der Praxis bedeutet dies, wenn mit einer Intervention gezielt ein bestimmter Symptombereich verändert werden soll (z.B. negative Selbstüberzeugungen) und das verwendete Messinstrument für die Symptomveränderung erfasst mehr als diesen Symptombereich, sind Rückschlüsse auf die Wirksamkeit der Intervention behindert (Brown & Barlow, 2009; Schulte, 1993). Interventionen sind oft auch spezifischer als nur auf "Depression" ausgerichtet, z.B. auf spezifischer Problembereiche, für die geprüft werden muss, ob die Intervention tatsächlich wirkt. Das können Therapieziele sein, eher somatische Probleme (Schlaflosigkeit als Teil der Depression) oder spezifische Kognitionen. Für diese Bereiche sollten jeweils Instrumente bereitgestellt werden, die es ermöglichen, kriterienorientiert und engmaschig zu prüfen, ob sich der Belastungsgrad ändert.

Aufgrund dieser Argumente wird in dieser Arbeit der Standpunkt vertreten, dass die Verwendung eindimensionaler Modelle für Fragebogendaten praktische wie forschungspraktische Vorteile hat, die genutzt werden sollten (Brown & Barlow, 2009; Kamphuis & Noordhof, 2009; G. T. Smith et al., 2009). Dies bedeutet nicht, dass neue Instrumente entwickelt werden müssen. Eine Prüfung und Analyse vorhandener Instrumente und eine Optimierung auf die entsprechenden Testzwecke reicht in diesem Fall (Doucette & Wolf, 2009; McFall, 2005) und würde eine hinreichenden Fortschritt in der Praxis bedeuten (M. C. Edwards et al., 2010). Der Einsatz verschiedener eindimensi-

[21] "...dann liegt keine Evidenz vor, dass die beiden Maße bedeutsam unterschiedliche psychologische Prozesse abbilden. Die Verwendung von zwei Begriffen und zwei Maßen wäre sowohl unnötig und möglicherweise irreführend." Übers. durch Autor

onaler und hinreichend verschiedener Instrumente kann sich dann effizient an den Gegebenheiten des diagnostischen Kontextes orientieren (s. 1.4.3; McFall, 2005; G. T. Smith et al., 2009) und auch die Integration von multidimensionalen Studienergebnissen ist in einer Form möglich, die für den Praktiker verwertbar ist (De Los Reyes & Kazdin, 2006; De Los Reyes et al., 2011).

1.5.3. Die Studien I und II: Das Rasch Modell

In dieser Arbeit werden zwei spezielle Mischverteilungsmodelle verwendet. Das erste, das Rasch-Modell, stammt aus der Gruppe von Mischverteilungsmodellen für kategoriale manifeste Variablen (Rasch, 1961). Dieses Modell prüft, ob die Untersuchungseinheiten in einer Stichprobe aus einer einzigen Population stammen und sich nur durch die Ausprägung der gemessenen latenten Variable unterscheiden. Das heißt, die Mischung der Individuen kommt nur dadurch zustande, dass sie entlang der gemessenen Fähigkeit in Gruppen unterschiedlicher Ausprägung eingeteilt werden können. Im vorliegenden Fall in Gruppen unterschiedlicher psychischer Belastung (Kempf, 2003, 2008; Moosbrugger & Kelava, 2007).

Das zweite Modell, die Latent Profile Analysis, untersucht kontinuierliche, manifeste Variablen und versucht diese auf eine latente kategoriale Mischverteilungsvariable zurückzuführen (Gibson, 1959). Die Mischung kommt in diesem Fall also dadurch zustande, dass es unterschiedliche Subpopulationen (Profile oder Klassen) gibt, die sich üblicherweise durch ihre spezifischen Mittelwertsprofile bei der Beantwortung von Variablen auszeichnen (kontinuierlicher Fall der Latent Class Analyse; Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968). Beide Modelle werden kurz vorgestellt; weitere Details finden sich in den empirischen Studien.

Rasch-Modelle nehmen an, dass die Information, die in einer Reihe von manifesten Variablen (typischerweise Items) enthalten ist, die numerisch kodiert werden, sich durch eine Aufsummierung der Kategorien-Nummern darstellen lässt. Rasch-Modelle nehmen an, dass es eine latente Dimension gibt, die die Antworten auf die Items durch die Personen "hervorruft" (Borsboom, Mellenbergh, & Van Heerden, 2003). Je höher die Ausprägung dieser latenten Dimension ist, desto höher fällt die Antwort der entsprechenden Person auf die Items aus.

Die mathematische Form für den dichotomen Fall ist:

$$P(X_i = 1) = \frac{\exp(\theta_v - \delta_i)}{1 + \exp(\theta_v - \delta_i)} \qquad \textbf{Formel 1-3}$$

Dabei bezeichnet θ_v die Ausprägung auf der latenten Dimension einer Person *v* und δ_i die sog. Schwierigkeit eines spezifischen Items *i*. Im Rasch-Modell werden Items und Personen auf einer Dimension gemeinsam bewertet. Im dichotomen Fall ist die Wahrscheinlichkeit $P(X_i=1)$, ein Item *i* in der Kategorie "1" zu beantworten, abhängig von den relativen Positionen der Person *v* und des Items *i* auf der latenten Dimension. Befinden sich Person und Item exakt an derselben Stelle der latenten Dimension, hat die Person eine Wahrscheinlichkeit von $P(X_i=1) = .5$ für die Antwort in Kategorie "1". Je weiter die Person oberhalb des Items auf der latenten Dimension liegt, desto mehr steigt diese Wahrscheinlichkeit gegen $P(X_i=1) = 1$ (bzw. gegen $P(X_i=1) = 0$, je weiter unterhalb die Person liegt). Diese Beziehung lässt sich durch die sog. Itemcharakteristiken (Rost, 2004; "Item Characteristic Curves") ausdrücken (siehe Abbildung 1-1).

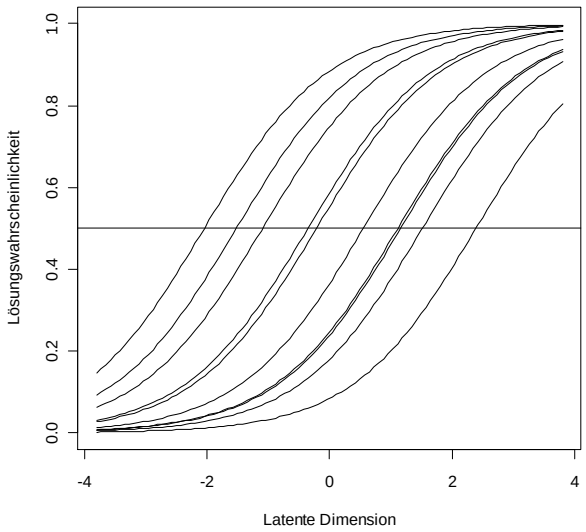

Abbildung 1-1: Beispiele für Itemcharakteristiken 10 dichotomer Items.

Für den polytomen Fall wird angenommen, dass nicht nur jedes Item durch seine "Schwierigkeit" auf dieser latenten Dimension verankert werden kann, sondern durch die Übergänge zwischen jeder der *m* Kategorien des Items. Ein mögliches Modell für diesen Fall ist das sog. "Partial Credit

Modell" (PCM; Masters, 1982). Das PCM drückt für jedes Item die Wahrscheinlichkeit aus, dass die Antwort einer Person in einer bestimmten Kategorie ausfällt, gegeben die Wahrscheinlichkeiten aller anderen Kategorien:

$$P(X_i = x) = \frac{\exp \sum_{k=1}^{x} (\theta_v - \tau_{ki})}{1 + \sum_{x=1}^{m} \exp \sum_{k=1}^{x} (\theta_v - \tau_{ki})} \qquad \textbf{Formel 1-4}$$

Das PCM geht von einer Beziehung der Schwellen τ_{ki} aller Kategorien k eines Items i und der latenten Dimension aus. Die Schwellen sind dabei definiert als der Punkt auf der latenten Dimension, an dem eine Person v mit exakt dieser latenten Ausprägung (θ_v) die gleiche Wahrscheinlichkeit hat, in den beiden an diese Schwelle grenzenden Kategorien des Items i zu fallen. Der Vorteil des PCM gegenüber anderen IRT-Modellen ist, dass wenn seine Gültigkeit geprüft und nicht falsifiziert wurde, der Summenscore über die verwendeten Items als Messung der latenten Eigenschaft verwendet werden kann ("Suffizienz des Summenscores"; z.B. Rost, 2004; Thissen & Orlando, 2001). Dies macht die Brücke zwischen der einfachen Summierung in der Testanwendung und der gleichzeitigen Verwendung der IRT-Modellschätzer möglich (s. Kapitel 3). Abbildung 1-2 zeigt Ergebnisse der Schätzung des PCM des FEP (Ergebnisse in Kapitel 3.3.3). Die Abbildungsteile zeigen die Kategorienwahrscheinlichkeiten (Kategoriencharakteristika) für jeweils ein Item. Dort, wo sich die Kurven für zwei benachbarte Kategorien schneiden, befindet sich der jeweilige Schwellenparameter. Zusätzlich ist durch die unterschiedliche Form der Kurven erkennbar, dass nicht für alle Items dieselben Abstände zwischen den Kategorien gelten müssen, was mit üblichen Antwortformaten von Fragebögen (so auch dem FEP) zwar angestrebt wird, aber nicht zwangsläufig realisiert wird (Rost, 2004).

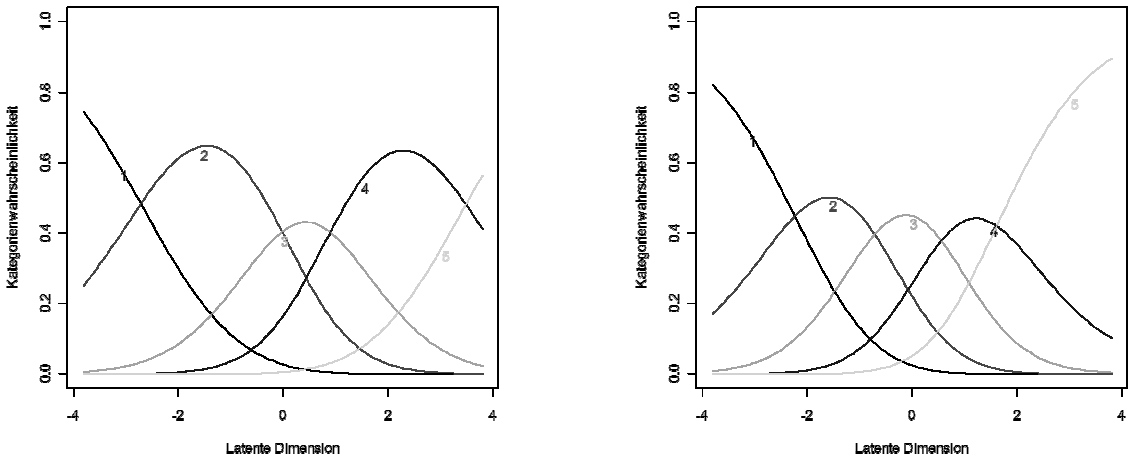

Abbildung 1-2: Kategoriencharakteristika für zwei unterschiedliche Items des FEP; links Item 1 "...fühlte ich mich wohl", von dem besonders Kategorien 2 ("selten") und 4 ("oft") einen großen Bereich der latenten Dimension abdecken; rechts Item 4 "...war ich nervös", das eine gleichmäßige Verteilung seiner Kategorien über den Bereich der latenten Dimension zeigt.

Das Rasch Modell beschreibt das Zustandekommen der Antworten auf die Items eines Tests, wenn angenommen wird, dass alle Items des Tests eine Dimension beschreiben. Die durchgeführten mathematischen Operationen spiegeln wider, wie sich die relativen Antworthäufigkeiten auf ein Item entlang der latenten Dimension ändern, wenn dem Instrument tatsächlich eine Dimension zugrunde liegt. Das Rasch Modell versucht einige Aspekte eines Fragebogens widerzuspiegeln: Neben der Eindimensionalität nimmt das Rasch-Modell an, dass sich die Antworten auf einen Fragebogen durch sehr wenige Parameter beschreiben lassen. Die Items variieren auf dieser einen Dimension nur in ihrer Schwierigkeit. Diese zeigt an, wie hoch die gemessene Eigenschaft einer Person ausgeprägt sein muss, damit diese in höheren Kategorien eines Items ihre Antwort macht. In einem Symptomfragebogen gibt die Schwierigkeit also an, wie stark die Belastung der Person ausgeprägt sein muss, damit sie ein bestimmtes Symptom aufweist (dichotomer Fall) oder aber von diesem stärker belastet wird (polytomer Fall). Der Zusammenhang zwischen Schwellenparametern und Itemschwierigkeit im polytomen Fall ist in Abbildung 1-3 dargestellt. Der linke Teil der Abbildung zeigt wie Abbildung 1-2 die Kategoriencharakteristika von Item 4 des FEP, deren Schnittpunkte die Schwellenparameter darstellen. Die senkrechte Linie zeigt, die Schwierigkeit des Items an, die im polytomen Fall etwa als Mittel der Schwellen modelliert wird (Details: Rost, 2004). Im

rechten Teil der Abbildung ist die dazugehörige Iteminformationsfunktion abgebildet, die den Verlauf der Messgenauigkeit für dieses Item zeigt (Details dazu siehe unten und Kapitel 3).

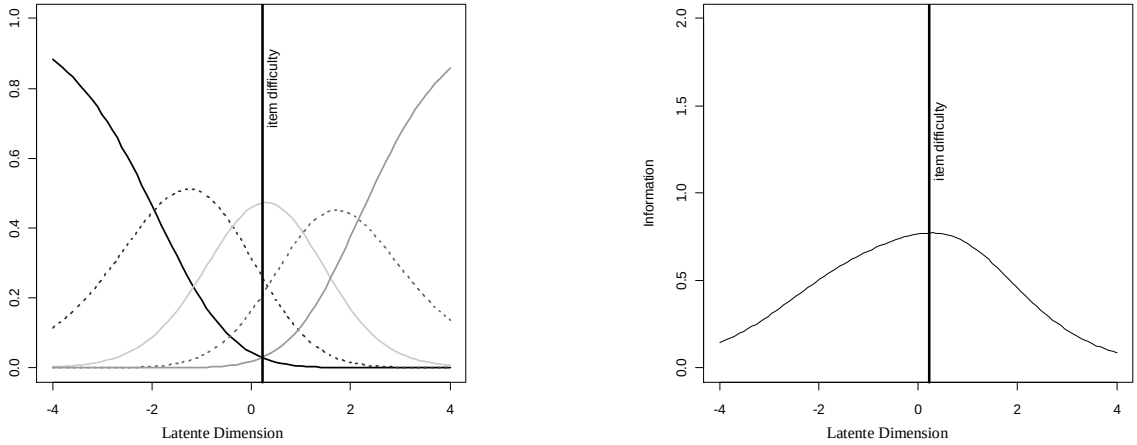

Abbildung 1-3: Der linke Teil der Abbildung zeigt die Kategoriencharakteristika für Item 4 des FEP mit eingetragener Schwierigkeit des Items ("item difficulty"); der rechte Teil zeigt die Informationsfunktion des Items, ebenfalls mit eingetragener Schwierigkeit.

Die Reduktion auf diese wenigen Parameter wird deshalb als gerechtfertigt angesehen, da Eindimensionalität eine für den Kontext wünschenswerte Eigenschaft ist (siehe 1.5.2) und die Schwierigkeits- und Schwellenparameter bereits einige Flexibilität in der Verbindung zwischen latenter Eigenschaft und dem Item ermöglichen. Zusätzlich geben nur diese Modelle die Möglichkeit, dass sich für die Testanwender im Vergleich zur derzeitigen Praxis möglichst wenig ändert: Wenn gezeigt wurde, dass das Rasch-Modell für eine Gruppe von Items gilt, kann der Summenscore als Repräsentant der latenten Dimension ausgewertet werden (Thissen & Orlando, 2001). Außerdem können die Summenscores direkt in die Personenparameter der latenten Dimension umgerechnet werden und somit mit den Vorteilen der Differenzenskalierbarkeit (Kempf, 2003) und den individuellen Standardfehlern für jeden Personenparameter ausgewertet werden (s. z.B. Kapitel 3).

Durch die mathematische Formulierung des Rasch-Modells wird angegeben, wie sich die Kategoriencharakteristika im Verhältnis zur latenten Dimension verhalten sollen. Aus der Falsifikation dieser Annahme kann die Funktionsweise der einzelnen Items untersucht werden (Adams, Wu,

& Wilson, 2012). Das Rasch-Modell liefert einen klar definierten Begriff der Testfairness: Die Schwierigkeiten der Items sollten sich nicht zwischen Subgruppen der Testpopulation unterscheiden, was sich durch sog. "Differential Item Functioning" zeigen würde: Dasselbe Item wäre für Personen gleicher Fähigkeit (hier also: gleicher Belastung) unterschiedlich schwer (Osterlind & Everson, 2009). Dadurch, dass Items und Personen auf derselben latenten Dimension verankert sind, kann auch beurteilt werden, für welche Bereiche des latenten Kontinuums ein Test besonders geeignet ist (viele Items/ hohe Genauigkeit) und für welche Bereiche noch weitere Items entwickelt werden müssen. Das Rasch-Modell macht es auch möglich, die Annahme eines gleichbleibenden Messfehlers über das Kontinuum der latenten Eigenschaft aufzugeben. Jeder Score erhält im Rasch-Modell einen eigenen Messfehler, der auch dazu genutzt werden kann, Veränderungen z.B. über den Verlauf von Therapien zu bewerten (siehe 1.4.1 und Reise & Haviland, 2005).

Zwei Implikationen des Modells für allgemeine Testgütemerkmale sind die bereits genannte Möglichkeit der Prüfung auf Differential Item Functioning und die Überprüfung der Modellgültigkeit durch entsprechende Statistiken. Hierzu stehen zum einen die Fit-Indizes zur Verfügung, die prüfen, wie viel nicht-aufgeklärte Varianz für jedes Item/jede Person verzeichnet wird (Bond & Fox, 2007) oder aber globale Fit-Tests, die auf der Eigenschaft des Rasch-Modells der "Spezifischen Objektivität" beruhen (Rost, 2004). Ein weiterer Aspekt ist der sich über das Spektrum der latenten Eigenschaft verändernde Messfehler, der in der sog. "Testinformationsfunktion" gefasst wird, die aus der Modellformulierung folgt: Mit ihr kann die Messgenauigkeit der Items an den einzelnen Abschnitten des latenten Kontinuums erfasst werden und auch so festgestellt werden, wo sich Zonen mit sehr niedriger Messgenauigkeit befinden (Embretson & Reise, 2000).

Eine weitere zentrale Eigenschaft der mathematischen Formulierung des Modells ist, dass ein Zusammenhang zwischen der latenten Dimension und jedem einzelnen Item formuliert wird. Es können so Items unabhängig von einander aus der Skala ausgewählt werden, die für einen bestimmten Erhebungszweck besonders geeignet sind. Dieses Prinzip wird beim computer-adaptiven Testen verwendet (Wainer, 2000), doch kann dies auch für statische Testkurzformen genutzt werden (Cella, Gershon, Lai, & Choi, 2007). Es können z.B. Kurzformen eines Tests dazu verwendet werden, um Screenings durchzuführen (Lehr, Hillert, Schmitz, & Sosnowsky, 2008) und eine ande-

re Kurzform kann dazu genutzt werden, um die Veränderung der Patienten im Verlauf der Therapie zu betrachten. So lange all diese Testformen aus einem Test (im IRT Sinne aus einer Item-Bank; Riley et al., 2010) stammen, können alle diese mit unterschiedlichen Verfahren bestimmten Personenparameter miteinander verglichen und benutzt werden.

Vor diesem Hintergrund sind übergreifende Fragen der folgenden Studien:

a) Können diese Modelle mit frei verfügbarer Software hinreichend genau geschätzt werden, damit sie in der Praxis tatsächlich verwendet werden können? (Kapitel 2)

b) Wie groß müssen die Stichproben sein, mit denen diese Parameter geschätzt werden und ist dies von der Verteilung der Stichprobendaten abhängig? (Kapitel 2)

c) Wie können die Eigenschaften dieser Modelle kombiniert werden, um mit ihnen Kurzformen von Fragebögen zu erstellen und zu testen, die für die Praxis und Verwendung in der Forschung angemessene Messeigenschaften haben? (Kapitel 3)

1.5.4. Die Studie III: Die Latent Profile Analysis

In der dritten Studie wird die Latent Profile Analysis (Gibson, 1959) eingesetzt. Während das Rasch-Modell als konfirmatorisches Modell gesehen werden kann, da in der Sprache der Strukturgleichungsmodelle ein einfaktorielles Modell mit auf "1" festgesetzten Ladungen der Items und nicht restringierter Struktur der polychoren Schwellen am nächsten kommt (Maydeu-Olivares & McArdle, 2005), ist die Latent Profile Analysis ein Verfahren, das in der Regel explorativ verwendet wird. Die Latent Profile Analysis zerlegt ausgehend von a) arbiträren Startwerten und b) einer vorgegebenen Anzahl von Profilen, die Stichprobe in Subgruppen, die sich durch ähnliche Mittelwerte und Standardabweichungen auf den Indikatoren auszeichnen. Da dieses Ergebnis oft als ein Verlauf von Mittelwerten dargestellt wird, werden die Subgruppen "Profile" genannt. Tatsächlich werden allerdings basierend auf dem Prinzip der lokal-stochastischen Unabhängigkeit Zentroide (multivariate Mittelwerte) identifiziert, die die Verteilung der Werte in der Stichprobe beschreiben. Neben der Annahme der lokal-stochastischen Unabhängigkeit wird bei der Extraktion die Annahme genutzt, dass die Fehlerverteilungen um die Mittelwerte normalverteilt sind und in der Regel

auch, dass sich diese Varianzen zwischen den Gruppen nicht unterscheiden (Fraley & Raftery, 2002; Gibson, 1959; Haughton, Legrand, & Woolford, 2009).

Das Modell beschreibt somit, ob und in wie viele Zentroide sich die vorhandenen Daten zerlegen lassen. Bei der Anwendung auf einen Fragebogen wird also davon ausgegangen, dass die Antworten auf die Items aus einer Normalverteilung mit *k* Dimensionen stammen (k als Anzahl der Items) und es Gruppen gibt, die sich durch gemeinsame Zentroide und Streuungen um diese auszeichnen. Die Zentroide werden dadurch gefunden, dass in jeder der Gruppen um einen solchen Zentroid das Prinzip der lokal stochastischen Unabhängigkeit gilt, d.h. die Korrelationen innerhalb einer Gruppe zwischen den Indikatoren gegen "0" streben (Gibson, 1959). Die Widerspiegelung der mathematischen Operationen in der wirklichen Welt ist damit derjenigen des Beispiels der Normalverteilung oben sehr ähnlich: Um die Stichprobe zu beschreiben, wird sie in verschiedene Zentroide mit multivariaten Standardabweichungen zerlegt und die resultierende Dichteverteilung formt einen Raum der größten Wahrscheinlichkeit für weitere Beobachtungen.

Bei der Latent Profile Analysis werden unter der Annahme von multivariaten Normalverteilungen nur Mittelwerte und Standardabweichungen als relevante Parameter gesehen und modelliert. Im Gegensatz zu Verfahren wie der Diskriminanzanalyse oder der MANOVA wird angenommen, dass ein multivariates Set an Indikatoren als Kennzeichnung für eine kategoriale Variable verwendet werden kann, die aber nicht wie bei diesen genannten Verfahren manifest, sondern latent ist. Diese latente Variable wird als Ursache für die verschiedenen Gruppen gesehen. Im Gegensatz zur klassischen Summenbildung wird hier angenommen, dass es in der Kombination dieser quantitativen Indikatoren auch qualitative Unterschiede geben kann, dass eben Muster entscheidend sind. In der vorliegenden Analyse wurde dieses Modell weiter eingeschränkt. Da für den verwendeten Fragebogen (FEP; Lutz, Schürch, et al., 2009) bereits Evidenz vorlag, dass die verwendeten Items im Wesentlichen als eindimensional angesehen werden können, wurde eine eindimensionale Latent Profile Analysis verwendet: Als Gruppenzentroide wurden die Mittelwerte von Prä- und Post-Erhebungen einer klinischen Stichprobe verwendet sowie eine Erhebung einer Normstichprobe. Die um diese drei Zentroide verbleibende Restvarianz wurde dann versucht durch latente Klassen von Varianzmustern um die Items aufzuklären (Dattatreya, 2002).

Ziel dieser Vorgehensweise ist es zu untersuchen, wie sich die Restvarianzen um diese Gruppenmittelwerte unter der Annahme der Eindimensionalität verhalten. Wäre der Test eindimensional und die drei Gruppen distinkte Populationen, sollten die Zentroide der Items reichen, um unterschiedliche Grade und Qualitäten psychischer Belastung zwischen diesen Gruppen zum Ausdruck zu bringen. Alle so nicht aufgeklärte Restvarianz muss sich also in den Varianzen der Zentroide niederschlagen. Die in Kapitel 4 verwendete Systematik erlaubt es, die Zentroide zu vergleichen und so Items zu identifizieren, die besonders sensitiv für die Unterschiede zwischen Prä- und Postwerten sowie die Bevölkerungsstichprobe sind. Dies ist nützlich, um die Sensitivität für Veränderungen (Lutz, Tholen, et al., 2006; siehe auch Debatte zu "responsiveness", z.B. Helmreich et al., 2011; Murawski & Miederhoff, 1998) in einem Mehrgruppensetting erfassen zu können: Hier wird nicht die Variabilität der Differenzen erhoben und bewertet, sondern wie stark die Items in diesen als separat behandelten Populationen (Prä, Post, Norm) variieren, um dann diejenigen auswählen zu können, die am zuverlässigsten eine Veränderung anzeigen.

2. Studie I: Abhängigkeit der Schätzer für Item- und Personenparameter von Itemzahl und Stichprobengröße[22]

Eine vergleichende Simulation mit drei R-Paketen

2.1. Einleitung

Testmodelle der Item Response Theorie (IRT) haben sich in vielerlei Hinsicht als den Betrachtungen der Klassischen Testtheorie überlegen erwiesen. Sie ermöglichen eine deutlich detailliertere Analyse der Items und damit eine erleichterte Optimierung der Tests (z.B. Doucette & Wolf, 2009; Hambleton, Swaminathan, & Rogers, 1991), flexiblere Verwendung der Tests und Items (z.B. computer-adaptives Testen; Wainer, 2000; z.B. auch Walker, Böhnke, Cerny, & Strasser, 2010) und die Berücksichtigung von individuellen Antwortstilen und -tendenzen (z.B. Böhnke & Lutz, 2008; Rost, 2004; Schürch et al., 2009) Differential Item Functioning: Osterlind & Everson, 2009). Eine gute Übersicht über die Vorteile von IRT Modellen im Kontext der Erhebung patientenbezogener Maße geben Chang & Reeve (2005). Doch trotz dieser Vorteile sind IRT Modelle in der Praxis noch nicht so verbreitet. Zwei wesentliche Gründe werden immer wieder angeführt, warum IRT nicht breit in praxisorientierterer Forschung angewendet wird: Kostenintensive Software und die nötigen Stichprobengrößen (Zickar & Broadfoot, 2009).

2.1.1. Software

Die Software-Begründung besagt, dass Spezialsoftware zur Schätzung der Modelle nötig ist, Zeit für die eigene Einarbeitung und zusätzliche Kosten (z.B. Lizenzgebühren) anfallen. In großen Statistiksoftware-Paketen sind IRT-Modelle bis heute nicht integriert (z.B. SPSS 19.0) oder nur als Ergänzungen zu haben (z.B. Stata 12; StataCorp, 2011).

Mit der freien Statistik-Umgebung "R" steht eine nicht-kommerzielle Lösung für dieses Problem zur Verfügung und es setzt derzeit ein Wandel zu einer stärkeren Verwendung dieser Softwareumgebung ein (Culpepper & Aguinis, 2010; Hubert & Wainer, 2011; R Development Core

[22] Diese Arbeit wurde vorgestellt auf dem 5th UK Rasch Users Day: Böhnke & Lutz, 2011c.

Team, 2010). Dies kann an verschiedenen Indikatoren festgemacht werden. Robert Muenchen präsentiert auf seiner Website (Muenchen, n.d.) ständig erneuerte und ergänzte Informationen zur Nutzung verschiedener Statistiksoftwaren. In ausgezählten Web-Diskussionen und anderen netzbasierten Nutzer-Statistiken sowie mehreren Nutzer-Befragungen belegt R bereits Spitzenpositionen. Auf der Website *kaggle.com*, die Wettbewerbe zwischen Datananalysten sponsert (72000 aktive Analysten), belegt R mit großem Abstand die Spitzenposition (vor Matlab und SAS). In der Zahl der *google scholar* Treffer der Nutzung in wissenschaftlichen Artikeln liegt R noch hinter SAS und SPSS, mittlerweile aber vor Stata, Statistica und S-Plus (alles bezogen auf die vorliegenden Daten zur Verfassung dieses Abschnittes am 12.12.2012).

In den letzten Jahren ist eine Zunahme an praxisorientierter Einführungsliteratur zu dieser Statistik-Umgebung zu verzeichnen (Everitt & Hothorn, 2010; Everitt, 2005; Field, Miles, & Field, 2012; Fox & Weisberg, 2011; Luhmann, 2010) und R hat auch Eingang in einführende Lehrbücher in den Sozialwissenschaften gefunden (Langdridge & Hagger-Johnson, 2009). Erste Universitäten in Deutschland haben in der Psychologie ihre Grundlagenausbildung von kommerziellen Paketen auf R umgestellt (z.B. U Tübingen, FU Berlin). Außerdem gibt es mittlerweile mehrere grafische Nutzeroberflächen und andere Erweiterungen, die den Start mit R ebenfalls erleichtern (einige Beispiele: R Commander, Fox, 2005; RStudio, Allaire, Cheng, Paulson, & DiCristina, n.d.; JGR, Helbig, Theus, & Urbanek, 2005; Deducer, Fellows, 2012). R kann damit zumindest in dem Bereich der Sozialwissenschaften nicht mehr als Spezialistensoftware gesehen werden.

Die Softwareumgebung R hat damit das Potenzial das beschriebene Software-Problem bei der Verwendung von IRT-Modellen zu lösen. Da R eine Freeware ist, sind Anschaffung und Betrieb mit keinerlei finanziellen Kosten verbunden. Das soll nicht darüber hinwegtäuschen, dass die Lernkurve am Anfang steil ist und trotz der neu entwickelten Materialien zeitliche Kosten verursacht. Wenn R den Weg in die Standardausbildung der Studiengänge findet, entfällt dies auch für die Einarbeitung in die IRT-Modelle in R, denn die Sprache R und die Verwendung des Programmes ändern sich nicht. Bei freier Software steht aber immer die Frage im Raum, wie gut die Software tatsächlich ist. Dies ist kein unerhebliches Problem, insbesondere in der Qualitätssicherung in der

Forschung. In den "Statistical Principles for Clinical Trials" (ICH E9, 1998) wird betont, dass die Glaubwürdigkeit statistischer Analyseergebnisse nicht nur auf der Glaubwürdigkeit der verwendeten Daten und Erhebungsmethoden beruht, sondern auch auf der Qualität und Validität der verwendeten Analysesoftware. Und dies unabhängig davon, ob die Software neu entwickelt wurde oder ein Standardsystem ist.

Daher ist die erste Fragestellung der vorliegenden Arbeit, ob die Schätzer von drei in R existierenden IRT-Pakete zu konsistenten Ergebnissen führen – und damit tatsächlich auch die Empfehlung ausgesprochen werden kann, diese zu verwenden. Die Frage dieser Arbeit hat explorativ-evaluativen Charakter: Sind die Pakete zur Schätzung von IRT Modellen in R ähnlich effizient in Bezug auf die Item- und Personenparameter und kann daher eine allgemeine (oder spezifische) Anwendungsempfehlung ausgesprochen werden (siehe Hypothese 1 und 2 unten)?

2.1.2. Stichprobengrößen und Schätzer für IRT Modelle

Der zweite Grund (neben der Kostenproblematik) gegen die Verwendung von IRT-Modellen ist die oft geäußerte Annahme, dass die Stichprobenzahlen nicht groß genug sind, um zu angemessenen Schätzern der Parameter zu kommen (z.B. Zickar & Broadfoot, 2009). In Überblicksarbeiten (Orlando Edelen & Reeve, 2007; Walker et al., 2010) wird festgehalten, dass je nach Itemzahl und Modell mehr als 200 Personen benötigt werden. Rost (2004) empfiehlt für gut passende Items im Rasch-Modell Stichprobenzahlen ab ca. 50 Personen. Da für die Forschungsfrage der Güte der IRT-Programme in R eine Evaluation der Schätzeffizienz der Programme unter verschiedenen Bedingungen nötig ist, d.h. verschiedene Item- und Personenzahlen getestet werden müssen, kann gleichzeitig bestimmt werden, ab welcher Stichprobengröße bei welcher Itemzahl eine Schätzung mit den vorhandenen Methoden als hinreichend genau angesehen werden kann.

Die drei untersuchten Programmroutinen in R zur Schätzung von IRT-Modellen verwenden unterschiedliche Schätzer für die Bestimmung der Personen- und Itemparameter. Zwischen diesen Schätzern ist keine analytische Entscheidung möglich, welcher besser geeignet ist, und bisherige Prüfungen haben keine abschließende Entscheidung gebracht (z.B. Bolt, 2005; J. Cohen, Chan,

Jiang, & Seburn, 2008; Embretson & Reise, 2000; van den Wollenberg, Wierda, & Jansen, 1988). Die folgenden drei Schätzverfahren und die implementierenden Pakete sind:

1) conditional maximum likelihood (eRm; Mair & Hatzinger, 2007a),
2) marginal maximum likelihood (ltm; Rizopoulos, 2006) und
3) joint maximum likelihood (mixRasch; Willse, 2011; weitere Verfahren existieren, s. z.B. Bolt, 2005).

Aus der Literatur gibt es widersprüchliche Hinweise, welches der Verfahren am geeignetsten ist (s.u. bei den jeweiligen Methoden). Das Schätzproblem bei den Item-Response-Modellen entsteht dadurch, dass die Anzahl der Parameter nicht gegen einen festen Wert strebt, wenn die Anzahl der Beobachtungen erhöht wird, und so sind die Ergebnisse von Schätzern nicht erwartungstreu, die sowohl Personen- als auch Itemparameter zusammen schätzen (Kempf, 2003; Neyman & Scott, 1948; Rost, 2004). Das erste Paket, eRm (Hatzinger & Rusch, 2009; Mair & Hatzinger, 2007a), benutzt die conditional maximum likelihood estimation (CML), die auf der Basis der Suffizienz des Summenscores dieses Problem umgeht (Kempf, 2008; Rost, 2004):

$$\boldsymbol{CML}(\mathbf{X}) = \prod_{v=1}^{n} \boldsymbol{prob}\{(x_{v1}, \ldots, x_{vk}) | x_{vo}\} \qquad \textbf{Formel 2-1}$$

Gilt das Rasch-Modell, enthält der Summenscore alle nötige Information über die Ausprägung der latenten Dimension für jede Person. Daher werden bei diesem Schätzer zunächst die Itemparameter geschätzt, während der Score x_{vo} als Angabe der Ausprägung der Fähigkeit verwendet wird. In einem zweiten Schritt werden dann auch die Personenparameter geschätzt. Zu beachten ist, dass das Modell in seiner Gültigkeit immer noch scheitern kann und mit der Annahme im ersten Schritt also nicht die automatische Annahme des Modells einhergeht (für Überprüfungen des Rasch-Modells siehe z.B. Kempf, 2003, 2008; Rost, 2004). Diese Schätzmethode hat insgesamt den Ruf, relativ langsam zu sein, aber die exaktesten Ergebnisse zu liefern (J. Cohen et al., 2008).

Eine zweite Möglichkeit ist es, die Personenparameter durch Werte aus einer Verteilung zu ersetzen. Diese als Marginal Maximum Likelihood (MML) bekannte Schätzmethode wird vom Paket ltm (Rizopoulos, 2006) verwendet. Die Schätzung wird dadurch schneller (verglichen mit CML),

doch verschlechtert sich die Passung des Modells umso mehr, je weniger die empirischen Daten zu der Verteilungsannahme passen (Kempf, 2003, 2008). Auch kann die Schätzung der Parameter nicht mehr als verteilungsfrei angesehen werden (im Vergleich zu CML und JML; DeMars, 2010; Linacre, 2007a). Üblich Annahmen sind die Normalverteilung (so z.B. in dem Paket ltm) oder die zweiparametrige Exponentialverteilung, die eine deutlich größere Flexibilität in der Modellierung der Verteilung gibt (in WINMIRA; Davier, 2000). Die Likelihood-Schätzfunktion dieser Methode ist (Rost, 2004):

$$\boldsymbol{MML}(\mathbf{X}) = \prod_{v=1}^{n} \boldsymbol{prob}\{(\boldsymbol{x_{v1}}, \ldots, \boldsymbol{x_{vk}})|\boldsymbol{\theta}\}\boldsymbol{g}(\boldsymbol{\theta}) \qquad \textbf{Formel 2-2}$$

Die Schätzung der Itemparameter erfolgt bei dieser Methode indem zunächst die Personenparameter auf die Verteilung $g(\theta)$ standardisiert werden und im Anschluss daran z.B. über empirische Bayes-Schätzer (Rizopoulos, 2006; Rost, 2004) die Personenparameter bestimmt werden. Diese Methode liefert asymptotisch dieselben Ergebnisse wie die CML-Schätzung, ist aber davon abhängig, ob die Verteilung $g(\theta)$ korrekt spezifiziert ist. Ist sie es nicht, sind die Schätzer nicht erwartungstreu (Mair & Hatzinger, 2007b; Pfanzagl, 1994).

Als dritte Methode schätzt die Joint Maximum Likelihood (JML) Personen- und Itemparameter zusammen. Diese Methode hat aufgrund des analytischen Befundes (also statistisch-mathematischen Faktums) einen relativ schlechten Ruf (Bolt, 2005; Kempf, 2003; Rost, 2004) und auch die Ergebnisse zur Konsistenz der Parameteschätzung sind gemischt (J. Cohen et al., 2008). Dennoch hat sie mit dem Computerprogramm WINSTEPS zu weiter Verbreitung gefunden (Linacre, 2007a; Wang & Chen, 2005), die sich darin begründet, dass die Methode sehr schnell ist und ebenfalls wie die CML ohne Verteilungsannahmen auskommt (Bond & Fox, 2007; Linacre, 2007a). Diese Methode ist in dem R-Paket "mixRasch" (Willse, 2011) umgesetzt. Sie basiert auf der folgenden Likelihood-Schätzfunktion (Rost, 2004):

$$\boldsymbol{JML}(\mathbf{X}) = \prod_{v=1}^{n} \boldsymbol{prob}\{(\boldsymbol{x_{v1}}, \ldots, \boldsymbol{x_{vk}})\} \qquad \textbf{Formel 2-3}$$

Hier bestehen unterschiedliche Vorgehensweisen zur Schätzung der Parameter. Üblich ist beispielsweise die Verwendung von Transformationen der Randverteilungen der Datenmatrix (Perso-

nen- & Itemsummen), um Startwerte zu generieren, von denen aus dann die Parameter geschätzt werden (Bond & Fox, 2007).

2.1.3. Anwendung von IRT-Modellen in der klinischen Forschung: Bimodalität

Neben den ersten beiden Fragen (Güte der Schätzer und nötige Stichprobengrößen) stellen sich auch praktische Fragen bei der Anwendung von IRT-Modellen in der Forschung. In der klinisch-psychologischen Forschung stellt sich oft die Frage, ob sich Bimodalität der verwendeten Daten auf die Schätzer der Modelle auswirken. Werden Tests auf Normstichproben und belastete Stichproben angepasst, liegen nur in den seltensten Fällen normalverteilte (oder wenigstens unimodal verteilte) Daten vor (Jacobson & Truax, 1991; Kraemer et al., 2003), da bei der dimensionalen Auffassung von psychischer Belastung davon auszugehen ist, dass sich eine oder mehrere Populationen unterschiedlicher Belastungsgrade entlang des gemessenen Kontinuums ausmachen lassen. Teilaufgabe der (Patientenorientierten) Versorgungsforschung und Epidemiologie ist daher die Generierung von Daten, die diese Hypothese prüfen und die Bestimmung von Cut Offs zwischen diesen Populationen ermöglichen (Jacobson & Truax, 1991; Lambert & Ogles, 2009; Tingey et al., 1996). Folgt man diesem Verständnis rein quantitativer Unterschiede zwischen Personen in Belastungsgraden (anderer Standpunkt s. z.B. Böhnke & Lutz, 2008; Kraemer et al., 2003; Schürch et al., 2009), dann folgt daraus, dass ein solcher Test nicht nur dazu geeignet sein sollte, in einer dieser Populationen eingesetzt zu werden. Stattdessen sollte er in beiden Populationen gleich gut messen und auch eine Schätzung des Überganges zwischen beiden Populationen ermöglichen. Die Güte des Schätzers sollte dabei nur wenig von der Verletzung der Annahme der Normalverteilung beeinflusst werden.

Als Beispiele hierfür wird auf einige gebräuchliche Tests zur Bestimmung allgemeiner Symptombelastung eingegangen. Bei der Normierung des "Clinical Outcomes in Routine Evaluation" (CORE, Connell et al., 2007) werden nicht nur Kennwerte der verschiedenen Populationen berichtet, sondern darüber hinaus auch die Verteilungen der verschiedenen Normierungsstichproben grafisch dargestellt. Dabei wird ersichtlich, dass die beiden in dem Paper verwendeten nicht-belasteten Stichproben stark rechtsschief sind. Dieses Ergebnis ist auch klinisch gut nachvollziehbar: Während in belasteten Stichproben neben einem höheren Mittelwert auch eine breite Kombi-

nation von Symptomen und damit der selbstberichteten Belastungen möglich ist, sollte bei einem entsprechend konstruktvaliden Test in der nicht-belasteten Bevölkerung nur eine sehr niedrige Belastung vorliegen, was eine sehr schmale Verteilung mit einem niedrigen Mittelwert impliziert. Da die Personen oft aber nicht danach selegiert werden, ob sie tatsächlich diagnostisch unauffällig sind, sondern in der Regel zwischen Personen in Behandlung und Personen außerhalb Behandlung unterschieden wird, gibt es auch immer wieder Personen, die eigentlich im Versorgungssystem aufgenommen sein müssten, aber die keine Hilfe in Anspruch nehmen oder aber falsch versorgt werden (Wittchen & Jacobi, 2001). Daher gibt es auch in Stichproben aus der "Normalbevölkerung" immer wieder Personen, die hohe Werte aufweisen. Dies sind aber vergleichsweise wenige, wodurch rechtsschiefe Verteilungen entstehen (siehe auch Margraf & Ehlers, 2007 für ein ähnliches Beispiel).

Zumeist präsentieren Testpublikationen lediglich Populationskennwerte wie Mittelwerte und Standardabweichungen. Diese sind basieren zwar auf der Annahme der Normalverteilung, liefern aber wertvolle Informationen zur Gestaltung der Bedingungen für die Überprüfung der Anfälligkeit der Schätzer für die Verteilungsart. Tabelle 2-1 präsentiert Kennwerte für drei gebräuchliche Maße psychischer Belastung. Die auf die Streuung der klinischen Stichprobe standardisierten Effektstärken sind ebenfalls in Tabelle 2-1 angegeben. Da in dieser Studie durch eine Simulation geprüft werden soll, wie stark die Schätzer der Programme von Verletzungen der Normalverteilung abhängen, kann aus diesen Werten abgeleitet werden, was ein plausibles Maß für den Unterschied zwischen zwei Populationen sein könnte, die klinische und nicht-klinische Daten repräsentieren sollen. Die niedrigste Effektstärke für den Unterschied zwischen den Stichproben beträgt in den in Tabelle 2-1 zitierten Studien $d = 1.70$. Da dies bereits einen sehr großen Unterschied zwischen den Stichproben darstellt (J. Cohen, 1988), wird dieser Wert zur Simulation unterschiedlicher Stichproben, die klinische und nicht-klinische Fälle repräsentieren sollen, verwendet. Dieser Wert erhebt keinen Anspruch darauf, repräsentativ für die Unterschiede zwischen klinischen und nicht-klinischen Populationen in psychometrischen Erhebungsinstrumenten zu sein, doch liefert er einen angemes-

senen Wert zum Vergleich mit simulierten Daten, in denen keine Unterschiede zwischen Populationen vorliegen[23].

Tabelle 2-1: Verteilungskennwerte von drei üblichen Instrumenten zur Messung psychischer Belastung aus den jeweiligen Normierungspublikationen.

Test (Publikation)	Klinische Stichprobe *M* (*SD*)	Nicht-klinische Stichprobe *M* (*SD*)	*d*
CORE (Connell et al., 2007)	18.3 (7.1)	4.8 (4.3)	1.90
engl. Version des OQ-45 (Lambert et al., 1996)	83.09 (22.23)	45.19 (18.57)	1.70
FEP (Lutz, Schürch, et al., 2009)	3.0 (.63)	1.9 (.46)	1.75

Anmerkung: *d* standardisiert auf die klinische Stichprobe; CORE = Clinical Outcomes in Routine Evaluation – Outcome Measure; OQ = Outcome Questionnaire; FEP = Fragebogen zur Evaluation von Psychotherapieverläufen; CORE & FEP nur mit einer Nachkommastelle berichtet; Zahl der Dezimalstellen richtet sich nach der Originalpublikation

2.1.4. Fragestellungen und Ziele der Studie

Die zentrale Frage dieser Studie ist, ob die Pakete zur Schätzung von IRT Modellen in R ähnlich effizient in Bezug auf die Item- und Personenparameter sind und ob eine allgemeine (oder spezifische) Anwendungsempfehlung ausgesprochen werden kann. Dies wird in den zwei Fragestellungen überprüft. Zunächst wird getestet, wie groß eine Stichprobe sein muss, um zu reliablen Schätzern der Item- und Personenparameter zu kommen. Diese Frage ist relativ zu beantworten (s. z.B. Cohen et al., 2008), indem die Schätzer der drei R-Pakete miteinander verglichen werden.

> *Fragestellung 1: Es wird vermutet, dass alle drei Pakete insgesamt zu strukturerhaltenden Schätzern kommen. Dennoch sollte CML die relativ besten Ergebnisse liefert, MML die zweitbesten und JML die relativ ineffizientesten.*

[23] Die Effektstärken von Instrumenten zur Ergebnismessung in der Psychotherapie hängen von Instrument und Population ab. Für den "Shorter Psychotherapy and Counselling Evaluation" (Halstead, Leach, & Rust, 2007) werden noch größere Unterschiede berichtet ($d = 2.96$); aus den Prozentrangtabellen für das Beck Anxiety Inventory und dem Fragebogen zu körperbezogenen Ängsten, Kognitionen und Vermeidung rekonstruierte Effektstärken liegen bei $d = 1.56$ bzw. $d = 1.28$ (Ehlers & Margraf, 2001; Margraf & Ehlers, 2007); für das Brief Symptom Inventory (Franke, 2000) finden sich Effektstärken von $d = 1.24$ in der Normstichprobe und von $d = 1.37$ im TK-Modellvorhaben (Lutz, Böhnke, Köck, & Bittermann, 2011).

Danach wird die Unabhängigkeit der Schätzgenauigkeit der Programme von der Stichprobenverteilung geprüft. Diese hat direkte Konsequenzen für die Empfehlung der Nutzung dieser Programme in der Forschung. Sollten sich die vermuteten Zusammenhänge zeigen, wären R-Pakte in der Praxisanwendung zu bevorzugen, die immer die relativ höchste Genauigkeit haben.

> *Fragestellung 2: Die verteilungsunabhängigen Schätzer der JML und CML-Methode sind in ihrer Schätzgenauigkeit weniger von der Verteilung der Daten beeinflusst als die MML-Schätzung.*

Aus beiden Hypothesen lässt sich zusammen die Frage beantworten, welches der Pakete sich unter den verwendeten Bedingungen am ehesten für den Einsatz in der Forschung eignet: Das Paket, das relativ die besten Schätzer liefert und am wenigsten von den Bedingungen (Stichprobengröße und Bimodalität) beeinflusst wird.

2.2. Methode

2.2.1. Monte Carlo Experiment

Zur Untersuchung beider Hypothesen wird eine Monte Carlo Studie durchgeführt. Eine Monte Carlo Studie ist ein Experiment, in dem nach den Vorgaben eines statistischen Modells Daten generiert und dann mit demselben oder anderen statistischen Modellen ausgewertet werden (Harwell, Stone, Hsu, & Kirisci, 1996; Harwell, 1997). So ist es in diesem Fall möglich, verschiedene Schätzmethoden zu vergleichen. Für die Untersuchung solcher Fragen haben Monte Carlo Studien bereits eine lange Tradition (Psychometric Society, 1979; Yen, 1987). Sie sind besonders dann sinnvoll, wenn a) der Vergleich nicht analytisch geschehen kann oder b) wenn verschiedene Algorithmen oder Schätzer verglichen werden sollen, die dieselbe Funktion erfüllen (Psychometric Society, 1979). Monte Carlo Studien haben gegenüber dem Vergleich verschiedener Programme an natürlichen Datensätzen zwei Vorteile. Zum Einen ist bei natürlichen Datensätzen der Prozess unbekannt, unter dem die Daten zustande gekommen sind. Das bedeutet, dass z.B. die Anwendung des Rasch-Modells auf die Daten falsch sein könnte und damit eigentlich keine Rückschlüsse auf die Verwendbarkeit der Programme möglich sind. Zum Anderen ist nicht nur *ein* Vergleich an

tatsächlich unter Bedingungen des Modells zustande gekommenen Daten möglich, sondern durch die wiederholte Datengenerierung wird der teststatistische Vergleich von Bedingungen möglich. In der Studie werden die folgenden Parameter variiert:

- **Stichprobengröße**: Es werden Stichproben mit N = 100, 250, 500, und 1000 verwendet;
- **Itemzahl:** Es werden k = 10, 25 und 50 dichotome Items pro Test simuliert, die gleichmäßig über das latent Kontinuum zwischen δ_{min} = -2 bis δ_{max} = 2 verteilt sind;
- **Modalität der Verteilungen:** Die Personenparameter, die zur Berechnung der Antwortmatritzen simuliert werden, stammen in der Bedingung ("unimodal") einer Normalverteilung, $\theta_G \sim$ NV(0,1); in der anderen Bedingung ("bimodal") wird die Hälfte der Personenparameter aus $\theta_1 \sim$ NV(-.85,1) und die andere Hälfte aus $\theta_2 \sim$ NV(.85,1) simuliert. Der Unterschied in der bimodalen Bedingung entspricht damit einem d = 1.7 (s.o.).

Die Stichprobengrößen bewegen sich im Rahmen dessen, was in der Validierungsforschung bei IRT-Modellen verwendet wird. Ein N = 100 wird in der Regel als zu gering angesehen (Ausnahmen z.B. Linacre, 1994; Rost, 2004), und stellt für die Forschung im klinischen Kontext eine relevante Größe dar, da Patientengruppen in dieser Größe relativ schnell erhebbar sind. Die restlichen Stichprobengrößen sind auch in anderen Untersuchungen dieser Art realisiert und N = 1000 kann als ein Konsenswert gesehen werden, ab der das Rasch Modell problemlos geschätzt werden können sollte (Hidalgo & López-Pina, 2011; Orlando Edelen & Reeve, 2007; Walker et al., 2010).

Die Anzahl der Items orientiert sich ebenfalls an diesen Überlegungen. Ein Test mit k = 10 dichotomen Items ist eher zu kurz für eine genauen Messung eines Konstruktes und k = 50 sind vermutlich genügend (Raîche, Blais, & Magis, 2007). Die Items wurden gleichmäßig auf δ_{min} = -2 bis δ_{max} = 2 verteilt, d.h. das leichteste und das schwierigste Item lagen exakt auf diesen Grenzen und die restlichen in gleichem Abstand dazwischen (ähnliche Simulationsbedingungen: J. Cohen et al., 2008; Hidalgo & López-Pina, 2011; Pina & Montesinos, 2005; Willse, 2011).

Die obere und untere Grenze wurden so bestimmt, dass eine Person v, die einen Personenparameter entsprechend dem Mittelwert der simulierten Verteilungen (θ_v = 0) hat, eine Lösungswahrscheinlichkeit von $p_{vi:\,\delta = -2}$ = .88 bzw. $p_{vi:\,\delta = 2}$ = .12 für das leichteste bzw. schwierigste Item

erreicht. Für die Schätzung von Rasch-Modellen werden für die Kategorienanteile in den Items Werte von mindestens .10 empfohlen (10% Nichtlösende bzw. nicht mehr als 90% Lösende; Holman, Lindeboom, Glas, Vermeulen, & de Haan, 2003; Linacre, 2007), was durch die Wahl der Simulationsparameter hier im Mittel gewährleistet wird.

Die Daten wurden mit einem erprobten Algorithmus simuliert (Linacre, 2007b; Pina & Montesinos, 2005; Willse, 2011). Bei diesem Algorithmus wird aufgrund von vorgegebenen Item- und Personenparametern eine Antwortmatrix unter Geltung des Modells simuliert. Die Vorgabe der Itemparameter wurde bereits beschrieben genauso wie die Verteilungen, aus denen die Personenparameter generiert werden. Basierend darauf sind die folgenden Schritte nötig:

1. Gemäß den Item- und Personenparametern wird die Wahrscheinlichkeit berechnet, mit der Person *v* dieses Item *i* nicht lösen würde ($P(fail)_{vi}$);
2. dann wird für jede Person-Itemkombination eine Zufallszahl aus [0, 1] generiert (u_{vi});
3. ist die generierte Zufallszahl größer als die Wahrscheinlichkeit, dass die Person das Item nicht lösen würde [$P(fail)_{vi} < u_{vi}$], dann wird davon ausgegangen, dass das Item gelöst wurde ("1"); sonst wird kodiert, dass das Item nicht gelöst wurde ("0").

Es werden pro Bedingung des Designs $b = 1000$ Samples realisiert. Die Zahl dieser Simulationsstichproben befindet sich in einem üblichen Bereich für eine Untersuchung von Verteilungsmerkmalen (etwa $b = 500$ Durchläufe gelten als ausreichend, z.B. Harwell et al., 1996).

2.2.2. Parameterrekonstruktion

Diese Arbeit ist vorrangig daran interessiert, ob die gewählten Verfahren dazu in der Lage sind, vorhandene Datenstrukturen hinreichend zu rekonstruieren. Wichtig für eine sinnvolle Evaluation der Passung der geschätzten Modelle, sind Evaluationskriterien, die eine Bewertung der einzelnen Methoden sowie einen Vergleich zwischen ihnen ermöglichen. Diese Ergebnis-Kriterien werden dann über die Zellen des Simulationsdesigns verglichen. Diese Zellen sind Itemzahl (drei Stufen: 10, 25, 50) X Anzahl der Personen im Schätzsample (vier Stufen: 100, 250, 500, 1000) X Modalität (zwei Stufen: unimodal, bimodal). Dies entspricht einem Design mit 24 Zellen plus den drei Paketen als messwiederholtem Faktor in jedem simulierten Datenset.

Bei IRT Modellen werden sowohl die Itemparameter, die die Schwierigkeit der einzelnen Items erfassen, wie auch die Personenparameter, die die Ausprägung der Personen auf der latenten Eigenschaft darstellen, in derselben Metrik erfasst (Doucette & Wolf, 2009; Hambleton et al., 1991). Für den Bereich der Messung psychischer Belastung korrespondieren diese beiden Parameter mit dem Belastungsgrad, den ein Item optimal erfasst (Itemparameter), und dem Belastungsgrad, den ein Patient aufweist (Personenparameter; Doucette & Wolf, 2009; Hays, Morales, & Reise, 2000; Reise & Haviland, 2005; Schürch et al., 2009; Wirtz & Böcker, 2007). Beide Parameter sind für die Verwendung der Ergebnisse von IRT Modellen in der Psychotherapieforschung von großer Wichtigkeit. Wenn ein Test als IRT-skaliert gilt, werden diese Parameter weitergegeben: Sie stellen damit einen zentralen Aspekt des Normierungsprozesses dar. Daher wird die Genauigkeit der Schätzung beider Parametergruppen im Folgenden überprüft.

Die Auswahl der Prüfstatistiken zur Evaluation ist darauf ausgerichtet, mit parametrischen und nicht-parametrischen Verfahren zu prüfen, ob die Rekonstruktion beider Parametergruppen gelingt. Hierzu eignen sich besonders korrelative Maße, die die Nähe zwischen den theoretischen und den geschätzten Parametern evaluieren können. Im Folgenden werden die herangezogenen Bewertungsparameter beschrieben (s. z.B. J. Cohen et al., 2008; Willse, 2011).

Personenparameter: Die "wahren" Personenparameter, die zur Simulation verwendet wurden, wurden als der jeweils zu rekonstruierende Parameter verwendet. Das erste Kriterium zur Bewertung der Güte der Rekonstruktion ist die Korrelation der geschätzten Parameter mit den echten Parametern, da sie erfasst, ob die Personen genauso geordnet werden wie sie simuliert wurden (vom niedrigsten bis zum höchsten Belastungsgrad). Das zweite Kriterium ist die mittlere Summe der individuellen Abweichungen zwischen echten und geschätzten Parametern. Dieser Abstand zwischen den wahren und den geschätzten Parametern wird durch den Root Mean Square Error (RMSE) gemessen (siehe z.B. J. Cohen et al., 2008):

$$\boldsymbol{RMSE} = \frac{\mathbf{1}}{\boldsymbol{r}} \sum_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{r}} \left(\sqrt{\frac{\mathbf{1}}{\boldsymbol{k}-\mathbf{1}} \left(\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{k}} \boldsymbol{\delta}_{\boldsymbol{bk}} - \boldsymbol{\delta}_{\boldsymbol{b0}} \right)^{\mathbf{2}}} \right) \qquad \textbf{Formel 2-4}$$

wobei b (1,...,*r*) die simulierten Datensets und *i* die Items 1,...,*k* bezeichnet; δ_{bk} bezeichnet den geschätzten Itemparamter des Items *k* in Datenset *b* und δ_{b0} den entsprechenden wahren Wert, der für die Simulation verwendet wurde. Der RMSE quantifiziert also die mittlere Abweichung der geschätzten Werte von den simulierten "wahren" Werten.

Itemparameter: Da die Betrachtung der Items und Personen in den IRT-Modellen symmetrisch ist, stellt sich dieselbe Frage wie bei den Personenparametern: Wie gut werden die Itemparameter geschätzt? Auch hier werden wie bei den Personenparametern zunächst die Korrelationen zwischen den simulierten und wahren Itemparametern ausgewertet. Als Zweites wird der Abstand zwischen den wahren und den geschätzten Parametern wiederum durch den "Root Mean Square Error" (RMSE) erfasst (z.B. Cohen et al., 2008 im selben Kontext):

$$\boldsymbol{RMSE = \frac{1}{N}\sqrt{\sum(\theta_{b0} - \theta_{bv})^2}} \qquad \textbf{Formel 2-5}$$

wobei *b* (1,...,*r*) die simulierten Datensets bezeichnet und θ die Personenparameter 1,...,*N*; θ_{bv} den geschätzten Personenparameter der Person *v* in Datenset *b* und θ_{v0} den jew. wahren Wert, der für die Simulation verwendet wurde (siehe auch oben).

Modellgeltungstest: Für das Rasch-Modell besteht bei Schätzung mit der CML-Methode die Möglichkeit eines Modellgeltungstestes, der sog. Andersen Test (Andersen, 1973). Für den Andersen Test wird die Stichprobe, an der das Rasch-Modell geschätzt wurde, in zwei (oder mehr) Teile geteilt. In diesen wird das Rasch-Modell getrennt bestimmt und die logarithmierte Likelihood über alle Substichproben hinweg aggregiert. Diese wird dann verglichen mit der logarithmierten Likelihood, die in der Gesamtstichprobe erreicht wird. Dies geschieht über einen Likelihood-Quotienten-Test: Gilt das Rasch-Modell, sollte in den Substichproben keine bessere Anpassung an die Daten gelingen als im Gesamtmodell, d.h. der Vergleich sollte nicht-signifikant ausfallen. Die Teststatistik berechnet sich aus:

$$\boldsymbol{\chi^2 = -2 * \left[ln(\text{Likelihood(Gesamt)}) + \sum_1^g \ln\left(\text{Likelihood(Subgruppe}_g)\right)\right]} \qquad \textbf{Formel 2-6}$$

und ist bei CML-Schätzung mit (*g*-1)(*k*-1) Freiheitsgraden χ^2-verteilt (*g* ist die Anzahl der Subgruppen; *k* die Anzahl der verwendeten Items). Für alle simulierten Stichproben wird dieser Test

angewendet, um die Korrektheit seines α-Fehler-Niveaus zu prüfen. Bei den Stichproben, die aus einer Verteilung simuliert werden, wird dazu ein zufälliger Split in zwei Hälften vorgenommen. Bei den Stichproben, die in der bimodalen Bedingung simuliert werden, wird die Zugehörigkeit zur Verteilung als Teilungskriterium verwendet.

2.2.3. Programme & Schätzer

In dieser Arbeit kommen drei in R (R Development Core Team, 2010) implementierte Routinen zur Schätzung von Rasch-Modellen zum Tragen. Neben der Variation von Programmen geht dies auch mit der Variation von Schätzmethoden einher. Generell sind die drei oben beschriebenen Schätzmethoden etabliert und in R abrufbar. Da sie bereits in Kapitel 2.1.2 beschrieben wurden, seien hier nur noch einmal die Namen, Abkürzungen und Pakete genannt: Conditional Maximum Likelihood (CML; eRm; Mair & Hatzinger, 2007), Marginal Maximum Likelihood (MML; ltm; Rizopoulos, 2006) und Joint Maximum Likelihood (JML; mixRasch; Willse, 2011).

Zur deskriptiven Beurteilung der relativen Güte der drei Programme werden am Ende Ränge innerhalb der zwei Bewertungskategorien (Güte Personen-/ Itemparameter) gebildet und danach diese Ränge gemittelt. Der niedrigste Rangplatz zeigt dabei das über alle Kriterien am Besten geeignete Programm an.

2.3. Ergebnisse

Zunächst werden die Kennwerte der simulierten Stichproben berichtet (2.3.1) und verschiedene Untersuchungen zur Güte der Skala durchgeführt (Reliabilität, Informationsfunktion, Andersen Test; 2.3.2). Daran anschließend folgen die Ergebnisse zum Vergleich der Schätzmethoden und Stichprobengrößen. Diese werden erst für die Itemparameter berichtet (2.3.3) und dann für die Personenparameter (2.3.5).

2.3.1. Deskriptive Daten für die simulierten Stichproben

Tabelle 2-2 und Tabelle 2-3 präsentieren die Statistiken für die simulierten Personenparameter im uni- bzw. bimodalen Fall. Dabei ist deutlich an den Standardabweichungen der simulierten Werte als auch den Mittelwerten zu erkennen, dass die Simulation in dieser Hinsicht das ge-

wünschte Ergebnis lieferte, die Verteilungen der Personenparameter entsprachen den vorgegebenen Parameterwerten.

Tabelle 2-2: Personenparameter der simulierten Stichproben im unimodalen Fall.

	Mittelwert der Mittelwerte	*SD* der Mittelwerte	Median der Mittelwerte	Mittlere *SD* der Durchläufe	*SD* der simulierten *SD*s	Median der *SD*s	Anzahl Stichproben (well-conditioned)[24]
für *k* = 50							
N = 1000	.001	.03	.001	1.00	.02	1.0002	1000
N = 500	.0002	.05	.0003	.9997	.03	.9999	1000
N = 250	-.002	.06	-.002	.995	.05	.994	999
N = 100	.003	.10	.003	.997	.07	.997	982
für *k* = 25							
N = 1000	.002	.03	.0008	1.0002	.02	.9991	1000
N = 500	-.0002	.04	.00	.9993	.03	.999	1000
N = 250	.00	.06	.00	1.001	.04	1.002	1000
N = 100	-.0004	.10	.001	.996	.07	.997	985
für *k* = 10							
N = 1000	-.0004	.03	-.0001	1.00	.02	.9997	1000
N = 500	-.0006	.04	-.002	1.001	.03	1.0004	1000
N = 250	.001	.06	.0007	1.001	.04	1.001	1000
N = 100	.00	.10	-.003	.997	.07	.999	988

Anmerkung. *k* = Itemzahl, *N* = Anzahl simulierter Personen; die Zahl der Nachkommastellen richtet sich nach der Anzahl nötiger Stellen, damit nicht "0" bzw. "1" angegeben werden, um die zufällige Struktur der Daten zu dokumentieren; wenn mehr als vier Nachkommastellen nötig wären, werden zwei angegeben.

[24] Unterschiedliche *N*'s in den Bedingungen kommen dadurch zustande, dass nur "well-conditioned matrices" (Fischer, 1981) ausgewertet wurden. Lösung und Nicht-Lösung der besonders leichten/ schweren Items sind seltene Ereignisse, was zur Konsequenz hat, dass die entstehenden Datenmatritzen insbesondere dann, wenn die simulierte Personenvektorzahl klein ist, "ill-conditioned" sein können (Fischer, 1981; Kubinger & Draxler, 2007). Eine Datenmatrix ist "ill-conditioned" wenn nicht jedes Item mindestens von einer Person gelöst bzw. nicht gelöst wurde. Matritzen, die diese Eigenschaft aufweisen werden von der Auswertung ausgeschlossen. Für die Studie selber ist die einzige Auswirkung eine reduzierte Power bei einigen Zellvergleichen, da aber pro Bedingung 1000 Durchläufe realisiert werden, hat dies keine große Bedeutung.

Tabelle 2-3: Personenparameter der simulierten Stichproben im bimodalen Fall.

	Mittelwert der Mittel-werte	*SD* der Mittelwerte	Median der Mittelwerte	Mittlere *SD* der Durch-läufe	*SD* der simulierten *SD*s	Median der *SD*s	Anzahl Stichproben (well-con-ditioned)
für *k* = 50							
N = 1000	-.001	.03	.0001	1.312	.03	1.313	1000
a1	-.851	.04	-.852	1.00	.03	1.00	
a2	.849	.04	.848	.998	.03	.998	
N = 500	-.001	.04	.0003	1.313	.04	1.312	1000
a1	-.852	.06	-.851	.999	.04	.999	
a2	.85	.06	.852	1.00	.05	.998	
N = 250	.001	.06	.001	1.312	.05	1.313	999
a1	-.852	.09	-.852	.993	.07	.994	
a2	.854	.09	.855	.997	.07	.995	
N = 100	-.002	.10	.00	1.308	.08	1.309	511
a1	-.83	.14	-.84	1.006	.10	1.004	
a2	.82	.14	.83	1.008	.10	1.006	
für k = 25							
N = 1000	-.0008	.03	-.0005	1.313	.03	1.314	1000
a1	-.85	.05	-.852	1.00	.03	1.00	
a2	.849	.05	.848	1.00	.03	.999	
N = 500	.0007	.04	.00	1.311	.04	1.312	1000
a1	-.848	.06	-.849	.999	.05	.998	
a2	.849	.06	.852	.998	.05	.998	
N = 250	-.002	.07	.0004	1.312	.05	1.312	999
a1	-.850	.09	-.849	.996	.07	.995	
a2	.846	.09	.850	1.004	.06	1.005	
N = 100	.004	.10	.0003	1.303	.08	1.305	701
a1	-.826	.14	-.83	.997	.10	.995	
a2	.835	.14	.83	1.002	.10	1.004	
für k = 10							
N = 1000	.00	.03	.0003	1.313	.03	1.314	1000
a1	-.851	.05	-.850	1.00	.03	1.00	
a2	.851	.05	.852	.999	.03	.999	
N = 500	-.0002	.05	.0003	1.312	.04	1.311	1000
a1	-.848	.07	-.848	.998	.05	.998	
a2	.847	.07	.848	1.002	.04	1.001	

Fortsetzung von Tabelle 2-3:

	Mittelwert der Mittelwerte	*SD* der Mittelwerte	Median der Mittelwerte	Mittlere *SD* der Durchläufe	*SD* der simulierten *SD*s	Median der *SD*s	Anzahl Stichproben (well-conditioned)
N = 250	.00	.06	.003	1.313	.06	1.312	1000
a1	-.852	.09	-.851	.998	.07	.996	
a2	.852	.09	.854	.996	.06	.996	
N = 100	.006	.10	.001	1.309	.08	1.307	807
a1	-.837	.14	-.839	.996	.10	1.00	
a2	.848	.14	.849	.998	.10	.998	

Anmerkung. *k* = Itemzahl, *N* = Anzahl simulierter Personen; in der Zeile "N" steht immer die Anzahl der Personen in der Gesamtstichprobe und die deskriptive Statistiken für diese Gesamtstichprobe; diese sind zur Hälfte auf die Substichproben a1 und a2 aufgeteilt, deren Verteilungswerte in den zwei Folgezeilen angegeben sind; die Zahl der Nachkommastellen richtet sich nach der Anzahl nötiger Stellen, damit nicht "0" bzw. "1" angegeben werden um die zufällige Struktur der Daten zu präsentieren; wenn mehr als vier Nachkommastellen nötig wären, werden zwei angegeben.

2.3.2. Untersuchungen zur Güte der Skala

Die Reliabilität der simulierten Daten wurde für jede Stichprobe mittels Kuder-Richardson-20 bestimmt (Holman et al., 2005; Kuder & Richardson, 1937). Die Ergebnisse werden getrennt für den unimodalen (Tabelle 2-4) und den bimodalen (Tabelle 2-5) Fall dargestellt.

Tabelle 2-4: Reliabilitäten (Kuder-Richardson-20) für den unimodalen Fall; Konfidenzintervalle geben Bootstrap-Perzentile aus den simulierten Stichproben an.

	k = 50	*k* = 25	*k* = 10
N = 1000			
Kuder-Richardson-20	.89	.81	.62
[2.5%, 97.5%]	.88; .90	.79; .82	.58; .65
N = 500			
Kuder-Richardson-20	.89	.81	0.62
[2.5%, 97.5%]	.88; .91	.78; .83	.57; .67
N = 250			
Kuder-Richardson-20	.89	.81	.62
[2.5%, 97.5%]	.87; .91	.77; .84	.55; .68
N = 100			
Kuder-Richardson-20	.89	.80	.61
[2.5%, 97.5%]	.86; .92	.74; .85	.48; .70

Tabelle 2-5: Reliabilitäten (Kuder-Richardson-20) für den unimodalen Fall; Konfidenzintervalle geben Bootstrap-Perzentile aus den simulierten Stichproben an.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
Kuder-Richardson-20	.93	.87	.73
[2.5%, 97.5%]	.93; .94	.86; .88	.70; 0.75
$N = 500$			
Kuder-Richardson-20	.93	.87	.73
[2.5%, 97.5%]	.93; .94	.86; .89	.69; .76
$N = 250$			
Kuder-Richardson-20	.93	.87	.73
[2.5%, 97.5%]	.92; .94	.85; .89	.68; .77
N = 100			
Kuder-Richardson-20	.93	.87	.73
[2.5%, 97.5%]	.91; .95	.84; .90	.64; .80

Aus den Mittelwerten für den unimodalen Fall ist erkennbar, dass mit einer Skala von knapp zehn dichotomen Items nach Klassischer Testtheorie nur in seltenen Fällen eine Skala mit akzeptabler Messgenauigkeit konstruiert werden kann. Selbst dann, wenn die Items wie in dieser Simulation einer breiten Verteilung auf einem latenten Kontinuum im Sinne des Rasch-Modells folgen (δ_{min} = -2 bis δ_{max} = 2). Als akzeptabel werden in der Literatur Werte ab .7 bezeichnet (Nunnally, 1978); von exzellenter Messgenauigkeit wird etwa ab .9 (Fliege et al., 2005) gesprochen. Mit diesen Referenzwerten ist deutlich zu erkennen, dass bei einem Test aus dichotomen Items, der dem Rasch-Modell gehorcht, eine solch hohe Messqualität im Mittel mit 50 Items erreicht wird (s. Konfidenzintervalle), doch Tests mit lediglich 10 dichotomen Items weit darunter bleiben.

Im bimodalen Fall ist die Verteilung des Merkmals breiter als im unimodalen Fall (vgl. Tabelle 2-2 und Tabelle 2-3), was sich direkt auf die Reliabilitäten auswirkt. Da die Fehlervarianz im unimodalen und bimodalen Fall der simulierten Daten als gleich angesehen werden kann (es wird derselbe Fehlermechanismus verwendet), die Varianz des Merkmals aber größer wird, muss die Reliabilität steigen (Kempf, 2003). In diesem Fall erreicht auch die kurze Skala mit 10 Items bereits eine akzeptable Messgenauigkeit und die Skala mit k = 25 Items zumindest im Mittel (s. Konfidenzintervalle) bereits eine exzellente Passung. Dies unterstreicht noch einmal, wie wichtig

in der Klassischen Testtheorie eine ausreichend große Variation des Merkmals in der untersuchten Stichprobe ist, um zu einer Aussage zu kommen, wie gut der Test misst (Junker & Sijtsma, 2000).

Auch die IRT bietet eine Möglichkeit, die Messgenauigkeit einer Skala zu untersuchen. Hier wird die sog. Informationsfunktion bestimmt, die ermittelt, wie groß der Beitrag der Items an einem bestimmten Punkt des latenten Kontinuums zur Differenzierung zwischen den Personen ist (Rost, 2004). Der Standardfehler ist der reziproke Wert der Informationsfunktion an der Stelle auf dem latenten Kontinuum des jeweiligen Personenparameters θ_v (siehe für weitere Details auch Kapitel 3). Für einen festgelegten Bereich auf dem latenten Kontinuum, kann die Fläche unter dieser Funktion als ein Maß der mittleren Messgenauigkeit verwendet werden. Dieses Maß ist zwar dimensionslos und gibt eine der Reliabilität nur bedingt vergleichbare Information. Dennoch kann diese Statistik zum Vergleich von verschiedenen Testformen herangezogen werden. Abbildung 2-1 zeigt für die drei in der Monte Carlo-Studie realisierten Itemzahlen bei $N = 500$ den Verlauf der Informationsfunktion sowie Kennwerte der Reliabilität nach Klassischen Testtheorie. Diese lassen sich durch die folgenden Beziehungen annähern (Raîche et al., 2007; Reeve & Fayers, 2005). Der Standardmessfehler ist definiert als:

$$\boldsymbol{SEM = s_{obs}(1 - \rho_{xx})^{1/2}} \qquad \textbf{Formel 2-7}$$

Unter der Annahme, dass die Streuung der standardisierten Skala $s_{obs} = 1$ (Raîche et al., 2007) beträgt, kann für verschiedene Reliabilitätswerte ρ_{xx} der SEM berechnet werden. Dieser steht in folgender Beziehung zur Informationsfunktion:

$$\boldsymbol{I(\theta) = \frac{1}{[SEM(\theta)]^2}} \qquad \textbf{Formel 2-8}$$

Damit ist eine Umrechnung von der Metrik in die andere möglich. Zu beachten ist hierbei, dass der Standardmessfehler bei der Informationsfunktion von der Ausprägung der latenten Variable abhängt und somit als Funktion gezeichnet werden muss. Dort wo diese Funktion den berechneten Wert übersteigt, kann gesagt werden, dass der korrespondierende Reliabilitätswert überschritten wurde. Deutlich ist zu erkennen, dass mit 10 dichotomen Items insgesamt keine befriedigende Messgenauigkeit erlangt werden kann: Über das ganze Kontinuum hinweg liegt die Reliabilität

unter einem Wert von .7. Bei 25 Items wird dieser Wert im mittleren Bereich der Skala überschritten. Erst bei 50 Items zeigt sich eine angemessen hohe Reliabilität über einen breiten Fähigkeitsbereich, doch Werte von .9 überschreitet die Funktion auch hier nicht.

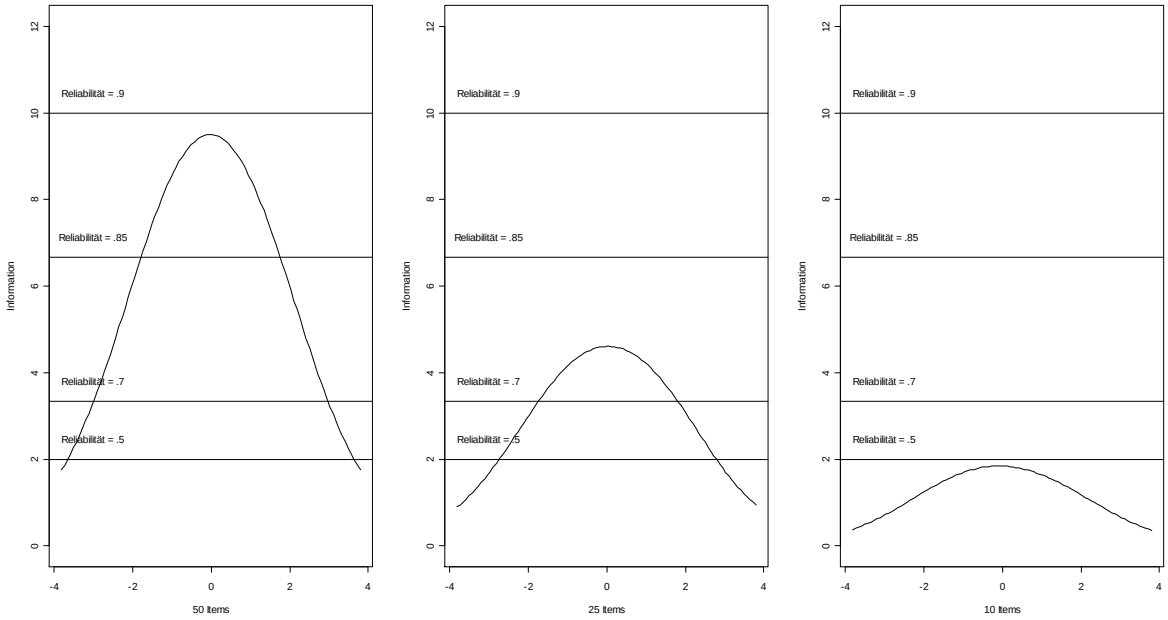

Abbildung 2-1: Informationsfunktionen für die drei Simulationsbedingungen anhand des Beispiels *N* = 500; eingetragen sind auch geschätzte Reliabilitäten nach Klassischer Testtheorie (Raîche et al., 2007) im Bereich -4 bis 4.

In den folgenden Tabellen (unimodal: Tabelle 2-6; bimodal: Tabelle 2-7) sind die Größen der Flächen für den Skalenberich -4 bis 4 der latenten Variable eingetragen (bestimmt mit ltm; Rizopoulos, 2006). Dies ist der gesamte Bereich, in dem plausiblerweise ein Messergebnis unter den gegebenen Simulationsbedingungen liegen kann. Über beide Tabellen hinweg (unimodal, bimodal) ist deutlich der Einfluss der Itemzahl zu erkennen. Über den Messbereich der Skala kommt es fast zu einer Verfünffachung der Fläche unter der Kurve, wenn die Testlänge um denselben Faktor erhöht wird. Die zweite auffällige Tatsache beim Vergleich der beiden Tabellen ist, dass es unter IRT-Gesichtspunkten keinen Unterschied macht, wie breit die latente Eigenschaft verteilt ist, sondern dass in den Zellen der bimodalen Stichprobe nahezu exakt dieselben Flächen unter der Kurve realisiert werden. Diese sog. Stichprobenunabhängigkeit der IRT-Modelle stellt einen weiteren Aspekt dar, in dem die IRT der Klassischen Testtheorie überlegen ist.

Tabelle 2-6: Mittlere Fläche unter der Informationsfunktion im Bereich von -4 bis 4 auf der latenten Dimension für den unimodalen Fall; Konfidenzintervalle geben die Perzentile aus den simulierten Stichproben wieder.

	$k = 50$	$k = 25$	$k = 10$
N = 1000			
Informationsfunktion [-4; 4]	46.81	23.37	9.37
[2.5%. 97.5%]	46.74; 46.90	23.32; 23.42	9.28; 9.34
N = 500			
Informationsfunktion [-4; 4]	46.80	23.37	9.30
[2.5%. 97.5%]	46.71; 46.89	23.30; 23.43	9.25; 9.35
N = 250			
Informationsfunktion [-4; 4]	46.77	23.35	9.30
[2.5%. 97.5%]	46.62; 46.91	23.25; 23.45	9.23; 9.36
N = 100			
Informationsfunktion [-4; 4]	46.68	23.31	9.28
[2.5%. 97.5%]	46.43; 49.90	23.12; 23.47	9.13; 9.38

Tabelle 2-7: Mittlere Fläche unter der Informationsfunktion im Bereich von -4 bis 4 auf der latenten Dimension für den bimodalen Fall; Konfidenzintervalle geben die Perzentile aus den simulierten Stichproben wieder.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
Informationsfunktion [-4; 4]	46.90	23.44	9.36
[2.5%. 97.5%]	46.83; 46.96	23.90; 23.49	9.33; 9.39
$N = 500$			
Informationsfunktion [-4; 4]	46.88	23.49	9.35
[2.5%. 97.5%]	46.79; 46.97	23.38; 23.50	9.31; 9.39
$N = 250$			
Informationsfunktion [-4; 4]	46.86	23.42	9.35
[2.5%. 97.5%]	46.73; 46.99	23.33; 23.51	9.29; 9.41
$N = 100$			
Informationsfunktion [-4; 4]	46.79	23.39	9.34
[2.5%. 97.5%]	46.53; 47.01	23.22; 23.54	9.23; 9.42

Ein letzter Test überprüft, ob alle Items zusammen eine rasch-skaliert sind. Dies ist der Andersen-Likelihood-Test verwendet werden (siehe Methoden, Formel 2-6). In Tabelle 2-8 sind die Anteile der als signifikant ausfallenden Modelltests abhängig von Stichprobengröße und Itemzahl für den unimodalen Fall eingetragen. Der signifikante Modelltest würde bedeuten, dass das Rasch-Modell die Daten nicht angemessen beschreibt. Insgesamt zeigt sich, dass dieser Test im unimodalen Fall für die Anzahl der Personen und Items invariant ist. Da die erreichten Anteile vom Niveau tendenziell oberhalb von .05 liegen, ist der Test wenn überhaupt etwas zu sensitiv.

Tabelle 2-8: Anteil signifikanter Modelltests und empirische 95%-Cut Offs der χ^2-Verteilung in der unimodalen Simulationsbedingung.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
Anteil signifikanter Tests	.061	.068	.056
95% der χ^2-Verteilung	67.83	37.35	17.23
$N = 500$			
Anteil signifikanter Tests	.051	.052	.051
95% der χ^2-Verteilung	66.48	36.48	17.02
$N = 250$			
Anteil signifikanter Tests	.055	.065	.056
95% der χ^2-Verteilung	67.59	37.998	17.48
$N = 100$			
Anteil signifikanter Tests	.055	.071	.059
95% der χ^2-Verteilung	67.14	37.56	17.33

Anmerkung. Anteil signifikanter Tests mit einer zusätzlichen Kommastelle angegeben.

Tabelle 2-9: Anteil signifikanter Modelltests und empirische 95%-Cut Offs der χ^2-Verteilung in der bimodalen Simulationsbedingung.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
Anteil signifikanter Tests	.076	.060	.058
95% der χ^2-Verteilung	67.44	37.29	17.30
$N = 500$			
Anteil signifikanter Tests	.053	.050	.039
95% der χ^2-Verteilung	66.86	36.37	16.30
$N = 250$			
Anteil signifikanter Tests	.066	.053	.060
95% der χ^2-Verteilung	68.60	36.87	17.45
$N = 100$			
Anteil signifikanter Tests	.039	.040	.046
95% der χ^2-Verteilung	65.00	35.88	16.83

Anmerkung. Anteil signifikanter Tests mit einer zusätzlichen Kommastelle angegeben.

Für den bimodalen Fall zeigt Tabelle 2-9 die erreichten Anteile und empirischen 95%-Werte der χ^2-Statistik. Das Bild, das sich hier ergibt, ist etwas differenzierter. Bei vielen Items und vielen Personen scheint der Test von der Tendenz her zu sensitiv zu sein und im Gegensatz dazu mit vielen Items, aber wenig Personen von der Tendenz eher zu wenig sensitiv. Je weniger Items der Test hat, desto geringer sind die Auswirkungen.

Insgesamt kann festgehalten werden, dass der Andersen-Test für die simulierten Daten angemessenes Verhalten zeigt. Aus diesen Gesamtergebnissen lässt sich zunächst festhalten, dass die simulierten Parameter den Bedingungen entsprachen und es plausibel ist, dass die Daten dem Rasch-Modell folgen.

2.3.3. Reproduzierbarkeit der Itemparameter

Die Darstellung der eigentlichen Simulationsergebnisse soll mit Fallbeispielen und einem Vergleich zu kommerziellen Programmen beginnen. Mit Referenz zu den beiden etablierten Schätzsoftware-Paketen WINSTEPS/JML (Linacre, 2007a) und WINMIRA/CML (Davier, 2000), wurde jeweils ein Datensatz mit 10, 25 und 50 Items für jeweils 500 Personen im unimodalen und bimodalen Fall generiert und an ihnen mit diesen Softwaren die Itemparameter geschätzt. Tabelle 2-10 bis Tabelle 2-15 präsentieren die entsprechenden Ergebnisse. Die Parameterschätzer für die Items von eRm und WINMIRA korrespondieren exakt, genauso wie die von WINSTEPS und mixRasch. Da diese beiden Programme denselben Schätzer verwenden, ist dieses Ergebnis wünschenswert. Es wird außerdem deutlich, dass die JML-Methode von der Mitte ausgehend zu den beiden Rändern der Verteilung hin zu extremeren Schätzern tendiert, als die anderen beiden Programme bzw. Methoden. Dadurch liegen die Items insgesamt etwas weiter auseinander. MML produziert sehr ähnliche Schätzer wie CML, doch neigt diese Methode vor allem bei den bimodalen Bedingungen und größeren Itemzahlen dazu, die extremen Items weniger gut zu schätzen. Die folgende Untersuchung mittels der Abweichungsstatistiken und ANOVAs soll zeigen, ob diese deskriptiven Unterschiede zwischen den Tabellen statistisch relevant sind.

Tabelle 2-10: Originale Itemparameter sowie Schätzungen der Programme für den Fall *N* = 500, *k* = 10, unimodal.

unimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i1	-2.00	-2.23	-2.22	-2.54	-2.23	-2.54
i2	-1.56	-1.66	-1.65	-1.89	-1.66	-1.88
i3	-1.11	-.89	-.88	-1.01	-.89	-1.01
i4	-.67	-.69	-.68	-.78	-.69	-.78
i5	-.22	-.16	-.16	-.18	-.16	-.18
i6	.22	.40	.39	.45	.40	.45
i7	.67	.63	.62	.71	.63	.71
i8	1.11	1.16	1.16	1.31	1.16	1.31
i9	1.56	1.49	1.50	1.69	1.49	1.69
i10	2.00	1.96	1.98	2.23	1.96	2.22

Anmerkung. i = Item

Tabelle 2-11: Originale Itemparameter sowie Schätzungen der Programme für den Fall *N* = 500, *k* = 10, bimodal.

bimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i1	-2.00	-2.09	-1.72	-2.40	-2.09	-2.39
i2	-1.56	-1.62	-1.30	-1.84	-1.62	-1.83
i3	-1.11	-1.14	-.87	-1.29	-1.14	-1.28
i4	-.67	-.71	-.49	-.80	-.07	-.80
i5	-.22	-.25	-.08	-.28	-.25	-.28
i6	.22	.35	.45	.40	.35	.40
i7	.67	.64	.71	.72	.64	.72
i8	1.11	.99	1.02	1.11	.99	1.11
i9	1.56	1.67	1.65	1.90	1.67	1.89
i10	2.00	2.15	2.11	2.47	2.15	2.46

Anmerkung. i = Item

Tabelle 2-12: Originale Itemparameter sowie Schätzungen der Programme für den Fall *N* = 500, *k* = 25, unimodal.

unimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i1	-2.00	-2.08	-2.03	-2.19	-2.08	-2.18
i2	-1.83	-1.93	-1.88	-2.03	-1.93	-2.02
i3	-1.67	-1.72	-1.67	-1.80	-1.72	-1.80
i4	-1.50	-1.67	-1.62	-1.76	-1.67	-1.75
i5	-1.33	-1.34	-1.29	-1.41	-1.34	-1.40
i6	-1.17	-1.16	-1.11	-1.22	-1.16	-1.22
i7	-1.00	-.92	-.87	-.97	-.92	-.97
i8	-.83	-.75	-.70	-.79	-.75	-.79
i9	-.67	-.72	-.67	-.76	-.72	-.76
i10	-.50	-.37	-.32	-.39	-.36	-.38
i11	-.33	-.17	-.12	-.18	-.17	-.18
i12	-.17	-.21	-.16	-.22	-.21	-.22
i13	.00	-.06	-.01	-.06	-.06	-.06
i14	.17	.33	.37	.34	.33	.34
i15	.33	.40	.44	.41	.40	.41
i16	.50	.42	.46	.44	.42	.43
i17	.67	.69	.74	.72	.69	.72
i18	.83	.88	.93	.93	.88	.92
i19	1.00	.99	1.04	1.04	.99	1.04
i20	1.17	1.24	1.29	1.30	1.24	1.30
i21	1.33	1.21	1.26	1.26	1.21	1.26
i22	1.50	1.42	1.47	1.49	1.42	1.49
i23	1.67	1.70	1.75	1.78	1.70	1.78
i24	1.83	1.92	1.98	2.02	1.92	2.01
i25	2.00	1.92	1.98	2.02	1.92	2.01

Anmerkung. i = Item

Tabelle 2-13: Originale Itemparameter sowie Schätzungen der Programme für den Fall *N* = 500, *k* = 25, bimodal.

bimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i1	-2.00	-2.00	-1.84	-2.11	-2.00	-2.10
i2	-1.83	-1.87	-1.71	-1.97	-1.87	-1.96
i3	-1.67	-1.61	-1.46	-1.68	-1.61	-1.68
i4	-1.50	-1.51	-1.36	-1.58	-1.51	-1.58
i5	-1.33	-1.29	-1.16	-1.36	-1.29	-1.35
i6	-1.17	-1.28	-1.15	-1.34	-1.28	-1.34
i7	-1.00	-1.19	-1.06	-1.25	-1.19	-1.24
i8	-.83	-.84	-.73	-.88	-.84	-.88
i9	-.67	-.32	-.24	-.34	-.32	-.33
i10	-.50	-.50	-.41	-.52	-.50	-.52
i11	-.33	-.11	-.04	-.12	-.11	-.12
i12	-.17	-.31	-.23	-.32	-.31	-.32
i13	.00	-.05	.02	-.05	-.05	-.05
i14	.17	.11	.17	.11	.11	.11
i15	.33	.27	.32	.28	.27	.28
i16	.50	.55	.59	.58	.55	.58
i17	.67	.60	.63	.62	.60	.62
i18	.83	.68	.71	.71	.68	.71
i19	1.00	1.03	1.04	1.07	1.03	1.07
i20	1.17	1.39	1.39	1.46	1.39	1.46
i21	1.33	1.24	1.25	1.30	1.24	1.30
i22	1.50	1.48	1.47	1.55	1.48	1.54
i23	1.67	1.76	1.74	1.85	1.76	1.84
i24	1.83	1.87	1.85	1.97	1.87	1.96
i25	2.00	1.91	1.88	2.00	1.91	2.00

Anmerkung. i = Item

Tabelle 2-14: Originale Itemparameter sowie Schätzungen der Programme für den Fall *N* = 500, *k* = 50, unimodal.

unimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i1	-2.00	-1.93	-1.83	-1.97	-1.93	-1.97
i2	-1.92	-2.06	-1.96	-2.11	-2.06	-2.10
i3	-1.84	-1.66	-1.57	-1.70	-1.66	-1.70
i4	-1.76	-1.66	-1.56	-1.70	-1.66	-1.70
i5	-1.67	-1.70	-1.61	-1.74	-1.70	-1.74
i6	-1.59	-1.59	-1.50	-1.63	-1.59	-1.63
i7	-1.51	-1.51	-1.42	-1.55	-1.51	-1.55
i8	-1.43	-1.40	-1.30	-1.43	-1.40	-1.43
i9	-1.35	-1.30	-1.20	-1.33	-1.30	-1.33
i10	-1.27	-1.19	-1.10	-1.22	-1.19	-1.22
i11	-1.18	-1.36	-1.27	-1.39	-1.36	-1.39
i12	-1.10	-1.32	-1.23	-1.36	-1.32	-1.35
i13	-1.02	-1.09	-1.00	-1.12	-1.09	-1.12
i14	-.94	-.91	-.81	-.93	-.91	-.93
i15	-.86	-.81	-.72	-.83	-.81	-.83
i16	-.78	-.83	-.74	-.85	-.83	-.85
i17	-.69	-.83	-.74	-.85	-.83	-.85
i18	-.61	-.77	-.67	-.79	-.77	-.79
i19	-.53	-.65	-.55	-.66	-.65	-.66
i20	-.45	-.29	-.19	-.29	-.29	-.29
i21	-.37	-.52	-.42	-.53	-.52	-.53
i22	-.29	-.32	-.23	-.33	-.32	-.33
i23	-.20	-.22	-.12	-.22	-.22	-.22
i24	-.12	.05	.15	.05	.05	.05
i25	-.04	.10	.20	.10	.09	.10
i26	.04	.07	.17	.07	.07	.07
i27	.12	.02	.12	.02	.02	.02
i28	.20	.12	.22	.12	.12	.12
i29	.29	.02	.12	.02	.02	.02
i30	.37	.44	.54	.45	.44	.45
i31	.45	.53	.63	.55	.53	.55
i32	.53	.41	.51	.42	.41	.42
i33	.61	.66	.76	.68	.66	.67
i34	.69	.78	.88	.79	.78	.79
i35	.78	.73	.83	.75	.73	.75
i36	.86	.78	.88	.79	.78	.79
i37	.94	1.02	1.12	1.05	1.02	1.05
i38	1.02	1.07	1.17	1.10	1.07	1.09
i39	1.10	1.19	1.29	1.22	1.19	1.22
i40	1.18	1.24	1.34	1.27	1.24	1.27
i41	1.27	1.24	1.34	1.27	1.24	1.27
i42	1.35	1.42	1.52	1.46	1.42	1.45
i43	1.43	1.41	1.51	1.44	1.41	1.44

Fortsetzung Tabelle 2-14:

unimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i44	1.51	1.56	1.66	1.60	1.56	1.60
i45	1.59	1.59	1.69	1.63	1.59	1.63
i46	1.67	1.78	1.88	1.82	1.78	1.82
i47	1.76	1.98	2.08	2.03	1.98	2.03
i48	1.84	1.89	1.99	1.94	1.89	1.94
i49	1.92	1.86	1.96	1.90	1.86	1.90
i50	2.00	1.95	2.05	1.99	1.95	1.99

Anmerkung. i = Item

Tabelle 2-15: Originale Itemparameter sowie Schätzungen der Programme für den Fall N = 500, k = 50, bimodal.

bimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i1	-2.00	-2.11	-2.02	-2.16	-2.10	-2.15
i2	-1.92	-1.83	-1.74	-1.87	-1.83	-1.87
i3	-1.84	-1.78	-1.70	-1.83	-1.78	-1.82
i4	-1.76	-1.74	-1.65	-1.78	-1.74	-1.78
i5	-1.67	-1.69	-1.61	-1.73	-1.69	-1.73
i6	-1.59	-1.80	-1.71	-1.84	-1.80	-1.84
i7	-1.51	-1.51	-1.43	-1.54	-1.51	-1.54
i8	-1.43	-1.48	-1.40	-1.52	-1.48	-1.51
i9	-1.35	-1.55	-1.47	-1.59	-1.55	-1.58
i10	-1.27	-1.32	-1.25	-1.35	-1.32	-1.35
i11	-1.18	-1.31	-1.24	-1.34	-1.31	-1.34
i12	-1.10	-1.40	-1.33	-1.43	-1.40	-1.43
i13	-1.02	-1.19	-1.12	-1.22	-1.19	-1.22
i14	-.94	-.93	-.87	-.95	-.93	-.95
i15	-.86	-.80	-.74	-.82	-.80	-.82
i16	-.78	-.53	-.48	-.54	-.53	-.54
i17	-.69	-.87	-.81	-.89	-.87	-.89
i18	-.61	-.78	-.72	-.79	-.78	-.79
i19	-.53	-.55	-.50	-.56	-.55	-.56
i20	-.45	-.40	-.35	-.40	-.40	-.40
i21	-.37	-.27	-.22	-.27	-.27	-.27
i22	-.29	-.25	-.21	-.26	-.25	-.26
i23	-.20	-.14	-.10	-.14	-.14	-.14
i24	-.12	-.19	-.15	-.19	-.19	-.19
i25	-.04	-.09	-.05	-.09	-.09	-.09
i26	.04	.06	.09	.06	.06	.06
i27	.12	.09	.12	.09	.09	.09
i28	.20	.18	.21	.18	.18	.18

Fortsetzung Tabelle 2-15:

bimodal	Originale Parameter	eRm/CML	ltm/MML	mixRasch/JML	WINMIRA/CML	WINSTEPS/JML
i29	.29	.31	.33	.31	.31	.31
i30	.37	.53	.55	.54	.53	.54
i31	.45	.43	.45	.44	.43	.44
i32	.53	.47	.49	.48	.47	.48
i33	.61	.58	.60	.60	.58	.60
i34	.69	.77	.78	.78	.77	.78
i35	.78	.70	.71	.71	.70	.71
i36	.86	.99	1.00	1.01	.99	1.01
i37	.94	.88	.89	.90	.88	.90
i38	1.02	1.10	1.11	1.13	1.10	1.13
i39	1.10	1.32	1.32	1.35	1.32	1.35
i40	1.18	1.31	1.31	1.34	1.31	1.34
i41	1.27	1.15	1.16	1.18	1.15	1.18
i42	1.35	1.29	1.29	1.32	1.29	1.32
i43	1.43	1.58	1.58	1.62	1.58	1.62
i44	1.51	1.48	1.48	1.52	1.48	1.52
i45	1.59	1.54	1.53	1.58	1.54	1.57
i46	1.67	1.57	1.56	1.61	1.57	1.60
i47	1.76	1.97	1.96	2.02	1.97	2.02
i48	1.84	2.04	2.03	2.09	2.04	2.09
i49	1.92	2.02	2.01	2.07	2.02	2.07
i50	2.00	2.11	2.10	2.16	2.11	2.16

Anmerkung. i = Item

Ergebnisse Itemparameter: RMSE

Bei der Implementierung eines Tests, der mittels probabilistischer Testtheorie gestaltet und validiert wurde, werden nicht generelle Skalenstatistiken, sondern die geschätzten Parameter der Items weitergegeben. Aufgrund dieser ist es dann möglich, die Personenparameter für neu untersuchte bzw. in die Klinik aufgenommene Patienten zu schätzen. Für solche Anwendungen steht die Frage im Vordergrund, ab wann die Itemparameter eines Test hinreichend genau geschätzt werden können, damit sich in der praktischen Anwendung darauf verlassen werden kann. Dies wird im Folgenden mittels der RMSEs und der Korrelationen der Parameter untersucht.

Die Tabelle 2-16 und Tabelle 2-17 zeigen zunächst das Ergebnis für die verschiedenen Itemzahlen, Stichprobengrößen und Modalität der Verteilung jeweils für die drei zur Schätzung verwendeten Programme für die Abweichungen gemessen in RMSEs.

Tabelle 2-16: Mittlere RMSEs zwischen wahren und geschätzten Itemparametern abhängig von der jeweils simulierten Itemzahl und der Schätzstichprobengröße, unimodal.

	k = 50	*k* = 25	*k* = 10
N = 1000			
eRm/CML	.077	0.077	.076
ltm/MML	.084	0.084	.085
mixRasch/JML	.084	0.098	.179
N = 500			
eRm/CML	.109	.109	.109
ltm/MML	.120	.119	.122
mixRasch/JML	.120	.128	.207
N = 250			
eRm/CML	.155	.154	.154
ltm/MML	.169	.168	.172
mixRasch/JML	.162	.173	.245
N = 100			
eRm/CML	.249	.248	.248
ltm/MML	.270	.272	.277
mixRasch/JML	.258	.270	.339

Tabelle 2-17: Mittlere RMSEs zwischen wahren und geschätzten Itemparametern abhängig von der jeweils simulierten Itemzahl und der Schätzstichprobengröße, bimodal.

	k = 50	*k* = 25	*k* = 10
N = 1000			
eRm/CML	.079	.079	.079
ltm/MML	.089	.096	.116
mixRasch/JML	.086	.100	.183
N = 500			
eRm/CML	.112	.111	.112
ltm/MML	.122	.126	.141
mixRasch/JML	.119	.131	.210
N = 250			
eRm/CML	.159	.160	.159
ltm/MML	.171	.173	.183
mixRasch/JML	.166	.179	.252
N = 100			
eRm/CML	.251	.251	.248
ltm/MML	.265	.265	.268
mixRasch/JML	.260	.274	.334

Aus beiden Tabellen ergibt sich:

a) Die Schätzer von eRm/CML zeigen durchweg die kleinsten RMSEs.

b) Die Schätzer von mixRasch/JML sind besonders groß für kleine Itemzahlen.

c) Die Schätzer von ltm/MML liegen zwischen den beiden Methoden: Sie sind schlechter als mixRasch im Falle vieler Items, aber besser im Falle weniger Items.

In Tabelle 2-10 bis Tabelle 2-15 war bereits zu sehen, dass mixRasch/JML die Schätzer für die Itemparameter, die weiter vom Skalenmittelpunkt entfernt liegen, nach außen treibt. Dies könnte eine Erklärung für diesen Effekt sein, der besonders bei wenigen Items stark ist.

Für die Gesamtbewertung dieses Kriteriums werden die folgenden Ränge vergeben: eRm/CML ist über alle Bedingungen konsistent mit dem kleinsten Fehler behaftet und erhält damit Rang 1. Für die anderen beiden Programme liegt eine Interaktion vor: Je weniger Items und (im unimodalen Fall) je mehr Personen, desto stärker liegt ltm/MML im Vorteil gegenüber mixRasch/JML. Da der Fokus der Hypothesen nicht auf einer differenzierten Bewertung für bestimmte Itemzahlen liegt, wird daher beiden Programmen der Rang 2.5 zugeteilt, da sie schlechter als eRm/CML sind und zusätzlich muss bei mixRasch/JML beachtet werden, dass insbesondere bei geringen Itemzahlen

Der RMSE lässt sich über die Simulationen in jeder Zelle des Designs für die Items insgesamt nur einmal bestimmen. Er gibt den erwarteten Fehler, der sich über alle Items über alle Stichproben einer Designzelle hinweg ergibt und so ein Schätzer für die Erwartungstreue der Methoden. Da dies aber nicht repliziert wird und damit keine Verteilung entsteht, ist er in diesem Fall anfällig sowohl für einen auftretenden Schätzbias als auch für fehlende Präzision (Harwell, 1997).

Ergebnisse Itemparameter: Nicht-parametrische Korrelationen

Diese eher deskriptive Auswertung, anhand derer die relative Genauigkeit der drei Methoden bewertet werden kann, soll im Folgenden durch die Auswertung mittels eines statistischen Tests ergänzt werden. Die Annäherung der geschätzten Itemparameter an die echten Werte kann auch durch die Verwendung von Korrelationen bestimmt werden. Hierzu wurden in jeder der simulierten

Stichproben nach der Schätzung der Parameter die parametrischen Korrelationen nach Pearson und die nicht-parametrischen Korrelationen nach Spearman verwendet. Korrelationen bieten verglichen mit dem RMSE den Vorteil, dass die Abweichungen gemessen mittels RMSE von den Simulationswerten Auskunft über die absolute Präzision der Schätzmethode (Harwell, 1997) geben. Die Itemparameter im Rasch-Modell sind aber differenzenskaliert, d.h. sie können beliebig durch die Addition von Konstanten verschoben werden, ohne dass sich die mit den Ergebnissen getroffenen Aussagen ändern (Kempf, 2003, 2008). In allen Simulationen wurde zwar dieselbe Festsetzung verwendet (der Wert "0" auf der latenten Dimension als Mittelwert der Verteilung der Personenparameter), doch um auszuschließen, dass diese Skalierungsfrage einen Einfluss auf die Vergleiche hat, werden zur Bestimmung des Effektes auf die relative Lage der Itemparameter zueinander durch die Korrelationen erfasst, die gegenüber einem solchen Effekt in der Skalierung invariant sind. Die parametrischen Korrelationen operationalisieren in diesem Vorgehen, wie nah die geschätzten Itemparameter einander sind und die nicht-parametrischen Korrelationen operationalisieren die Frage der Strukturerhaltung am deutlichsten, da sie erfassen, ob die geschätzten Itemparameter genauso geordnet sind, wie sie simuliert wurden.

Zunächst werden die nicht-parametrischen Korrelationen deskriptiv beschrieben (unimodal: Tabelle 2-18; bimodal: Tabelle 2-19) und im Anschluss teststatistisch mittels ANOVA ausgewertet. Weiter unten folgt dieselbe Darstellung für die parametrischen Korrelationen. Die nicht-parametrischen Korrelationen fallen alle sehr hoch und die Standardabweichungen dieser Korrelationen sind über die Durchläufe gering. Die Korrelationen nehmen ab, je weniger Personen zur Schätzung verwendet werden, und die Standardabweichungen nehmen zu. Je weniger Items verwendet werden, desto besser ist die Strukturwiedergabe innerhalb einer Stichprobengröße. Dieser Effekt ist dadurch zu erklären, dass die Bandbreite des latenten Kontinuums über alle Bedingungen unverändert bleibt. Je weniger Items sich nun auf diesem Abschnitt befinden, desto unwahrscheinlicher wird es, dass die Position zweier Items allein durch den Messfehler des Itemparameters in einer der Realisationen der Monte Carlo Studie vertauscht wird. Umgekehrt: Je mehr Items auf diesem Abschnitt liegen, desto wahrscheinlicher wird es, dass zwei Items den Platz tauschen. Der Messfehler der Itemparameter hängt von der Zahl der verwendeten Personen ab: Je mehr Personen

zur Schätzung verwendet werden, desto geringer fällt der Messfehler aus. Im direkten Vergleich der Werte scheint es außerdem keinen Unterschied zwischen den Modalitäten zu geben: Unabhängig von der Verteilung der Stichproben werden die Items sehr ähnlich geordnet.

Tabelle 2-18: Mittlere nicht-parametrische Korrelationen zwischen wahren und geschätzten Itemparametern abhängig von der jeweils simulierten Itemzahl und der Schätzstichprobengröße; unimodaler Fall (*SD* in Klammern).

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.998 (.001)	.999 (.001)	.99998 (.001)
ltm/MML	.998 (.001)	.999 (.001)	.99998 (.001)
mixRasch/JML	.998 (.001)	.999 (.001)	.99998 (.001)
$N = 500$			
eRm/CML	.996 (.001)	.997 (.002)	.9998 (.002)
ltm/MML	.996 (.001)	.997 (.002)	.9998 (.002)
mixRasch/JML	.996 (.004)	.997 (.002)	.9998 (.003)
$N = 250$			
eRm/CML	.992 (.002)	.993 (.003)	.998 (.005)
ltm/MML	.992 (.002)	.993 (.003)	.998 (.005)
mixRasch/JML	.992 (.002)	.993 (.003)	.998 (.004)
$N = 100$			
eRm/CML	.981 (.004)	.982 (.006)	.988 (.011)
ltm/MML	.981 (.004)	.983 (.006)	.988 (.012)
mixRasch/JML	.982 (.004)	.983 (.006)	.989 (.011)

Tabelle 2-19: Mittlere nicht-parametrische Korrelationen zwischen wahren und geschätzten Itemparametern abhängig von der jeweils simulierten Itemzahl und der Schätzstichprobengröße; bimodaler Fall (*SD* in Klammern).

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.998 (.001)	.999 (.001)	1.00 (0.00)
ltm/MML	.998 (.001)	.999 (.001)	1.00 (0.00)
mixRasch/JML	.998 (.001)	.999 (.001)	1.00 (0.00)
$N = 500$			
eRm/CML	.996 (.001)	.997 (.002)	.9998 (.002)
ltm/MML	.996 (.001)	.997 (.002)	.9998 (.002)
mixRasch/JML	.996 (.001)	.997 (.002)	.9998 (.001)
$N = 250$			
eRm/CML	.992 (.002)	.993 (.003)	.997 (.006)
ltm/MML	.992 (.002)	.993 (.003)	.997 (.006)
mixRasch/JML	.992 (.002)	.993 (.003)	.998 (.005)
$N = 100$			
eRm/CML	.980 (.005)	.982 (.007)	.988 (.012)
ltm/MML	.980 (.005)	.982 (.007)	.988 (.012)
mixRasch/JML	.981 (.004)	.982 (.007)	.989 (.011)

Um diesen Eindruck statistisch zu überprüfen, wurde eine Varianzanalyse gerechnet, mit den Zwischensubjektfaktoren Modalität, Itemzahl und Stichprobengröße und innerhalb jeder simulierten Stichprobe dem Programm zur Schätzung der Parameter als messwiederholtem Faktor. Tabelle 2-20 zeigt das Ergebnis zunächst für die Zwischensubjekteffekte. Während alle Haupt- und Interaktionseffekte außer dem höchsten signifikant werden ($p < .05$), zeigen sich in der Effektstärke (partielles η^2) deutliche Unterschiede. Diese Unterschiede in den Effektstärken sind bei der Bewertung einer Monte Carlo Studie aufgrund der hohen Stichprobengröße weit aussagekräftiger als die Signifikanzen selber. Das partielle η^2 für die Stichprobengröße beträgt .62 und ist damit der stärkste Effekt. Dieser ist gefolgt von dem Haupteffekt für die Zahl der Items in der Skala (partielles $\eta^2 = .17$). Schließlich zeigt der Interaktionsterm dieser beiden Variablen noch ein partielles η^2 von .04. Die anderen Effekte sind ihrer Größe nach vernachlässigbar (Grenze zur Interpretation bei $\eta^2 > .01$), dementsprechend wird hier nur auf die Variablen dieser Interaktion eingegangen.

Stichprobengröße. Für die sechs möglichen Vergleiche wird mittels Dunn's Prozedur die kritische Differenz ermittelt (*Diff* = .0003)[25]. Alle sechs möglichen Vergleiche zwischen den mittleren nicht-parametrischen Korrelationen, die sich für diesen Faktor ergeben, fallen größer aus, als die so bestimmte Differenz. Dies bedeutet, dass bis zu einer Stichprobengröße von $N = 1000$ ein signifikanter Zuwachs in der Genauigkeit der Schätzung der Reihenfolge der Itemparameter erreicht wird.

Zahl der Items. Für die drei möglichen Vergleiche (Dunn's *Diff* = .0002) zwischen den mittleren nicht-parametrischen Korrelationen, die sich für diesen Faktor ergeben, fallen größer aus, als die so bestimmte Differenz. Dies bedeutet, dass je weniger Items die Skala hat, im Bereich 50 bis 10 Items eine Verbesserung der Präzision erreicht wird.

Interaktion Itemzahl und Stichprobengröße. Werden innerhalb jeder Stichprobengröße die Korrelationen verglichen (*Diff* = .0006), so ergeben sich bei 10 Items immer die höchsten (und bei 50 die niedrigsten) Korrelationen. Innerhalb der Itemzahlen ergeben sich für 25 und 50 Items mit stei-

[25] Die kritischen Differenzen werden alle mit der kleinsten Zellenbelegung gerechnet (s. Tabelle 2-2, Tabelle 2-3 und FN 24); aufgrund der hohen Anzahl der Fehlerfreiheitsgrade werden die tabellierten Werte für ∞ verwendet.

gender Stichprobengröße immer höhere Korrelationen. Nur bei 10 Items unterscheiden sich die Korrelationen zwischen $N = 500$ und $N = 1000$ nicht mehr signifikant voneinander.

Tabelle 2-20: Zwischensubjekteffekte der ANOVA zum Vergleich der mittleren nicht-parametrischen Korrelationen der Itemparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Konstanter Term	1	1.18×10^{9}	< .001	1.00
Modalität	1	35.26	< .001	.002
Itemzahl	2	2358.88	< .001	.17
Stichprobe	3	12603.00	< .001	.62
Modalität X Itemzahl	2	3.41	.03	.00
Modalität X Stichprobengröße	3	5.91	< .001	.001
Itemzahl X Stichprobengröße	6	153.55	< .001	.001
Modalität X Itemzahl X Stichprobengröße	6	.23	.97	.00
Fehler	22948	MSQ = 5.65×10^{-5}		

Für die Innersubjekteffekte ergibt sich folgendes Bild (Tabelle 2-21). Das zur Schätzung verwendete Programm hat einen Effekt auf die Güte der Schätzer (partielles $\eta^2 = .02$). Die Stichprobengröße interagiert mit dem Programm (partielles $\eta^2 = .02$). Alle anderen Effekte sind entweder nicht signifikant oder klären weniger als 1% der Varianz auf.

Programm. Unterschiede, die größer als die kritische Differenz sind (*Diff* = .00003) ergeben sich zwischen den Korrelationen des mixRasch/JML-Schätzers ($r = .99426$) auf der einen und den Schätzern der anderen beiden Programme (eRm/CML: $r = .99405$; ltm/MML: $r = .99404$) auf der anderen Seite.

Programm und Stichprobengröße. Bei der kritischen Differenz von *Diff* = .0006 werden innerhalb der Stichprobengrößen nur zwei Vergleiche signifikant. Bei $N = 100$ macht es einen Unterschied, welches Programm verwendet wird: mixRasch/JML zeigt bei dieser Stichprobengröße im Mittel signifikant höhere Korrelationen ($r = .98400$) als die anderen beiden Programme (eRm/CML: $r = .98386$; ltm/MML: $r = .98383$). Sonst ist es unerheblich, welches Programm verwendet wird, zwischen den Korrelationen wird kein Unterschied gefunden. Innerhalb jedes der

Programme ist allerdings ein signifikanter Zuwachs in den Korrelationen zu beobachten: Alle Vergleiche zwischen den mittleren Korrelationen der Itemparameter fallen signifikant aus. Bis inkl. $N = 1000$ ist somit also ein Genauigkeitszuwachs in der Schätzung zu beobachten. Dieser Effekt ist allerdings insgesamt sehr klein.

Tabelle 2-21: Innersubjekteffekte der ANOVA zum Vergleich der mittleren nicht-parametrischen Korrelationen der Itemparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Programm	2	361.75	< .001	.02
Programm X Modalität	2	1.44	.24	.00
Programm X Itemzahl	4	7.63	< .001	.001
Programm X Stichprobengröße	6	125.47	< .001	.02
Programm X Modalität X Stichprobengröße	6	.49	.74	.00
Programm X Itemzahl X Stichprobengröße	12	7.09	< .001	.002
Programm X Modalität X Itemzahl X Stichprobengröße	12	.31	.99	.00
Fehler	45896	MSQ = 1.12×10^{-6}		

Wie aus Tabelle 2-19 ersichtlich, ist die Bedingung mit $N = 1000$ und $i = 10$ in der bimodalen Simulation ohne Varianz: Überall wird die Reihenfolge der Items exakt repliziert. Da fraglich ist, ob eine solche Bedingung ohne Varianz mit einer ANOVA auswertbar ist, wurde diese Bedingung ausgeschlossen, um die Robustheit der Ergebnisse zu prüfen. Die Ergebnisse für die Zwischensubjekteffekte blieben erhalten (partielle η^2 für Stichprobengröße .62; Itemzahl .18 und deren Interaktion .02). Für die Innersubjekt Effekte ebenfalls (partielle η^2 für Programm .02; Interaktion mit Stichprobengröße .014).

Insgesamt lässt sich festhalten, dass die Schätzer von mixRasch/JML die höchsten nicht-parametrischen Korrelationen aufweisen. Durch die Betrachtung der Ergebnisse der Fallstudien und die RMSEs wird allerdings deutlich, dass dieser Effekt durch eine übermäßige Verschiebung der Itemparameter zu den Rändern der Skala erreicht wird (siehe RMSE). Daher wird in der Gesamtbewertung für JML ein Rang von 3 vergeben; die anderen zwei Programme unterscheiden sich nicht (daher derselbe Rang: 1.5; siehe Tabelle 2-34 zur Gesamtbewertung).

Ergebnisse Itemparameter: Parametrische Korrelationen

Tabelle 2-22 und Tabelle 2-23 zeigen die mittleren Korrelationen (inkl. SDs) zwischen den zur Simulation verwendeten wahren und den geschätzten Itemparametern.

Tabelle 2-22: Mittlere parametrische Korrelationen zwischen wahren und geschätzten Itemparametern abhängig von der jeweils simulierten Itemzahl und der Schätzstichprobengröße; unimodaler Fall (*SD* in Klammern).

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.998 (.0004)	.998 (.001)	.998 (.001)
ltm/MML	.998 (.0004)	.998 (.001)	.998 (.001)
mixRasch/JML	.998 (.0004)	.998 (.001)	.998 (.001)
$N = 500$			
eRm/CML	.996 (.001)	.996 (.001)	.997 (.001)
ltm/MML	.996 (.001)	.996 (.001)	.997 (.001)
mixRasch/JML	.996 (.001)	.996 (.001)	.997 (.002)
$N = 250$			
eRm/CML	.992 (.002)	.992 (.002)	.994 (.003)
ltm/MML	.992 (.002)	.992 (.002)	.994 (.003)
mixRasch/JML	.992 (.002)	.992 (.002)	.994 (.003)
$N = 100$			
eRm/CML	.979 (.004)	.981 (.006)	.985 (.008)
ltm/MML	.979 (.004)	.981 (.006)	.985 (.008)
mixRasch/JML	.979 (.004)	.981 (.006)	.985 (.008)

Tabelle 2-23: Mittlere parametrische Korrelationen zwischen wahren und geschätzten Itemparametern abhängig von der jeweils simulierten Itemzahl und der Schätzstichprobengröße; bimodaler Fall (*SD* in Klammern).

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.998 (.0004)	.998 (.001)	.998 (.001)
ltm/MML	.998 (.0004)	.998 (.001)	.998 (.001)
mixRasch/JML	.998 (.0005)	.998 (.001)	.998 (.001)
$N = 500$			
eRm/CML	.996 (.001)	.996 (.001)	.997 (.002)
ltm/MML	.996 (.001)	.996 (.001)	.997 (.002)
mixRasch/JML	.996 (.001)	.996 (.001)	.997 (.002)
$N = 250$			
eRm/CML	.991 (.002)	.992 (.002)	.994 (.003)
ltm/MML	.991 (.002)	.992 (.002)	.994 (.003)
mixRasch/JML	.991 (.002)	.992 (.002)	.994 (.003)
$N = 100$			
eRm/CML	.979 (.004)	.980 (.006)	.985 (.008)
ltm/MML	.979 (.004)	.980 (.006)	.985 (.008)
mixRasch/JML	.979 (.004)	.980 (.006)	.984 (.008)

Auch bei den parametrischen Korrelationen ist zu sehen, dass die mittleren Korrelationen durchwegs hoch sind und auch diese steigen wie bei den nicht-parametrischen Korrelationen an, je weniger Items verwendet werden. Die mittleren Korrelationen zwischen den wahren und geschätzten Itemparametern sind zwischen den Programmen weder im unimodalen, noch im bimodalen Fall deskriptiv unterschiedlich. Auch für die parametrischen Korrelationen wurde eine ANOVA durchgeführt mit dem Programm/Schätzverfahren als messwiederholtem Faktor, um den Eindruck der Unabhängigkeit von Programm und Verteilung zu prüfen.

Tabelle 2-24: Zwischensubjekteffekte der ANOVA zum Vergleich der mittleren parametrischen Korrelationen der Itemparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Konstanter Term	1	2.10×10^9	< .001	1.00
Modalität	1	56.09	< .001	.002
Itemzahl	2	1098.59	< .001	.09
Stichprobengröße	3	26318.43	< .001	.78
Modalität X Itemzahl	2	1.21	.30	.00
Modalität X Stichprobengröße	3	5.00	.002	.001
Itemzahl X Stichprobengröße	6	228.97	< .001	.06
Modalität X Itemzahl X Stichprobengröße	6	.44	.86	.00
Fehler	22948	MSQ = 3.15×10^{-5}		

Für die Zwischensubjekteffekte (Tabelle 2-24) sind nahezu alle Haupt- und Interaktionseffekte signifikant; es werden nur sowohl der höchste Interaktionsterm als auch die Interaktion zwischen Modalität und Itemzahl nicht signifikant. Den deutlichsten Einfluss auf die Korrelationen hat erneut die Stichprobengröße (partielles $\eta^2 = .78$), gefolgt von der Itemzahl (partielles $\eta^2 = .09$) und der Interaktion dieser beiden Faktoren (partielles $\eta^2 = .06$).

Stichprobengröße. Alle sechs möglichen Vergleiche (Dunn's *Diff* = .0002) zwischen den mittleren parametrischen Korrelationen, die sich für diesen Faktor ergeben, fallen größer aus, als die so bestimmte Differenz. Dies bedeutet, dass bis zu einer Stichprobengröße von $N = 1000$ ein signifikanter Zuwachs in der Genauigkeit der Schätzung der Position der Itemparameter erreicht wird.

Zahl der Items. Alle drei möglichen Vergleiche zwischen den mittleren parametrischen Korrelationen, die sich für diesen Faktor ergeben, fallen größer aus, als die Differenz (*Diff* = .0002). Dies bedeutet auch hier, dass je weniger Items die Skala hat, im Bereich 50 bis 10 Items eine Verbesserung der Präzision erreicht wird.

Interaktion Itemzahl und Stichprobengröße. Werden innerhalb jeder Stichprobengröße die Korrelationen verglichen (*Diff* = .0004), zeigen sich bei den Stichprobengrößen $N = 100$ und $N = 250$ bei 10 Items die höchsten (und bei 50 die niedrigsten) Korrelationen. Bei $N = 500$ unterscheiden sich die Korrelationen für $k = 50$ und $k = 25$ nicht mehr voneinander. Bei $N = 1000$ ist nur noch der Vergleich von $k = 50$ mit $k = 10$ signifikant. Innerhalb der Itemzahlen im Gegensatz zum Ergebnis bei den nicht-parametrischen Korrelationen kann bei allen Skalenlängen bis $N = 1000$ ein Genauigkeitszuwachs beobachtet werden.

Durch den messwiederholten Faktor "Programm/Schätzer" wird wie oben nur wenig Varianz aufgeklärt. Die Grenze $\eta^2 > .01$ wird beim Haupteffekt sowie dem Interaktionseffekt mit der Itemzahl erreicht (Tabelle 2-25).

Tabelle 2-25: Innersubjekteffekte der ANOVA zum Vergleich der mittleren parametrischen Korrelationen der Itemparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Programm	2	300.12	< .001	.01
Programm X Modalität	2	101.38	< .001	.004
Programm X Itemzahl	4	204.44	< .001	.02
Programm X Stichprobengröße	6	53.50	< .001	.007
Programm X Modalität X Itemzahl	4	11.71	< .001	.001
Programm X Modalität X Stichprobengröße	6	14.97	< .001	.002
Programm X Itemzahl X Stichprobengröße	12	40.25	< .001	.01
Programm X Modalität X Itemzahl X Stichprobengröße	12	1.22	.27	.00
Fehler	45896	MSQ = 3.18×10^{-8}		

Programm. Unterschiede, die größer die kritische Differenz (*Diff* = .000005) sind, zeigen, dass eRm/CML und ltm/MML (nummerisch gleich: *r* = .992626) die höchsten Korrelationen erreichen und diese unterscheiden sich signifikant von mixRasch/JML (*r* = .992591).

Interaktion Programm und Stichprobengröße. Innerhalb der Programme sind alle Vergleiche signifikant (*Diff* = .00001), dass bedeutet, unabhängig vom Programm ist bis inkl. *N* = 1000 eine signifikante Steigerung der Messgenauigkeit zu beobachten. Innerhalb der Stichprobengrößen fallen aber nur wenige Vergleiche zwischen den Programmen signifikant aus:

- Bei *N* = 100 erreicht ltm/MML die höchsten Korrelation (*r* = .98184), gefolgt von eRm/CML (*r* = .98183) und mixRasch/JML (*r* = 98175); alle Vergleiche signifikant.
- Bei *N* = 250 erreichen eRm/CML (*r* = .99249) und ltm/MML (*r* = .99248) höhere Korrelationen als mixRasch/JML (*r* = .99245).
- Bei *N* = 500 erreichen eRm/CML (*r* = .99622) und ltm/MML (*r* = .99622) höhere Korrelationen als mixRasch/JML (*r* = .99620).
- Bei *N* = 1000 besteht kein signifikanter Unterschied zwischen den Programmen mehr.

Aufgrund der kleinen Unterschiede zwischen ltm/MML und eRm/CML wird daher in der Gesamtbewertung wieder beiden Programmen derselbe Rang gegeben (1.5; siehe Tabelle 2-34) und mixRasch/JML auf den dritten Rang eingeordnet.

2.3.4. Fazit Itemparameter

Insgesamt kann festgehalten werden, dass alle drei Programme in der Lage sind, die Itemparameter zuverlässig auf der latenten Dimension anzuordnen. Itemzahl und Stichprobengröße haben deutliche Einflüsse auf die Genauigkeit der Schätzung, doch betreffen diese alle drei Programme gleichmäßig. Die Modalität der Personenparameterverteilung zeigte keinen Einfluss auf die Genauigkeit der Schätzung der Itemparameter. Insgesamt kann daher allen Programmen für den Zweck der Schätzung der Itemparameter eine Anwendungsempfehlung ausgesprochen werden.

Im Detail muss dieses Gesamturteil etwas relativiert werden. Wenn davon abgesehen wird, dass die Unterschiede zwischen den Programmen marginal in ihrer absoluten Größe sind, so kann

doch gesagt werden, dass mixRasch/JML tendenziell schlechter bei der Schätzung der Itemparameter abschneidet. Wie in den Fallbeispielen deutlich zu erkennen (s. Tabelle 2-10 bis Tabelle 2-15) werden bei JML die Schätzer für die Itemparameter der Items, die vom Mittelpunkt der Skala entfernter liegen, stärker nach außen gedrängt, d.h. extremer. Dadurch gelingt es mit JML tendenziell besser, die Reihenfolge der Items zu replizieren (s. nicht-parametrische Korrelationen), aber die Exaktheit der Wiedergabe auf dem Kontinuum (RMSE und parametrische Korrelationen) ist weniger gut. Dieser Effekt ist besonders groß für kleine Stichproben und wenige Items. Daher schneidet in der Gesamtbewertung (Tabelle 2-34) eRm für die Aspekte der Itemparameterschätzung am besten ab, gefolgt von ltm und schließlich mixRasch.

2.3.5. Reproduzierbarkeit der Personenparameter

Von besonderer Relevanz für die Diagnostik ist die Frage, wie genau der Personenparameter, d.h. die Ausprägung der zu messenden latenten Eigenschaft, geschätzt wird. Dabei kann entweder gefordert werden, dass der Parameter möglichst exakt geschätzt werden soll (im Folgenden mittels des RMSE geprüft) oder aber dass die Ordnung der Personen zumindest möglichst unverändert bleibt (Thissen & Wainer, 2001; im Folgenden mittels nicht-parametrischer Korrelationen geprüft). In dieser Simulation wird von dem Fall ausgegangen, dass das gesamte Spektrum der latenten Fähigkeit von Interesse ist und ebenso die gesamte Personenverteilung. Wie in Kapitel 3 vorgestellt wird, können auch nur spezifische Abschnitte des latenten Kontinuums interessant sein oder Teile der untersuchten Stichprobe bzw. Population. Zunächst werden die Ergebnisse für den RMSE präsentiert, danach für die nicht-parametrischen Korrelationen.

Ergebnisse Personenparameter: RMSE

Im unimodalen Fall (Tabelle 2-26) zeigt sich eine deutliche Zunahme der RMSEs bei allen drei Programmen mit abnehmender Itemzahl. Auch unterscheiden sich die Programme deutlich in den RMSEs: ltm/MML produziert konsequent die niedrigsten und mixRasch/JML die höchsten Abweichungen. Für den bimodalen Fall (Tabelle 2-27) zeigen sich dieselben Tendenzen wie im unimodalen Fall. Zusätzlich zeigt der Vergleich zwischen den beiden Modalitäten, dass sich im bimodalen Fall insgesamt größere Abweichungen ergeben.

Tabelle 2-26: Mittlerer RMSE der Personenparameter zwischen den wahren und den geschätzten (eRm, ltm, mixRasch) Personenparametern; unimodaler Fall, *SD*s in Klammern.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.357 (.009)	.517 (.014)	.838 (.022)
ltm/MML	.328 (.008)	.441 (.010)	.616 (.014)
mixRasch/JML	.362 (.010)	.532 (.015)	.911 (.026)
$N = 500$			
eRm/CML	.357 (.013)	.517 (.020)	.840 (.032)
ltm/MML	.330 (.012)	.441 (.015)	.616 (.020)
mixRasch/JML	.362 (.014)	.532 (.022)	.913 (.039)
$N = 250$			
eRm/CML	.357 (.018)	.518 (.030)	.841 (.044)
ltm/MML	.332 (.017)	.444 (.021)	.617 (.028)
mixRasch/JML	.363 (.020)	.535 (.033)	.915 (.052)
$N = 100$			
eRm/CML	.360 (.031)	.520 (.045)	.850 (.0739)
ltm/MML	.340 (.031)	.448 (.036)	.622 (.0458)
mixRasch/JML	.366 (.033)	.538 (.050)	.927 (.088)

Tabelle 2-27: Mittlerer RMSE der Personenparameter zwischen den wahren und den geschätzten (eRm, ltm, mixRasch) Personenparametern; bimodaler Fall, *SD*s in Klammern.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.383 (.012)	.550 (.016)	.849 (.023)
ltm/MML	.362 (.009)	.497 (.012)	.723 (.017)
mixRasch/JML	.390 (.013)	.572 (.018)	.937 (.028)
$N = 500$			
eRm/CML	.383 (.017)	.552 (.022)	.852 (.033)
ltm/MML	.363 (.013)	.498 (.017)	.723 (.024)
mixRasch/JML	.391 (.019)	.574 (.025)	.941 (.040)
$N = 250$			
eRm/CML	.384 (.023)	.551 (.031)	.857 (.045)
ltm/MML	.367 (.019)	.500 (.025)	.724 (.033)
mixRasch/JML	.393 (.026)	.574 (.036)	.947 (.055)
$N = 100$			
eRm/CML	.382 (.034)	.551 (.051)	.862 (.075)
ltm/MML	.369 (.031)	.501 (.038)	.727 (.054)
mixRasch/JML	.392 (.040)	.579 (.059)	.954 (.093)

Im Gegensatz zur Untersuchung der Itemparameter, bei der nur über alle Items und über alle Simulationsstichproben aggregiert werden konnte, wurde die mittlere Abweichung der Personenparameter für jede Realisation in der Monte Carlo Studie bestimmt. So wurden die RMSEs der Personenparameter analog zu obigem Vorgehen mit einer messwiederholten ANOVA ausgewertet. Für

die Zwischensubjektanalysen (Tabelle 2-28) erwiesen sich alle Haupt- und Interaktionseffekte außer dem höchsten als signifikant. Relevant waren die Effekte der Itemzahl (partielles η^2 = .977), der Modalität (partielles η^2 = .33) und die Interaktion dieser beiden Faktoren (partielles η^2 = .02). Diese sollen im Folgenden genauer untersucht werden.

Itemzahl. Wieder wurde mittels der Prozedur von Dunn die kritische Differenz bestimmt (*Diff* = .001). Alle drei möglichen Vergleiche fallen signifikant aus, d.h. der RMSE nimmt bis zur Zahl von 50 Items signifikant ab.

Modalität. Bei dem zweistufigen Faktor ist die mittlere Abweichung in den unimodalen Stichproben kleiner (RMSE = .547) als in den bimodalen (RMSE = .593).

Interaktion Itemzahl X Modalität. Die kritische Differenz für die neun Vergleiche innerhalb der Stufen der Faktoren beträgt *Diff* = .002. Diese Differenz wird bei allen Skalenlängen beim Vergleich uni- vs. bimodal überschritten, d.h. in den bimodalen Bedingungen ist die Abweichung von den wahren Werten immer signifikant größer, diese Differenz wird aber stetig kleiner bei Zunahme der Itemzahl. Auch innerhalb der Modalitäten fallen alle Vergleiche signifikant aus: Die größten Fehler gibt es bei *k* = 10, die kleinsten bei *k* = 50.

Tabelle 2-28: Zwischensubjektfaktoren der ANOVA zum Vergleich der mittleren RMSEs der Personenparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Konstanter Term	1	8874645.67	< .001	.997
Modalität	1	27.40	< .001	.33
Itemzahl	2	484590.24	< .001	.98
Stichprobengröße	3	57.92	< .001	.008
Modalität X Itemzahl	2	243.02	< .001	.02
Modalität X Stichprobengröße	3	3.19	.02	.00
Itemzahl X Stichprobengröße	6	6.14	< .001	.002
Modalität X Itemzahl X Stichprobengröße	6	.71	.64	.00
Fehler	22948	MSQ = .002		

Bei den Vergleichen für den messwiederholten Faktor sind alle Haupt- und Interaktionseffekte außer dem höchsten und der Interaktion Programm X Modalität X Stichprobengröße signifikant. Varianz oberhalb der festgelegten Grenze klären auf: Das Programm (partielles η^2 = .89), die Interaktion zwischen Programm und Itemzahl (partielles η^2 = .83), gefolgt von der Interaktion zwischen Programm und Modalität (partielles η^2 = .19) und der Interaktion zwischen Programm, Modalität und Itemzahl (partielles η^2 = .17).

Tabelle 2-29: Innersubjektfaktoren der ANOVA zum Vergleich der mittleren RMSEs der Personenparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Programm	2	180551.07	< .001	.89
Programm X Modalität	2	5283.67	< .001	.19
Programm X Itemzahl	4	56854.10	< .001	.83
Programm X Stichprobengröße	6	6.76	< .001	.001
Programm Modalität X Itemzahl	4	2339.36	< .001	.17
Programm X Modalität X Stichprobengröße	6	1.27	.27	.00
Programm X Itemzahl X Stichprobengröße	12	16.24	< .001	.004
Programm X Modalität X Itemzahl X Stichprobengröße	12	1.11	.35	.00
Fehler	45896	MSQ = .001		

Programm. Die kritische Differenz (*Diff* = .0001) wird bei allen drei Vergleichen überschritten. ltm/MML hat die im Mittel geringsten Abweichungen (RMSE = .498), gefolgt von eRm/CML (RMSE = .587) und schließlich von mixRasch/JML (RMSE = .623).

Interaktion Programm X Itemzahl. Die kritische Differenz für die achtzehn Vergleiche innerhalb der Faktorstufen beträgt *Diff* = .002 und diese Differenz wird in allen überschritten. Die RMSEs nehmen innerhalb der Programme mit zunehmender Itemzahl signifikant ab und ltm/MML produziert immer die geringsten Abweichungen, gefolgt von eRm/CML und schließlich mixRasch/JML.

Interaktion Programm X Modalität. Die kritische Differenz (*Diff* = .0016) wird in allen neun möglichen Vergleichen innerhalb der Faktoren überschritten. Die Reihenfolge der Programme ist

wie oben bereits beschrieben ltm/MML < eRm/CML < mixRasch/JML und auch der bereits beschriebene Effekt für die Modalität bleibt erhalten (RMSE unimodal < RMSE bimodal).

Interaktion Programm X Modalität X Itemzahl. Die kritische Differenz für die 45 möglichen direkten Vergleiche, die die oben berichteten Haupt- und Interaktionseffekte einschließen (*Diff* = .003) wird bei allen überschritten. Das bedeutet, innerhalb jedes Programmes nimmt die Schätzgenauigkeit mit steigender Itemzahl zu und dies unabhängig von der Modalität. Ebenfalls unabhängig von der Modalität zeigt ltm/MML im Vergleich die niedrigsten Fehler, eRm/CML die mittleren und mixRasch/JML die höchsten. Im Vergleich über die Modalitäten zeigt sich zusätzlich, dass alle Programme bei der bimodalen Verteilung größere Abweichungen von den simulierten Werten zeigen.

Personenparameter: Nicht-parametrische Korrelationen

Zusätzlich zur Genauigkeit sollte noch das diagnostische Kriterium der Reihung der Personen untersucht werden. Da die Parameter eines IRT Modells nicht auf dem Niveau einer Absolutskala (z.B. Kempf, 2003) gemessen werden, wäre in der praktischen Anwendung ggf. eine exakte Reproduktion nicht nötig. Daher wurde mit nicht-parametrischen Korrelationen getestet, ob sich die Programme in der relativen Ordnung der simulierten Personen unterscheiden. Innerhalb jeder der simulierten Stichproben wurden die wahren Personenparameter mit den Ergebnissen der Schätzungen der Programme korreliert und diese Korrelationen verglichen. Tabelle 2-30 und Tabelle 2-31 zeigen die Mittelwerte und Standardabweichungen aus den Durchläufen der Simulationen. Wird die klassische Definition der Reliabilität verwendet (Kempf, 2008; Lord & Novick, 1968), dann können die abgebildeten Korrelationen zwischen den Schätzern und den wahren Werten als Reliabilitäten interpretiert werden.

In beiden Tabellen sind bereits Tendenzen zu erkennen. Da die Korrelation zwischen den zur Simulation genutzten und den geschätzten Parametern als Reliabilität interpretiert werden kann, ist nachvollziehbar, dass die Werte mit steigender Itemzahl zunehmen. Auch unter diesem Gesichtspunkt ist deutlich, dass wenn die Verteilung bimodal und breiter ist, die Korrelationen höher ausfallen: Eine Erhöhung der True-Score-Varianz führt bei gleichbleibender Fehlervarianz zur Erhöhung

der Reliabilität (z.B. Lord & Novick, 1968). Beim Vergleich zwischen den Programmen zeigt sich, dass eRm/CML und mixRasch/JML in der Regel identische Werte liefern und diese etwas oberhalb der von ltm/MML erreichten Werte liegen.

Tabelle 2-30: Mittlere nicht-parametrische Korrelationen zwischen den wahren und den geschätzten (eRm, ltm, mixRasch) Personenparametern; unimodaler Fall, *SD*s in Klammern.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.944 (.004)	.896 (.007)	.782 (.013)
ltm/MML	.943 (.004)	.894 (.007)	.775 (.014)
mixRasch/JML	.944 (.004)	.896 (.007)	.782 (.013)
$N = 500$			
eRm/CML	.943 (.006)	.895 (.010)	.784 (.019)
ltm/MML	.943 (.006)	.894 (.010)	.776 (.019)
mixRasch/JML	.943 (.006)	.895 (.010)	.784 (.019)
$N = 250$			
eRm/CML	,942 (.008)	.895 (.014)	.783 (.026)
ltm/MML	.942 (.008)	.893 (.014)	.776 (.027)
mixRasch/JML	,942 (.008)	.895 (.014)	.783 (.026)
$N = 100$			
eRm/CML	.940 (.013)	.892 (.023)	.779 (.044)
ltm/MML	.940 (.013)	.889 (.024)	.771 (.045)
mixRasch/JML	.941 (.013)	.892 (.023)	.779 (.044)

Tabelle 2-31: Mittlere nicht-parametrische Korrelationen zwischen den wahren und den geschätzten (eRm, ltm, mixRasch) Personenparametern; bimodaler Fall, *SD*s in Klammern.

	$k = 50$	$k = 25$	$k = 10$
$N = 1000$			
eRm/CML	.967 (.002)	.937 (.004)	.859 (.009)
ltm/MML	.967 (.002)	.936 (.004)	.853 (.009)
mixRasch/JML	.967 (.002)	.937 (.004)	.859 (.009)
$N = 500$			
eRm/CML	.967 (.003)	.937 (.006)	.858 (.012)
ltm/MML	.967 (.003)	.935 (.006)	.852 (.013)
mixRasch/JML	.967 (.003)	.937 (.006)	.858 (.012)
$N = 250$			
eRm/CML	.966 (.005)	.936 (.008)	.858 (.018)
ltm/MML	.966 (.005)	.935 (.008)	.852 (.019)
mixRasch/JML	.966 (.005)	.936 (.008)	.858 (.018)
$N = 100$			
eRm/CML	.965 (.007)	.933 (.013)	.853 (.030)
ltm/MML	.964 (.007)	.931 (.013)	.847 (.031)
mixRasch/JML	.965 (.007)	.933 (.013)	.853 (.030)

Für die Überprüfung, ob sich die Korrelationen zwischen den wahren und geschätzten Personenparametern abhängig von den Bedingungen unterscheiden, wurde eine Varianzanalyse mit dem Programm als messwiederholtem Faktor angewendet (s.a. oben). Bei den Zwischensubjektfaktoren (Tabelle 2-32) werden alle Haupt- und Interaktionseffekte signifikant außer der Interaktion Modalität X Stichprobengröße und der Interaktion aller Faktoren. Als relevante Zwischensubjektfaktoren bezogen auf das partielle η^2 erweisen sich hier die Itemzahl (partielles $\eta^2 = .93$), die Modalität (partielles $\eta^2 = .68$) und die Interaktion dieser beiden Faktoren (partielles $\eta^2 = .30$).

Tabelle 2-32: Zwischensubjektfaktoren der ANOVA zum Vergleich der mittleren nichtparametrischen Korrelationen der Personenparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Konstanter Term	1	69912133.21	< .001	1.00
Modalität	1	47947.70	< .001	.68
Itemzahl	2	145491.84	< .001	.93
Stichprobengröße	3	68.95	< .001	.009
Modalität X Itemzahl	2	5006.32	< .001	.30
Modalität X Stichprobengröße	3	.66	.58	.00
Itemzahl X Stichprobengröße	6	2.67	.01	.001
Modalität X Itemzahl X Stichprobengröße	6	1.04	.40	.00
Fehler	22948	MSQ = .001		

Itemzahl. Die kritische Differenz nach Dunn für die drei möglichen Vergleiche beträgt *Diff* = .00097 und sie wird bei allen Vergleichen überschritten. Die Korrelationen steigen also signifikant von $k = 10$ ($r = .82$) bis $k = 50$ ($r = .95$) Items an.

Modalität. Der Faktor Modalität besitzt nur zwei Stufen, die sich daher signifikant voneinander unterscheiden müssen. In der bimodalen Bedingung fallen die Korrelationen signifikant höher aus ($r = .92$) als in der unimodalen ($r = .87$).

Interaktion Modalität X Itemzahl. Die kritische Differenz (*Diff* = .002) für die neun möglichen Vergleiche innerhalb beider Faktoren wird bei allen überschritten. Dies bestätigt die beiden Haupteffekte: Die Korrelationen sind immer in der bimodalen Bedingung höher und steigen außerdem

innerhalb jeder Modalität mit der Zahl der Items. Dieser Unterschied zwischen den Modalitäten schwächt sich zwar mit steigender Zahl der Items ab, aber nicht genügend, um nicht mehr signifikant zu sein.

Für den messwiederholten Faktor "Programm" zeigen sich differenziertere Ergebnisse als für die Reproduktion der Reihenfolge der Itemparameter. Alle Haupt- und Interaktionseffekte sind signifikant – außer der Interaktion Programm X Modalität X Stichprobengröße sowie der Vierfachinteraktion. Zunächst gibt es hier einen starken Haupteffekt für das Programm (partielles $\eta^2 = .43$). Die nächststärkere Interaktion ist die für das verwendete Schätzprogramm mit der Itemzahl (partielles $\eta^2 = .40$), gefolgt von der Interaktion Programm mit der Modalität (partielles $\eta^2 = .02$).

Tabelle 2-33: Innersubjektfaktoren der ANOVA zum Vergleich der mittleren nicht-parametrischen Korrelationen der Personenparameter in den Bedingungen.

	Freiheitsgrade	F-Wert	Signifikanz	Partielles η^2
Programm	2	17495.12	< .001	.43
Programm X Modalität	2	337.93	< .001	.02
Programm X Itemzahl	4	7619.11	< .001	.40
Programm X Stichprobengröße	6	29.41	< .001	.004
Programm X Modalität X Itemzahl	4	111.97	< .001	.01
Programm X Modalität X Stichprobengröße	6	1.44	.20	.00
Programm X Itemzahl X Stichprobengröße	12	3.16	< .001	.001
Programm X Modalität X Itemzahl X Stichprobengröße	12	.66	.79	.00
Fehler	45896	MSQ = 3.84×10^{-6}		

Programm. Die kritische Differenz (*Diff* = .00008) wird nicht bei allen drei Vergleichen überschritten. Im Mittel erreichen die beiden Programme eRm/CML und mixRasch/JML höhere Korrelationen (beide $r = .895$) als ltm/MML ($r = .892$).

Interaktion Programm X Itemzahl. Die kritische Differenz für die 18 Vergleiche innerhalb der Gruppen (*Diff* = .0002) wird ebenfalls nicht bei allen Vergleichen überschritten. Die Programme mixRasch/JML und eRm/CML haben dieselben Korrelationen und unterscheiden sich daher auch

bei den verändernden Stichprobengrößen nicht. ltm/MML zeigt dafür im Mittel leicht niedrigere Korrelationen über alle Itemzahlen hinweg. Die Vergleiche bei den Itemzahlen fallen alle signifikant aus, d.h. innerhalb der Programme ist über die Breite der getesteten Bedingungen ein Genauigkeitszuwachs bei der Ordnung der Personen zu beobachten.

Interaktion Programm X Modalität. Die kritische Differenz beträgt *Diff* = .0001 und wird bei allen Vergleichen außer beim Vergleich zwischen eRm und mixRasch signifikant (diese erreichen wieder dieselben Werte). Beide Programme erreichen also unabhängig von der Modalität höhere Korrelationen und in der bimodalen Stichprobe bleiben die Korrelationen unabhängig vom Programm höher.

2.3.6. Fazit Personenparameter

In der Bewertung der Genauigkeit der Schätzung der Personenparameter erbringt keines der Programme eindeutig die besten Ergebnisse. Bei der Bewertung der Präzision der Personenparameter erreicht ltm/MML (mit Emprical Bayes als Schätzung für den Personenparameter; Rizopoulos, 2006) die besten Werte, gefolgt von eRm/CML und dies wiederum von mixRasch/JML. Diese Unterschiede sind signifikant, selbst bei der Kontrolle für den höchsten Interaktionseffekt mit Modalität und Itemzahl. Für die nicht-parametrischen Korrelationen, die statt der Präzision der Schätzer die Stabilität der Personenordnung im Vergleich zur Originalrangreihe prüfen, zeigt sich, dass eRm/CML und mixRasch die Reihung exakt gleich gut replizieren und dies besser schaffen als ltm/MML. Die vergebenen Ränge sind entsprechend Tabelle 2-34 zu entnehmen.

2.4. Diskussion

Die Studie untersuchte, ob drei Pakete in der Softwareumgebung R (ltm, eRm und mixRasch), die jeweils einen anderen Schätzer nutzen, zu gleichen Ergebnissen bei der Rekonstruktion der Parameter des Rasch-Modells kommen. Hierzu wurden in den Bedingungen Modalität (2) X Itemzahl (3) X Stichprobengröße (4) jeweils 1000 Datensätze simuliert, die dem Rasch-Modell entsprechen, und jeweils mit den drei Paketen analysiert. Getestet wurden zwei Hypothesen, a) ob eines der Pakete besser zur Schätzung der Parameter geeignet ist, und b) ob eines der Pakete anfälliger ist für eine in der klinischen Forschung typische Verteilung des Personenmerkmals

(Bimodalität). Die Ergebnisse wurden bereits nach den jeweiligen Analysen knapp zusammengefasst. Zum Überblick sind die Ergebnisse in Tabelle 2-34 noch einmal zusammengefasst und sollen hier noch einmal kurz umrissen werden.

2.4.1. Itemparameter

Für die Schätzung der Itemparameter war das Ergebnis, dass die Programme eRm und ltm (mit den Schätzern CML und MML) am besten abschnitten. In den teststatistisch abgesicherten Kriterien der nicht-parametrischen (Rangfolge der Items) und parametrischen (lineare Exaktheit der Position) Korrelationen zwischen Schätzung und den simulierten Werten gab es zwischen den beiden Programmen keinen Unterschied. In der mittleren Abweichung der Itemparameter über alle Replikationen (RMSE) erwies sich eRm als leicht überlegen. Das Paket mixRasch zeigte insgesamt schlechtere Eigenschaften, da die Itemparameter am wenigsten exakt wiedergegeben wurden. Je weiter ein Item vom Mittelpunkt der Skala entfernt liegt, desto stärker wird der Schätzer für den Itemparameter nach außen getrieben.

2.4.2. Personenparameter

Da bei den Personenparametern ein Verteilungsparameter (mittlere Lage der Einzelwerte) bewertet werden kann, lag der RMSE als Präzisionsmaß für alle simulierten Stichproben vor und konnte teststatistisch überprüft werden. Hier ergab sich ein klarer Vorteil des Programms ltm mit MML und Empirical Bayes als Schätzer der Personenparameter. Die Abweichung der geschätzten Personenparameter war bei der Verwendung dieses Paketes am kleinsten. Das Paket eRm/CML erwies sich als signifikant schlechter und mixRasch/JML erwies sich wiederum als signifikant schlechter im Vergleich zu den anderen beiden Pakete. Die Reihung der Personen wurde dagegen durch die beiden Pakete eRm/CML und mixRasch/JML am exaktesten wiedergegeben und ltm erwies sich als signifikant schlechter. Dieses Ergebnis ist darauf zurückzuführen, dass mixRasch und eRm den Summenscore als Reihungskriterium verwenden und nur zu unterschiedlichen Ergebnissen kommen könnten, wenn die Itemparameter deutlich andere Schätzer erreichen würden. Das Paket ltm/MML berechnet diese Parameter aufgrund der Itemparameter und der Verteilungsannahme im Empirical Bayes Schätzer (hier Normalverteilungsannahme). Die Interaktion zwischen

Programm und Modalität ist nicht disordinal, das bedeutet, dass dieser Schätzer auch nicht schlechter wird, wenn diese Annahme verletzt ist (zumindest unter den simulierten Bedingungen).

2.4.3. Modalität

Wie in Hypothese 2 formuliert, wurde untersucht, ob die Verteilung der Personenparameter einen Einfluss auf die Schätzung der Modelle hat. Es wurde ein deutlicher Einfluss festgestellt, der daraus resultiert, dass die bimodale Verteilung breiter ist als die unimodale – dadurch erhöht sich die Varianz des latenten Konstruktes, nicht aber die Fehlervarianz. Dadurch werden die Personenparameter reliabler schätzbar (z.B. Lord & Novick, 1968). Nach der Kontrolle für diesen Effekt zeigt sich aber kein besonderer Vor-/ Nachteil einer Schätzmethode gegenüber den anderen.

Parameter, die in dieser Studie nicht variiert wurden und in weiteren Untersuchungen manipuliert werden könnten, sind gleich breit gestreute uni- und bimodale Verteilungen, andere Verteilungsformen (siehe Beispiel bei Connell et al., 2007), sowie andere Mischverhältnisse zwischen den Populationen (hier immer 50-50). Unter den verwendeten Bedingungen kann festgehalten werden, dass auch bei zwei deutlich getrennten Verteilungen (d = 1.70) eine hinreichend gute Parameterschätzung erreicht werden kann. Wie in der Einleitung dargestellt, war diese als plausible Effektstärke für den Unterschied zwischen klinischen und nicht-klinischen Populationen in einem Messinstrument für allgemeine psychologische Belastung gewählt worden. Diese Einschätzung wird auch bestätigt, wenn zum Vergleich Effektstärken für die Veränderung in der Psychotherapie herangezogen werden, die in der Regel in Meta-Analysen im Mittel unter diesem Wert liegen (Lambert & Ogles, 2004) genauso wie Werte, die in der psychotherapeutischen Routineversorgung erreicht werden (z.B. Lutz, Böhnke, Köck, et al., 2011; Schindler & Hiller, 2010). Somit können die Ergebnisse der Parameterschätzungen aller drei Programme als für in der klinisch-psychologischen Forschung relevante Bedingungen überprüft angesehen werden.

2.4.4. Gesamtbewertung

In einer Gesamtbewertung der mittleren Ränge aus der Schätzung der Itemparameter und der Schätzung der Personenparameter (s. Tabelle 2-34) schneidet das Paket eRm/CML am besten ab, gefolgt von ltm/MML und zuletzt mixRasch/JML. Aufgrund der diskutierten Vergleiche ist damit

die Empfehlung auszusprechen, eRm (und damit CML) zu verwenden – eher als ltm mit MML und dies wiederum eher als mixRasch. Der Befund, dass die JML-Methode nicht zu den optimalsten Parameterschätzern führt, fügt sich gut in die existierende analytische Literatur (Kempf, 2008; Neyman & Scott, 1948; Rost, 2004) wie auch in die empirischen Belege ein. Cohen und Kollegen zeigten zum Beispiel denselben Befund, der auch in dieser Studie gefunden wurde: Niedrige Itemparameter werden unterschätzt, hohe werden überschätzt (Cohen et al., 2008: Abbildung 1).

Tabelle 2-34: Auflistung der Kriterien und Rangergebnisse aus der Studie.

	eRm	ltm	mixRasch	Ergebnisverweis
Itemparameter				
RMSE (deskriptiv)	1	2.5	2.5	siehe Tabelle 2-16 und Tabelle 2-17
nicht-parametrische Korrelationen	1.5	1.5	3	siehe Tabelle 2-18 und Tabelle 2-19
parametrische Korrelationen	1.5	1.5	3	siehe Tabelle 2-22 und Tabelle 2-23
Mittel Itemparameter	1.33	1.83	2.83	
Personenparameter				
nicht-parametrische Korrelationen	1.5	3	1.5	siehe Tabelle 2-30 und Tabelle 2-31
RMSE	2	1	3	siehe Tabelle 2-26 und Tabelle 2-27
Mittel Personenparameter	1.75	2	2.25	
Mittel der mittleren Ränge	**1.54**	**1.92**	**2.54**	

Das Ergebnis der Unterschiede muss allerdings relativiert werden. Vor dem Hintergrund der geringen Variation der Ergebnisse in den Simulationsläufen, bedeuten die großen Effektstärken nur große Unterschiede bezogen auf eine geringe Variabilität der herangezogenen Kriterien. Numerisch sind die Ergebnisse sowohl für die Itemparameter wie auch für die Personenparameter in Bereichen, die als sehr gut bezeichnet werden können, und unterscheiden sich in ihren Absolutwerten kaum. Ob es Fälle gibt, in denen die Unterschiede zwischen den Programmen und Schätzern praktische Relevanz erlangen, muss in der Anwendung gezeigt werden. Die Ergebnisse dieser Studie zeigen, dass die Unterschiede zwischen den Programmen statistisch zwar reliabel sind, aber von geringer Größe. Alle drei Pakete können zur Schätzung der Modelle gut verwendet werden.

2.4.5. Stichprobengröße & Modellwahl

Wie eingangs beschrieben, gehen die Schätzer für die nötige Stichprobengröße weit auseinander. Die Ergebnisse dieser Studie weisen darauf hin, dass selbst bei für Simulationsstudien üblichen, aber im Vergleich zu existierenden Tests niedrigen, Itemzahlen eine Steigerung der Genauigkeit erreicht wird, wenn die Stichprobengröße von 500 auf 1000 Personen steigt. Außerdem betont die Untersuchung, dass die Präzision in der Schätzung der Itemparameter sinkt, je mehr Items ein Test hat (da sich auf der latenten Dimension mehr Items anordnen lassen müssen). Die Stichprobenplanung für eine IRT-Instrumentenentwicklung sollte sich am Zweck der orientieren. Für eine Testung der psychometrischen Eigenschaften eines Instrumentes mögen relativ kleine Stichproben genügen (für dieses Argument s. z.B. Orlando Edelen & Reeve, 2007). Sollen die Ergebnisse aber weiter verwendet werden, wie z.B. in einer Itembank, dann sind größere Stichproben nötig. Wie groß diese Stichproben sein müssten, müsste Ziel weiterer Studien sein, die den Trade-Off zwischen der Anzahl Items (und damit der Genauigkeit der Reproduktion der Itemordnung aber auch der Personenordnung) und der nötigen Stichprobengröße untersuchen. Diese Studien sollten auch untersuchen, wann fehlende Genauigkeit auch praktische Konsequenzen hat.

Diese Frage tritt noch stärker in den Vordergrund, wenn statt dichotomer polytome Items untersucht werden. Der Zusammenhang zwischen den Ergebnissen dieser Arbeit und polytomen Ergebnissen ergibt sich dadurch, dass unter Verwendung des Partial Credit Models (Masters, 1982) für ein polytomes Item die Schwellenparameter (d.h. die Übergänge zwischen zwei benachbarten Kategorien) auf der latenten Dimension verortet werden. Für zehn polytome Items mit jeweils 5 Kategorien ist dies etwa ähnlich den 50 dichotomen Items, die in dieser Arbeit verwendet wurden (s. Kapitel 3 und Huynh, 1994, 1996). Dies zeigt, dass für ein übliches Instrument in der Versorgung mehr Parameter zu schätzen sind, als in dieser Studie realisiert. Andere Parametrisierungen des Rasch-Modells stehen nicht vor diesem Problem. In dem Rating Scale Modell (Andersen, 1997) wird davon ausgegangen, dass die Schwellen aller Items in gleicher Weise um die Schwierigkeit des Items angeordnet sind. Hier würde es ausreichen, wenn sich die Schwierigkeit (Vergleichbar mit dem Mittelwert des Items) reliabel anordnen lassen. Eine solche Situation ist wieder analog zu den in dieser Studie untersuchten Bedingungen.

Probleme der Modellwahl und der Frage der Relevanz der Unterschiede zwischen Parameterschätzern können nur über Simulationsstudien gepaart mit naturalistischen Daten rekonstruiert werden. Nach Bolt (2005: S. 48) sind neben der Untersuchung von Daten, bei denen das Modell passt, echte empirische Datensets informativ, die groß sind und an denen an wiederholt gezogenen Stichproben das Verhalten der Schätzer untersucht werden kann.

2.4.6. Reliabilität und Skalenlänge

Die in der Untersuchung festgestellten Reliabilitäten und Verläufe der Informationsfunktionen zeigen, dass ab einer Zahl von 50 dichotomen Items über einen breiten Skalenbereich zufriedenstellende Messergebnisse erreichbar sind. Wird gemäß Lord & Novick (1968) die Reliabilität als der Anteil der True Score Varianz an der Gesamtvarianz definiert und über die Korrelation der wahren Werte mit den geschätzten berechnet, so liegt dieser Wert über .90. Wird dies auf das ordinale Partial Credit Model übertragen, wäre also zu erwarten, dass mit bereits ca. 10 fünfstufigen Items eine solch hohe Messqualität mit rasch-skalierten Instrumenten erreichen ließe. Dies betont ein weiteres Mal die wünschenswerten Eigenschaften der Rasch-Skalierung.

Die Studie untersuchte die Messgenauigkeit von Items, die über ein Spektrum der Skala verteilt sind. Der Verlauf der Informationsfunktionen zeigt noch einmal deutlich, dass eine Optimierung der Skala auf ein bestimmtes Spektrum es auch ermöglicht, bei gleichbleibender Messgenauigkeit deutlich weniger Items zu verwenden. Die Forschung zum Verhalten computer-adaptiver Tests zeigt, dass bei der Schätzung der Lage eines Individuums auf der latenten Dimension, also einem festen Wert, bereits sehr geringe Itemzahlen ausreichen, da durch sie lokal eine hohe Messgenauigkeit erreicht werden kann (Raîche et al., 2007; Wainer, 2000; Kapitel 3).

Im Zusammenhang mit der Frage, wie viele Items verwendet werden sollten, stellt sich die Frage, wie relevant die verwendeten Itemzahlen sind. Die verwendeten Anzahlen sind in Simulationsstudien durchaus üblich (z.B. Emons, Sijtsma, & Meijer, 2007; Hidalgo & López-Pina, 2011; Orlando Edelen & Reeve, 2007). Bezogen auf die Fragestellung der Unterschiede zwischen den Programmen zeigt sich, dass die Ergebnisse der Schätzer der Programme konvergieren. Für die Verwendung im Bereich der computer-adaptiven Tests zeigt sich auch, dass so geringe Zahlen für

die Schätzung fester Personenwerte vermutlich genügen (Babcock & Weiss, 2009). Die Forschung zur Erstellung von allgemeinen Kurzformen, die für ein ganzes Spektrum geeignet sein sollen, zeigt jedoch, dass dies nicht ohne weiteres generalisierbar ist. In einer aktuellen Arbeit werden Ergebnisse zur Messqualität einer Kurzfassung des Brief Symptom Inventory mit 18 Items diskutiert (Derogatis, 2001; Meijer et al., 2011). Die Autoren können zeigen, dass die Messqualität für dieses Instrument insgesamt (wie auch für die Subskalen) angemessen ist. Aber die Analyse der Testinformationsfunktionen zeigte, dass das Instrument besonders zur Messung im mittleren bis hohen Belastungsbereich geeignet ist. Für diesen Bereich mögen die 18 Items also geeignet sein, zur Differenzierung zwischen niedrigeren Belastungsgraden eignet sich das Instrument aber nicht. Emons und Kollegen (2007) zeigten bereits früher auf, dass Klassifizierungsentscheidungen mit wenigen Items nur schwer zu treffen sind. Sie testeten verschiedene Kurzskalenlängen (sechs bis 40 Items) und zeigten, dass selbst mit Skalen zwischen 20 bis 40 dichotomen Items die Klassifikationskonsistenz nicht zufriedenstellend war und dass es nur geringe Unterschiede zwischen dichotomen und polytomen Items gab.

2.4.7. Abschluss und Ausblick

Bezogen auf die Nutzbarkeitsevaluation kann festgehalten werden, dass die Pakete der freien Softwareumgebung R zur Schätzung des Rasch-Modells verwendet werden können. Die Schätzungen der Programme sind alle hoch mit den wahren Werten, die zur Simulation genutzt wurden, korreliert. Damit stehen diese Pakete zur Verwendung in der Routineversorgung bereit und können auch in den eingangs umrissenen, eher ressourcenarmen Kontexten eingesetzt werden und so zu einer Erleichterung der Wissenschaftler-Praktiker-Zusammenarbeit beitragen.

Im Rahmen der Qualitätssicherung unterstreicht diese Studie auch, dass sich in der freien Softwareumgebung R (R Development Core Team, 2010) Simulationen zur Überprüfung statistischer Methoden leicht umsetzen lassen und so die Qualität der in der Praxis angewendeten Entscheidungs- und Auswertungsroutinen gut prüfen lässt (ICH E9, 1998).

Diese Ergebnisse gepaart mit der stärkeren Verbreitung von R in der Wissenschaftswelt und der Tatsache, dass dort auch viele Modelle schätzbar sind, die gerade in der Patientenorientierten Ver-

sorgungsforschung von Interesse sind (konditionale und nicht-konditionale Wachstumsmodelle; Gelman & Hill, 2007; Long, 2012; Lutz et al., 1999; Lutz, 2002; Lutz, Leach, et al., 2005) weisen auf das weitere Potential von R, da es möglich ist, alle Analysemethoden in einer Umgebung durchzuführen und die Ergebnisse der verschiedenen Analysen direkt in einander zu überführen, ohne die Statistiksoftware wechseln zu müssen.

3. Studie II: Die Verwendung von Item Response Modellen und Bootstrap-Techniken zur Entwicklung von Fragebogenkurzformen[26]

3.1. Einleitung

Verschiedene Entwicklungen in den letzten Jahren führten dazu, dass es einen erhöhten Bedarf an der Messung und Dokumentation von Fortschritten während Therapieprozessen gibt (s. Einleitung, 1.4). Dieser Trend zeichnete sich bereits in den Bereichen der Psychotherapieforschung ab, da relevante Ergebnis- und Prozessmerkmale reliabel und valide gemessen werden müssen (Doucette & Wolf, 2009; Kazdin, 1998; Lambert & Ogles, 2004; Stiles et al., 1986). Parallel zu der Betonung dieser unterschiedlichen Forschungsstrategien fand eine deutlich stärkere Hinwendung zur praktischen Ausrichtung und Verwertbarkeit der Forschung statt. Wissenschaftler-Praktiker-Netzwerke (Borkovec et al., 2001; Locke et al., 2011; Lutz, Böhnke, Köck, et al., 2011; Lutz, 2011; Steffanowski et al., 2011) wurden etabliert und durch sie wurde eine stärkere Rückbindung des Forschungsprozesses an die Wünsche, aber auch an die Anforderungen dieser Settings erreicht. Dies führte zu einer stärkeren Verwendung von psychometrischen Maßen in der Praxis (Lambert & Ogles, 2004), aber umgekehrt auch der stärkeren Forderung, die Instrumente besser auf diese Kontexte zuzuschneiden (Cahill et al., 2011; Lutz, Tholen, et al., 2006; Lutz, Schürch, et al., 2009; Meier, 1997). Auch die Entwicklung von Lösungen für spezifische Erhebungssettings im Sinne der Qualitätssicherung und Feedbackforschung macht es nötig, dass die Instrumente nicht einfach "von der Stange" gekauft werden, sondern auf den spezifischen Kontext optimal zugeschnitten werden (Härter et al., 2003; Krampen, 2010; Laireiter & Vogel, 1998; Lutz, 2002). Das TK Modellvorhaben unterstreicht diesen Punkt: Auch wenn die Instrumentenbatterie auf einige wenige Fragebögen konzentriert war und nur wenige Erhebungen stattfanden (kürzestes Intervall: alle 10 Sitzungen), so wird dennoch im Abschlussbericht festgehalten, dass dies bereits einen zu kompensierenden Zusatzaufwand darstelle (Wittmann et al., 2011).

[26] Diese Studie wurde als Konferenzbeitrag vorgestellt (Böhnke & Lutz, 2010a, 2010b); derzeit in gekürzter Fassung eingereicht als (Böhnke & Lutz, submitted)

Doch auch außerhalb der Psychotherapieforschung sind diese Bewegungen bemerkbar. In der Medizin gibt es bereits seit längerer Zeit einen erhöhten Bedarf an der Verwendung von Fragebogeninstrumenten. Unter der Überschrift "Patient Reported Outcomes" ("vom Patienten berichtete Behandlungsergebnisse") gibt es bereits seit längerem den Versuch, systematisch die Perspektive der Patienten auf Behandlungsprozess und Wohlbefinden zu erheben. Zentrale Bedeutung kommt dabei der PROMIS-Initiative zu ("Patient-Reported Outcome Information System"), die darauf abzielt, eine Vielzahl von Belastungsdimensionen auf Seiten der Patienten im Selbstreport systematisch zu erheben (Ader, 2007; Ellwood, 1988; Riley et al., 2010; U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, & U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health, 2006; Willke et al., 2004). All diese Bewegungen entspringen verschiedenen Motiven, die sich in der konkreten Anwendung durchaus auch überschneiden können:

1. Entwicklung und Etablierung von systematischen Routinen zur Verlaufsdiagnostik (Böhnke & Lutz, 2010b; Howard et al., 1996; Lambert, 2007; Lueger et al., 2001; Lutz, Böhnke, Köck, et al., 2011; Lutz, 2002).

2. Systematische Einbindung von verschiedenen Perspektiven auf den Therapieverlauf (siehe Kapitel 1; Cuijpers et al., 2010; Grosse Holtforth et al., 2010; Seidenstücker & Baumann, 1987; Seidenstücker, 1995).

3. Qualitätssicherung und Qualitätsmanagement (Laireiter & Vogel, 1998; Lutz, 1997, 2011).

4. Transparenz und Empowerment der Patienten (McAllister, Dunn, Payne, Davies, & Todd, 2012).

5. Verbesserung der Qualität/Qualitätsentwicklung des psychotherapeutischen Angebotes selber (z.B. der Effekt von Feedback für negativ entwickelnde Patienten und Nutzung

ökonomischer Spareffekte, s. Kapitel 1 und Lambert & Shimokawa, 2011; Lambert, 2001; Lord Darzi, 2008; Lutz, 2002).

Diese zusätzlichen Erhebungen lassen sich aus verschiedenen Perspektiven normativ als wünschenswert bezeichnen und haben auch empirische Bedeutung. Der Effekt der Wirkung von Dokumentation und Verlaufsfeedback auf das Therapieergebnis ist beispielsweise breit dokumentiert (Lambert & Shimokawa, 2011; Poston & Hanson, 2010) und stellt damit eine Rechtfertigung für den inkrementellen Wert zusätzlicher Erhebungen dar (siehe Kapitel 1; Hunsley & Mash, 2005; Meyer et al., 2001). Auch nehmen verschiedene Stakeholder und oft auch die Patienten diese Maßnahmen als positiv wahr (Lutz, Böhnke, Köck, et al., 2011; Lutz, 1997; Steffanowski et al., 2011; The Future Vision Coalition, 2009).

Dem gegenüber steht, dass diese Systeme ab einem gewissen Grad zu einer deutlich erhöhten (zumindest zeitlichen) Belastung der Patienten führen: Von ihnen müssen über den Verlauf einer Behandlung mehrmals (und unter Umständen jeweils eine ganze Reihe) von Erhebungsinstrumenten bearbeitet werden (Gilbody et al., 2002b; Lutz, Böhnke, & Köck, 2011; Newnham & Page, 2010; Walker et al., 2010). Aus diesem Grund ist es nötig über kürzere Erhebungsinstrumente nachzudenken und solche ggf. zu konstruieren (Böhnke & Lutz, 2010b; Forkmann, Boecker, Wirtz, Glaesmer, et al., 2010; Lutz, Tholen, et al., 2006; Meier, 1997; Meijer et al., 2011). Eine Möglichkeit ist die Entwicklung neuer Fragebogen, die die gewünschten Eigenschaften haben (s. z.B. Tabelle 4-1 mit einer Übersicht zu Kriterien starker Veränderungssensitivität). Über die vergangenen Jahrzehnte der Psychotherapieforschung wurden eine ganze Reihe von Fragebögen entwickelt (Lambert & Ogles, 2004), doch nur wenige wurden auch in mehreren Studien eingesetzt. Die Entwicklung neuer Instrumente für spezifische Settings ist günstig für die jeweilige Studie, da sie genau den jeweiligen Wünschen und Zwecken gerecht werden. Dieser Vorteil macht auf der anderen Seite aber Vergleiche über verschiedene Studien hinweg eher kompliziert und eine Erweiterung der Testlandschaft nur aus diesem Grund scheint nicht wünschenswert.

Die Entwicklung neuer Tests kann aus vielen Gründen immer wieder notwendig sein: So gibt es beispielsweise bisher keine Instrumente für ein Konstrukt oder der Fragebogen ist aus anderen

Gründen, wie z.B. Urheberrechtsfragen nicht verwendbar (z.B. Lutz, Tholen, et al., 2006; Lutz, Schürch, et al., 2009; Reise & Henson, 2003). Doch in Forschungskontexten, in denen Nutzer auf eine Vergleichbarkeit von Ergebnissen angewiesen sind, sollte von weiteren Entwicklungen eher abgesehen werden. In der Patientenorientierten Versorgungsforschung ist es unerlässlich, dass die Verlaufsdaten eines Patienten vor dem Hintergrund der Forschung betrachtet werden können, damit das Urteil getroffen werden kann, ob die Therapie anschlägt oder eben nicht und bei der Qualitätssicherung in Anwendungskontexten ist es wichtig, dass möglichst einfach klare Rückbezüge zu den Wirkungsweisen der Therapie in anderen Institutionen gemacht werden können (Böhnke & Lutz, 2010b; Howard et al., 1996; Lutz & Grawe, 2007; Schindler & Hiller, 2010)[27].

Eine andere Möglichkeit ist die Entwicklung von Kurzformen bereits existierender Instrumente. Basierend auf vorher definierten Kriterien und durch die gezielte Erhebung von Stichproben zur Testung, ob diese Kurzformen die spezifischen Ziele erfüllen, werden Items aus existierenden Fragebögen ausgewählt, die dann zu einer verkürzten Skala zusammengestellt werden (Lutz, Tholen, et al., 2006). Es existieren verschiedene Listen für die Auswahl von spezifischen Items (Meier, 1997; Vermeersch, Lambert, & Burlingame, 2000), statistischen Verfahren, wie ein Fragebogen für bestimmte Zwecke optimiert werden kann (s. Kapitel 4) oder aber auch von testpsychologischen Kriterien, die herangezogen werden können (Doucette & Wolf, 2009; Lutz, Tholen, et al., 2006; Reise & Haviland, 2005). Dieser Ansatz führt zu verkürzten Skalen, die ggf. auch hohe Messqualität aufweisen, doch bleibt die Vergleichbarkeit der Scores aus der Kurz- und der Langform fraglich (Thissen & Wainer, 2001).

Auch wenn Tests nach Kriterien der Item Response Theorie (IRT, siehe Kapitel 1; Doucette & Wolf, 2009; Hambleton et al., 1991) erstellt werden, werden oftmals nur solche Kriterien herangezogen, die die Messqualität untersuchen. Ein zentraler Unterschied zwischen IRT und der Klassischen Testtheorie (Kempf, 2003, 2008; Rost, 2004) besteht darin, dass bei der IRT Personen und Items zusammen auf der latenten Eigenschaft positioniert werden. Dadurch kommt es dazu, dass

[27] Meta-Analysen stellen zwar eine Möglichkeit dar, Vergleiche über Studien hinweg zu ziehen, doch erlauben diese kaum einen blick auf den konkreten Einzelfall, der in der Qualitätssicherung und patientenorientierten Versorgungsforschung von Bedeutung ist.

nicht nur der gesamte erreichte Testscore bereits Informationen über die Ausprägung der latenten Eigenschaft enthält (wie in der Klassischen Testtheorie), sondern auch jede einzelne Itemantwort bereits Informationen darüber bereitstellt, welche Ausprägung der latenten Eigenschaft bei einer Person vorliegt. Dieses Prinzip macht den Kern von computer-adaptivem Testen aus (Meijer & Nering, 1999; Reise & Henson, 2003; Wainer, 2000; Walker et al., 2010). Beim computer-adaptiven Testen wird eine sog. "Item-Bank" verwendet, die eine Fülle von Items enthält, die *eine* latente Eigenschaft messen. Bei jeder Testsituation werden aus dieser Datenbank Items ausgewählt und dem Probanden vorgegeben und je nach dem, welche Antworten er gibt, werden weitere Items ausgewählt. Dieser Prozess wird einige Male durchgeführt, bis ein Abbruchkriterium erfüllt ist, z.B. eine maximale Zahl Items vorgelegt wurde oder der Messfehler unter eine bestimmte Grenze gefallen ist (Meijer & Nering, 1999; Raîche et al., 2007; Wainer, 2000; Walker et al., 2010).

Ein solches Vorgehen wäre mit Mitteln der Klassischen Testtheorie nicht möglich, da nur eine Gruppe Items gemeinsam als "Score" ausgewertet werden können (Moosbrugger & Kelava, 2007; Steyer & Eid, 2001). In der Klassischen Testtheorie muss daher jede Kurzversion aus einer Reihe festgelegter Items bestehen, für die dann Normen entwickelt werden, damit sie in der Praxis verwendet werden können (Lutz, Tholen, et al., 2006). Dies ist bei computer-adaptiven Tests nicht nötig: Wenn gezeigt wurde, dass die Items verlässlich eine einzelne Dimension messen, dann kann jede beliebige Kombination dieser Items verwendet werden, um das latente Konstrukt zu messen (Holman, Lindeboom, et al., 2003).

Computer-adaptive Anwendungen sind in der empirischen Bildungsforschung, Zulassungsprüfungen und Bewerberauswahlen zwar weit verbreitet (Embretson & Reise, 2000), doch trotz der Vorteile kann computer-adaptives Testen nicht als ein Standardverfahren in der klinischen Versorgung (oder anderen Teilen der Psychologie) gesehen werden (Reise & Haviland, 2005; Reise & Henson, 2003; Zickar & Broadfoot, 2009). Besonders die bereits angesprochene PROMIS Initiative ist dadurch hervorzuheben, dass sie sich dem Problem annimmt, wie eine Vielzahl verschiedener Ergebnis-Dimensionen effektiv durch die Verwendung von computer-adaptiven Tests gemessen werden kann (Riley et al., 2010). Weitere Beispiele aus dem Gebiet der Psychotherapie sind die

Entwicklungen computer-adaptiver Tests zur Erhebung von depressiver Belastung (Fliege et al., 2005; Forkmann, Boecker, Wirtz, Glaesmer, et al., 2010), das zur wiederholten Messung der Symptomschwere bei depressiven Erkrankungen eingesetzt wird. Die Autoren konstruierten im ersten Schritt ihrer Arbeit einen "Itempool" (eine Sammlung von möglichen Kandidatenitems) zur Messung von "Depression" in Übereinstimmung mit DSM-IV (Saß, Wittchen, Zaudig, & Houben, 2003). Dieser bestand in einer ersten Stufe aus 320 Items, die aus einer Reihe von Fragebögen stammten. Nach einer Bewertung durch unabhängige Juroren, wie indikativ die Items jeweils für "Depression" waren, blieben noch 144 Items, die Eingang in die Kalibrierungsstudie fanden. Diese wurden aufgeteilt in zwei Itemsets mit 30 überlappenden Items, die dann insgesamt $N = 3270$ Patienten vorgelegt wurden. Nach Anwendung mehrerer Schritte zur Prüfung der Konstruktvalidität (Faktorenanalyse, Generalized Partial Credit Model) wurden insgesamt 80 der Items ausgeschlossen und es verblieb damit eine Item Bank von 64 Items, die sich in angeschlossenen Simulations- und Kreuzvalidierungsstudien als sehr effektiv erwies (z.B. im Zentralbereich des Beschwerdespektrums waren unter 10 Items zu einer effizienten Messung ($SE \leq .32$) nötig)[28].

Die Gründe dafür, dass computer-adaptive Tests bislang nicht als Standardtechnologie oder Vorgehensweise bei der Erfassung von Ergebnis- und Prozessdimensionen bei der Psychotherapie gesehen werden kann, sind vielfältig. Ein Grund ist die mangelnde Vertrautheit mit den statistischen Modellen, die zum Einsatz dieser Technologie nötig sind, denn obwohl diese Modelle zumindest von der Ableitung her seit über 50 Jahren zur Verfügung stehen (Lord & Novick, 1968), so kämpfen sie dennoch um Anerkennung und Einsatz (Reise & Haviland, 2005). Ein anderer Grund liegt auf der Seite der Konsumenten, in diesem Fall Psychotherapeuten (und ggf. anderem klinischen Personal), die dem Zusatznutzen von Fragebogenerhebungen generell skeptisch gegenüber steht und oft auch nicht die technische Ausstattung besitzt, um computer-adaptive Tests anwenden zu können (siehe auch Kapitel 1.4). In der Routineversorgung und insbesondere in Privatpraxen werden viele Erhebungen noch regelhaft mit dem klassischen Papier-und-Bleistift-Verfahren erho-

[28] Zu einer weiteren Diskussion und einer zusätzlichen Anwendung in der klinischen Psychologie aus dem deutschen Sprachraum am Beispiel der Depression sei verwiesen auf: Forkmann, Böcker, Wirtz, Norra, & Gauggel, 2012; Forkmann, Boecker, Wirtz, Glaesmer, et al., 2010.

ben. In dem TK Modellvorhaben Psychotherapie, in dem vermutlich eher Technik-offene und Fragebogen freundliche Therapeuten teilnahmen, administrierten nur etwa 2/3 der teilnehmenden Therapeuten den Bogen per PC (Wittmann et al., 2011). Dies wird zusätzlich noch dadurch relativiert, dass insgesamt der Anteil der Psychologen in Deutschland, die Fragebögen in der Therapie einsetzen, nicht besonders hoch ist (Hagemeister et al., 2010; D. R. Hatfield & Ogles, 2004; Jensen-Doss & Hawley, 2011; Steck, 1997).

In Kapitel 1.4.2 wurde aufgezeigt, welchen Nutzen psychometrische Erhebungen in verschiedenen Perspektiven erfüllen können (Patientenorientierte Versorgungsforschung, Diagnostik und Qualitätssicherung). Aus diesen wäre es durchaus wünschenswert, wenn solche Erhebungen in Form von Monitoring-Systemen größere Verbreitung finden. Die Debatte um die Qualitätssicherung in der Psychotherapie sowie die Durchführung von Großprojekten (Lutz, Wittmann, Böhnke, Rubel, & Steffanowski, 2012; Puschner & Kordy, 2010; Steffanowski et al., 2011) zeigen, dass in Deutschland ein Umdenken stattfindet, wenn auch bislang eher kritisch begleitet (Dold, Lenz, Demal, & Aigner, 2010; Padberg, 2012; Scheidt et al., 2012). Die Entwicklung von Software, die nicht nur eine reine Dokumentation von soziodemografischen Variablen und Kassendaten der Patienten erlaubt, sondern darüber hinaus auch Status- und Verlaufsrückmeldungen erlaubt, ist mittlerweile vorhanden (Grawe & Baltensberger, 1998; Hänsgen, 2006; Lutz, 1997; Steinkamp & Schulte, 2008; cibait iQ/5[29]; psychoEQ[30]). Ein zentraler Faktor, der über den Einsatz solcher Instrumente entscheidet, ist die Praktikabilität und Effizienz (Gilbody et al., 2002b; D. R. Hatfield & Ogles, 2004). Daher wären Möglichkeiten, die Erhebungen möglichst kurz zu halten in der Praxis sehr erwünscht, aber z.B. der Schritt, in Dokumentationssystemen auch die Messqualität durch die Einbettung von IRT-Modellen oder computer-adaptive Tests zu erhöhen, wurde noch nicht vollzogen. Und selbst dann bliebe fraglich, ob die so effizienteren Erhebungen einen Eingang in die Praxis finden würden, da eine große Breite der Psychotherapeuten die notwendige Infrastruktur nicht bereitstellen kann. Um gerade Wissenschaftler-Praktiker-Netzwerke und eine größere

[29] http://www.cibait.net/ (13.12.2012)
[30] http://www.psychoware.de/index.php/psychoeq.html (13.12.2012)

Evidenzbasierung auch in solchen Kontexten etablieren zu können, müssen daher andere Wege angedacht werden, die diese Probleme angehen (Chang & Reeve, 2005).

3.1.1. Ziele der Studie

Das Ziel der Studie ist es, die beiden oft als gegensätzlich konstruierten Positionen der Klassischen Testtheorie und IRT zu einem Zweck zu kombinieren. Auf der einen Seite sind in Praxissettings Erhebungen der relevanten Dimensionen nötig, die kurz und einfach durchzuführen sind. Dies wird durch Tests, die nach Klassischer Testtheorie konstruiert und vertrieben werden, erreicht: Es ist möglich Papier und Bleistift zu benutzen und der Summenscore ist mit Hilfe von Normtabellen einfach zu interpretieren. IRT-Techniken dagegen erlauben eine flexible Kombination von Items, die für einen bestimmten Erhebungszweck relevant sind, doch sind diese nicht ganz so einfach umzusetzen (Bechger, Maris, Verstralen, & Béguin, 2003; Cook et al., 2008; DeVellis, 2006; Zickar & Broadfoot, 2009). Das zentrale Ziel dieser Untersuchung ist daher die Entwicklung von IRT-basierten Kurzformen, die sich das Prinzip der Verwendbarkeit einzelner Items nach der IRT-Skalierung einer Skala zu Nutze machen, aber ähnlich einfach einzusetzen und auszuwerten sein sollen, wie ein Bogen nach Klassischer Testtheorie.

Die Schritte zur Erreichung dieses Ziels werden an einer existierenden Skala zur Messung psychologischer Belastung vorgenommen und daher muss dieses globale Ziel in Unterziele zerlegt werden. Dies ist zunächst die Identifizierung derjenigen Items dieser Skala, die als IRT-skalierbar angesehen werden können. Hierzu werden die Items auf Eindimensionalität, Monotonie, Ordnung der Schwellenparameter und Differential Item Functioning (Osterlind & Everson, 2009) untersucht. Anhand der Items, die diese Bedingungen erfüllen und damit als IRT-skalierbar angesehen werden können, sollen dann zwei Kurzversionen konstruiert, die für die Zwecke der Belastungsmessung im hohen Belastungsbereich und die Entscheidung im Screeningintervall besonders geeignet sind.

Zwei letzte Ziele sind dann die Demonstration der Verwendung der Kurz- und Langformen an einem Fallbeispiel und die Diskussion unterschiedlicher Vorgehensweisen bei der Auswahl von Items, die entweder eher populations- oder kriteriumsorientiert sind.

Die Entwicklung der Kurzfassung wird für den "Fragebogen zur Evaluation von Psychotherapieverläufen" (Lutz, Schürch, et al., 2009) vorgenommen. Der Fragebogen wurde an einer Stichprobe von $N = 708$ bestehend aus ambulanten Therapiefällen und nicht-behandelten Personen erhoben. Diese Stichprobe wurde zufällig in zwei Gruppen geteilt, eine zur Schätzung der Parameter der Modelle (Schätzstichprobe) und eine zur Überprüfung der Gültigkeit dieser Parameter (Validierungsstichprobe). Als IRT-Modell wurde das "Partial Credit"-Modell verwendet (Masters, 1982), das sich zur Modellierung polytomer Antworten eignet. Nach der Schätzung und Validierung des Modells, werden zwei Subgruppen von Items aus der Skala gewählt: a) der Auswahl von Items für eine Kurzfassung zum Screening von Personen und b) für eine Kurzfassung zur wiederholten Messung des Verlaufs.

3.2. Methoden

3.2.1. Stichprobe

Insgesamt wurden die Antworten einer Stichprobe von $N = 788$ Personen genutzt. Diese kamen aus zwei verschiedenen Gruppen, einer ambulanten Behandlungsstichprobe und einer Bevölkerungsstichprobe. Die ambulante Stichprobe bestand aus $n = 426$ Aufnahmemessungen in der Psychotherapieambulanz der Universität Trier. Die Patienten wurden diagnostiziert mittels des Strukturierten Klinischen Interviews für DSM IV (First, Spitzer, Gibbon, & Williams, 1996) und die häufigsten Diagnosen waren Majore Depressionen (296.3; 24.6%), Angststörungen (300.3 ohne PTSD und Zwang; 15.9%) und akute Belastungsstörungen (308.3; 13.0%). Diese Stichprobe war im Mittel 36.5 ($SD = 12.4$) Jahre alt und 62.1% der Stichprobe waren weiblich.

Die nicht-klinische Stichprobe bestand aus $n = 362$ Personen, die sich derzeit nicht in Psychotherapie befanden. Sie wurden zwischen Januar 2009 und Oktober 2009 im Westen von Deutschland erhoben (im Rahmen von drei Diplomarbeiten: Bottler, 2009; Jung, 2008; Mockenhaupt, 2009). Diese Stichprobe war in Bezug auf Alter (in Gruppen von 18-30, 31-50 und 51-65 Jahre), Geschlecht und Schulabschluss (Hauptschule, Realschule, Gymnasium) an die im Mikrozensus

erhobenen Verteilungen angepasst (Lüttinger & Riede, 1997)[31]. Das mittlere Alter dieser Stichprobe war 43.3 (*SD* = 12.4) Jahre und 49.7% waren weiblich.

Von den $N = 788$ Personen lagen für $n = 708$ komplette Fragebogen vor (klinisch: $n = 359$; 84.3%; nicht-klinisch: $n = 349$; 96.4% der Gesamtstichprobe). Diese Personen wurden in der folgenden Analyse benutzt. Die Stichprobe wurde mittels einer binomialen Zufallsvariable ($\pi = .5$) in zwei Gruppen geteilt ($n_1 = 355$; $n_2 = 353$). Die erste Stichprobe ist die Schätzstichprobe in der alle Modellschätzungen, Itemausschlüsse etc. vorgenommen werden. Die zweite Stichprobe ist die Validierungsstichprobe, in der die Resultate aus der Schätzstichprobe überprüft werden.

3.2.2. Instrument

Das verwendete Instrument ist der "Fragebogen zur Erfassung von Psychotherapieverläufen", der ein *public domain* Instrument zur Erfassung psychischer Belastungen ist (Lutz & Böhnke, 2008; Lutz, Schürch, et al., 2009). Der Fragebogen hat gute Reliabilitätswerte ($\alpha = .96$) und gute konvergente Validitäten mit anderen existierenden Messinstrumenten wie dem "Outcome Questionnaire" (OQ-45; $r = .81$; Lambert, Hannöver, Nisslmüller, Richard, & Kordy, 2002; Lambert et al., 1996) und dem "Treatment Evaluation and Management" (TEaM; $r = .78$; Grissom et al., 2002). Eine erste Konstruktvalidierung nach dem Rasch-Modell zeigte, dass für 35 der 40 Items gesagt werden kann, dass sie Rasch-skalierbar sind (Schürch et al., 2009). In keiner der vorigen Studien zur Validität des Fragebogens wurden die Daten der vorliegenden Studie verwendet.

Der Fragebogen erfasst vier Dimensionen psycho-sozialer Belastung. Wohlbefinden, Symptome und interpersonelle Probleme orientiert am Phasenmodell der psychotherapeutischen Veränderung (Howard et al., 1993). Zusätzlich wird in Anlehnung an Grawe (1998, 2004) Inkongruenz erfasst: Motivationale Inkongruenz zwischen verschiedenen Zielen wird demnach als eine Ursache oder aufrechterhaltender Faktor psychischer Störungen oder dysfunktionaler Denkmuster und Ver-

[31] Der Mikrozensus ist eine jährliche Erhebung von 1% der deutschen Bevölkerung und es wurden die Daten des 2002er Campus-Use-Files verwendet (Statistische Ämter des Bundes und der Länder. Datenangebot | CAMPUS-Files. Retrieved May 24, 2012, from http://www.forschungsdatenzentrum.de/campus-file.asp)

haltensweisen gesehen. Obwohl der Fragebogen auf die Erfassung von vier Dimensionen abzielt, sind diese vier Dimensionen v.a. von einem allgemeinen Belastungsfaktor geprägt (Lutz & Böhnke, 2008; Lutz, Schürch, et al., 2009), der als allgemeine Psychopathologie/psychische Belastung bezeichnet werden kann. Dieser Befund liegt auch für viele der Referenzinstrumente vor (Barkham et al., 2001; Halstead et al., 2007; Meijer et al., 2011; Tran, Walter, & Remmel, 2012). Die einzigen Items, die bereits in der Vergangenheit als reliabel eine andere Dimension erfassend demonstriert wurden, sind die Items der Skala interpersonelle Probleme. Diese zwölf Items werden in den folgenden Analysen nicht verwendet. Die Reliabilität der verbleibenden 28 Items beträgt Cronbach-α = .96 (Bootstrap 95%-Konfidenzintervall: .96 -.97) in der Schätzstichprobe.

3.2.3. Überprüfung der Modellgeltung

Um Items aus der Skala basierend auf IRT-Prinzipien auswählen zu können, muss zunächst geprüft werden, ob die Items überhaupt als IRT-skalierbar angesehen werden können. Dazu werden üblicherweise die Tests auf Eindimensionalität, Monotonie und differentielle Effekte bestimmter Populationen auf die Beantwortung der Items oder der Skala (*Differential Item/ Test Functioning*) verwendet (Doucette & Wolf, 2009). Weitere Möglichkeiten der Prüfung bestehen, doch können diese als ein Konsens gesehen werden, der sich derzeit als eine Art Mindeststandard durchsetzt (Fliege et al., 2005; Reise, Moore, & Haviland, 2010; Reise, Morizot, & Hays, 2007). Diese Tests werden im Folgenden kurz beschrieben und zwar in der Reihenfolge, in der sie durchgeführt wurden. Eindimensionalität und die differentiellen Effekte können getestet werden, bevor überhaupt ein IRT Modell geschätzt wurde. Die Annahme der Monotonie (d.h. die Ordnung der Itemkategorien entlang des latenten Kontinuums) wird nach der Anpassung des IRT-Modells getestet (für einen anderen Ansatz für dichotome Items: Junker & Sijtsma, 2000). Daher wird das verwendete IRT-Modell (Partial Credit-Modell; Masters, 1982) erst nach den Tests für Eindimensionalität und differentielle Effekte beschrieben.

3.2.4. Eindimensionalität

Eindimensionalität wird mittels der sog. "*Parallel Analysis*" geprüft. Parallel Analysis ist ein Verfahren, bei dem *b* Stichproben simuliert werden (in dieser Arbeit *b* = 500), die in Anzahl Items

und Personen genau gleich der Ausgangsstichprobe sind. Der Unterschied zur Ausgangsstichprobe besteht jedoch darin, dass die Items als (stochastisch) unabhängig voneinander simuliert werden. In jeder dieser Stichproben wird dann eine Faktorenanalyse durchgeführt. Die Eigenwerte der Faktoren werden für jeden der Durchläufe gespeichert. Die Eigenwerte der Faktorenanalyse in der Schätzstichprobe werden dann mit einem Quantil der Eigenwerte aus den simulierten Stichproben verglichen, in dieser Arbeit mit dem 95%-Quantil. Empirische Eigenwerte von Faktoren, die oberhalb dieser Grenze liegen, können interpretiert werden als Faktoren, die mehr Varianz binden als bei rein zufälligen Beziehungen zwischen den Items zu erwarten wäre (Drasgow & Lissak, 1983; Hayton, Allen, & Scarpello, 2004). In dieser Arbeit werden die polychoren Korrelationen verwendet, die lediglich Ordinalität und kein Intervallskalenniveau der Items annehmen. Hierfür werden die R Pakte psych (Revelle, 2010) und random.polychor.pa (Presaghi & Desimoni, 2010) verwendet (R Development Core Team, 2010).

3.2.5. Differential Item und Test Functioning

Eine Fehlerquelle bei der Auswertung von Tests und dem Vergleich der Resultate zwischen Personen ist, dass der ganze Test oder Teile (also Items) von ihm in unterschiedlichen Subpopulationen unterschiedliche Antwortcharakteristika aufweisen. Dies ist mit dem Begriff des *Differential Item Functioning* (DIF) gemeint: Ein Item hat unterschiedliche Lösungswahrscheinlichkeiten in verschiedenen Subpopulationen bei gleicher Fähigkeitsausprägung der Personen (Osterlind & Everson, 2009). Zusätzlich gibt es *Differential Test Functioning* (DTF): Der gesamte Test wird von einer Subpopulation anders bearbeitet als von der anderen (Penfield & Algina, 2006). Liegen DIF oder DTF vor, liegt damit ein Hinweis auf mögliche Multidimensionalität eines Items vor, die sich bei dieser Gruppenteilung bemerkbar macht, das bedeutet, dass dieses Item nicht nur die latente Personeneigenschaft misst, sondern die Antwort auf das Item auch von anderen Einflüssen abhängt, die sich beim Vergleich der Gruppen bemerkbar machen.

Für den Fall von Erhebungen von Belastungs-/Symptommaßen in klinischen und nicht-klinischen Populationen bedeutet dies: Sind in einer klinischen und einer nicht-klinischen Stichprobe zwei Personen mit gleichem Belastungsgrad (Ausprägung auf der latenten Dimension), dann

müssen sie für jedes abgefragte Belastungssymptom dieselbe Wahrscheinlichkeit dafür haben, dass das Symptom vorliegt bzw. bei polytomen Items in derselben Ausprägung. Nur dann kann für diese DIF Variable (z.B. klinisch vs. nicht-klinisch) plausibel ausgeschlossen werden, dass die Unterschiede zwischen den Personen, die in diesem Instrument festgestellt werden, auf Unterschiede in der latent Eigenschaft zurückzuführen sind und nicht auf Unterschiede zwischen den Populationen. Für eine detaillierte Darstellung von DIF und DTF sei auf Osterlind und Everson (2009) sowie Penfield und Algina (2006) verwiesen.

DIF und DTF werden mit der Freeware DIFAS (Penfield, 2005) geprüft. DIFAS stellt verschiedene klassische DIF Tests sowie einen Test für die Bewertung von DTF zur Verfügung und in der Schätzstichprobe wurden die Items mittels der Mantel-Haenszel Prozedur getestet. Diese Prozedur prüft, ob bei gleichbleibender Ausprägung der latenten Variable die Wahrscheinlichkeit des Auftretens der Itemlösung (im polytomen Fall das Auftreten einer Kategorie) die Wahrscheinlichkeit dieses Ereignisses (Itemlösung bzw. bestimmte Kategorie) unverändert bleibt. Für eine Annäherung an die gleiche Ausprägung auf der latenten Variable werden üblicherweise die Scores (oder Scoregruppen z.B. je fünf Punkte) verwendet. Innerhalb dieser Gruppierungen wird dann geprüft, ob die Odds-Ratios konstant bleiben. Dieses Vorgehen ist nicht-parametrisch und wird darüberhinaus als eines der verlässlichsten zur Erkennung von DIF angesehen (Osterlind & Everson, 2009; Wainer, 2010). DTF wurde mittels der v^2-Statistik bewertet. Diese Statistik prüft, ob die Varianz der einzelnen DIF-Effekte der Items insgesamt ungleich "0" ist und damit davon ausgegangen werden muss, dass DIF Effekte einzelner Items das Testergebnis beeinflussen (Penfield & Algina, 2006; Penfield & Camilli, 2007).

Die Überprüfung wurde bezogen auf zwei Kriterien durchgeführt, die im Zusammenhang der Anwendung des Fragebogens in der Routineversorgung wichtig erscheinen. Das erste Kriterium war das Geschlecht, da die Wahrscheinlichkeiten a) eine Diagnose zu haben und b) sich in Behandlung zu befinden mit dem Geschlecht korreliert sind (etwa 1:2 Männer/Frauen). Das zweite Kriterium war die Zugehörigkeit zur Stichprobe (ambulant/Bevölkerung): Wenn der Bogen ggf. zum Screening eingesetzt werden soll, sollte er keine systematischen Vor-/Nachteile gegenüber diesem

Kriterium und damit einer der beiden Populationen haben. Für multiples Testen wurden das α-Fehler-Niveau der einzelnen Tests entsprechend angepasst (Pallant & Tennant, 2007).

3.2.6. Das IRT Modell: Partial-Credit Model (PCM)

Wie oben beschrieben können die Tests für Eindimensionalität und differentielle Effekte durchgeführt werden, bevor ein IRT-Modell an die Daten angepasst wurde. Nach dieser Prüfung . wurde das PCM auf die Daten angewendet (PCM; Masters, 1982). Das PCM gehört zur Familie der Rasch-Modelle (Rost, 2001) und wird verwendet, da diese Modelle den Vorteil bieten, dass der Summenscore die "suffiziente Statistik" ist (Kempf, 2008; Rost, 2004): Die Summe der über alle bei den Items angekreuzten Kategorien enthält alle Information über die Ausprägung der latenten Dimension bei der spezifischen Testperson. Dies ist wichtig für die Anwendung, die im folgenden präsentiert wird, da so sichergestellt wird, dass jeder Summenscore eines Tests (des vollständigen oder einer Kurzform) direkt mit einem Personenparameter verbunden werden kann. Nur so kann eine einfache Umrechnung von Testergebnis in die Metrik der latenten Dimension sichergestellt werden (Thissen & Wainer, 2001). Das PCM erreicht diese Eigenschaft darüber, dass nur die Schwellenparameter der Items (s.u.) variieren dürfen, nicht aber z.B. die Trennschärfen, die die Bildung einer gewichteten Summe über die Items für die Bestimmung des latenten Personenwertes nötig machen würden (Embretson & Reise, 2000; Hambleton et al., 1991). Die Modellformel ist:

$$P(X_i = x) = \frac{\exp \sum_{k=1}^{x} (\theta_v - \tau_{ki})}{1 + \sum_{x=1}^{m} \exp \sum_{k=1}^{x} (\theta_v - \tau_{ki})} \qquad \textbf{Formel 3-1}$$

Das PCM geht von einer Beziehung der Schwellen τ_{ki} zwischen allen Kategorien *k* Items *i* und der latenten Dimension aus. Die Schwellen sind dabei definiert als der Punkt auf der latenten Dimension, an dem eine Person v mit exakt dieser latenten Ausprägung (θ_v, "Personenparameter") die gleiche Wahrscheinlichkeit hat in den beiden an diese Schwelle grenzenden Kategorien des Items *i* zu fallen. Die Wahrscheinlichkeit $P(X_i = x)$ ein item *i* in einer bestimmten Kategorie *x* zu beantworten, ist dann also abhängig davon, ob der Personenparameter oberhalb der Schwelle liegt (Zäh-

ler in Formel 3-1) und in welchem Verhältnis dieser Wert zu den anderen Schwellenparametern zu sehen ist (Nenner in Formel 3-1).

Das PCM wurde mit dem ltm Paket (Rizopoulos, 2006) geschätzt. Das Paket ltm schätzt verschiedene IRT-Modelle mittels Marginal Maximum Likelihood. Studie I (Kapitel 2) zeigte, dass die Unterschiede zu anderen Schätzmethoden eher gering sind und dieses R-Paket ermöglicht die einfache Bestimmung der Informationsfunktionen, die unten weiter erläutert werden.

3.2.7. Monotonie

In den vorigen Analysen auf Eindimensionalität und das Nicht-Vorhandensein von DIF/DTF sollte sich gezeigt haben, dass das PCM überhaupt die Möglichkeit hat, die vorliegenden Daten angemessen zu beschreiben. Als letztes steht nun noch die Frage aus, ob die Bedingung der Monotonie erfüllt ist (Doucette & Wolf, 2009). Bei der Überprüfung der Modelleigenschaft der Monotonie geht es darum, ob die Schwellenparameter für jedes Item a) entlang der latenten Dimension geordnet sind und b) jede Kategorie eine Wahrscheinlichkeit besitzt, überhaupt benutzt zu werden. Niedrige ´Kategorien eines Items sollten also mit niedrigen Abschnitten auf der latenten Eigenschaft korrespondieren und hohe Kategorien entsprechend mit hohen Ausprägungen.

Es wird derzeit darüber debattiert, wie wichtig diese Bedingung tatsächlich zur Beurteilung der Modellpassung ist (Adams et al., 2012; Cho, Cohen, Kim, & Bottge, 2010; Linacre, 1999). Prinzipiell wären drei verschiedene Ergebnisse bei einem jeden Item möglich:

a) die Schwellen sind geordnet, wie sie es nach der Konstruktion der Skala sein sollten (von der niedrigsten bis zur höchsten) und jede Kategorie hat eine Wahrscheinlichkeit an einem bestimmten Punkt des latenten Kontinuums ausgewählt zu werden (Abbildung 1-2);

b) die Schwellen sind korrekt geordnet aber es gibt Kategorien, deren Wahrscheinlichkeit nie größer ist als die einer anderen Kategorie, ausgewählt zu werden (Abbildung 1-2);

c) die Schwellen sind nicht geordnet und auch die Wahrscheinlichkeit der Auswahl der Kategorien ist nicht in monotoner Weise mit einem Anstieg des latenten Kontinuums verbunden. Nur Items, die Bedingung c erfüllen, wurden entfernt.

3.2.8. Validierung der Modellpassung

Alle Analysen bis hierhin werden mit der Schätzstichprobe durchgeführt. Mit ihr werden die Items entfernt, die die Kriterien nicht erfüllten. Die Überprüfung, ob die verbliebenen Items gemäß dem IRT-Modell passen, vervollständigt den Prozess der Modellprüfung. Für diesen Zweck werden b = 500 Bootstrap-Stichproben mit je n = 300 (150 klinische, 150 nicht-klinische) aus der Validierungsstichprobe gezogen. In diesen wurde dann das PCM zweimal geschätzt:

a) einmal basierend mit den Parametern, die sich aus der Schätzstichprobe ergaben;

b) einmal neu angepasst auf die Fluktuationen der Validierungsstichprobe.

In jeder dieser 500 Replikationen werden die geschätzten Personenparameter für die Personen aus a) und b) miteinander korreliert. Dabei geben Spearman-Rang Korrelationen an, wie sicher die Reihenfolge entlang der latenten Dimension repliziert wird. Niedrige Korrelationen oder eine breite Streuung der Korrelationen würden bedeuten, dass die Ordnung der Items, wie sie im Schätzsample identifiziert wurde, nicht das widergibt, was sich an Variation in der Validierungsstichprobe ergibt. Auch variierende Itemschwierigkeiten oder Schwellenparameter würden dazu führen, dass die geschätzten Personenparameter in Prozess b) anders ausfallen als in a). Korrelationen der Personenparameter sind auch aus diagnostischer Sicht ein interessantes Kriterium, da die Rangkorrelationen angeben, ob sich ähnliche Ordnungsreihenfolgen ergeben, wenn unterschiedliche Schätzergebnisse verwendet werden und die Produkt-Moment-Korrelationen geben im Vergleich dazu an, wie exakt gleich die geschätzten Personenparameter sind.

3.2.9. Die Entwicklung von Kurzformen

Wenn gezeigt ist, dass das PCM die vorhandenen Daten angemessen beschreibt, wird die Auswahl von Items mittels der Informationsfunktionen für bestimmte Erhebungszwecke an zwei Beispielen. Bei dem ersten Beispiel geht es um die Auswahl von Items, die das Instrument für

Screenings optimieren, die also die Entscheidung ob ein Fall eher aus der klinischen oder der nicht-klinischen Stichprobe stammt, stützen sollen. Beim zweiten Fall geht es darum, die verbleibenden Items so auszuwählen, dass sie optimal über den ganzen Bereich der Ausprägungen in der klinischen Stichprobe verteilt sind, wie es beispielsweise bei wiederholten Erhebungen nützlich sein kann. Für beide sollen im Folgenden die Auswahlmethoden und -kriterien beschrieben werden.

Ein Element von IRT-Modellen, das im Folgenden benutzt wird, ist die Informationsfunktion, die bei einem IRT-Modell die Messgenauigkeit angibt: Diese ist entlang der latenten Dimension nicht konstant (wie es in der Klassischen Testtheorie angenommen wird, eine Reliabilität, egal, welcher Score erreicht wird), sondern sie verändert sich entlang des Kontinuums abhängig von Qualität und Lage der Items. Jedes Item besitzt eine eigene Informationsfunktion (siehe Abbildung 1-3) und diese beschreibt, wie hoch die Messgenauigkeit eines einzelnen Items an jedem Punkt der latenten Eigenschaft ist. Hohe Werte bedeuten dabei eine höhere Präzision (= niedriger Standardfehler). Alle Iteminformationsfunktionen können außerdem zu einer Testinformationskurve integriert werden. Diese gibt dann Auskunft darüber, wie hoch die Messgenauigkeit des Gesamttests am jeweiligen Punkt des latenten Kontinuums ist (s. Abbildung 2-1; Hambleton et al., 1991; Thissen & Wainer, 2001).

Um Items für Kurzversionen auszuwählen, wird im Folgenden so vorgegangen, dass Abschnitte auf der latenten Dimension definiert werden, die als interessant angesehen werden. Für diese im folgenden als "Zielregion" bezeichneten Abschnitte werden dann die Items ausgewählt, deren Messgenauigkeit in diesem Bereich besonders hoch ist, d.h. die besonders hohe Iteminformationsfunktionen in diesem Bereich haben. Dies wird operationalisiert als die Fläche unter der Kurve innerhalb der Zielregion: Da der Bereich auf der x-Achse für alle Items konstant ist, ist die Fläche nur von der Höhe der Informationsfunktion abhängig und größere Flächen damit gleichbedeutend mit höherer Messgenauigkeit.

Die erste Kurzskala wird daraufhin optimiert, in dem Bereich des Cut Offs zwischen der Normstichprobe und den Behandlungsfällen zu unterscheiden. Hierfür wird zunächst ein nicht-parametrischer Algorithmus genutzt, um den Cut Off zwischen diesen beiden Stichproben in der

Schätzstichprobe zu finden. Die Verteilungen der beiden Gruppen sind distinkt in ihrer Form und Lage und weisen auch eine große Effektstärke in der Differenz auf (s. Abbildung 3-1). Der Algorithmus bestimmt den Cut Off zwischen den Stichproben, indem er die beiden kumulierten Häufigkeitsverteilungen entlang der latenten Dimension vergleicht:

$$F(x > t_{clin}) \leq F(x \leq t_{non\text{-}clin}) \qquad \textbf{Formel 3-2}$$

wobei *F()* die kumulierte Häufigkeitsverteilung der Fälle angibt, die entweder gleich oder unterhalb ($x \leq t$) bzw. über ($x > t$) einem spezifischen Wert auf der latenten Dimension liegen. Die Werte der latenten Dimension *t* werden in aufsteigender Folge iterativ durchgegangen und als Cut Off wird dasjenige *t** definiert, an dem die Ungleichung als letztes erfüllt ist (Klotsche, Ferger, Pieper, Rehm, & Wittchen, 2009).

Um diesen Cut Off wird ein Bootstrap Konfidenzintervall konstruiert. Hierzu werden $b = 500$ Stichproben mit jeweils $n = 300$ (150 klinische/150 nicht-klinische Fälle) gezogen. Der Bereich des Konfidenzintervalls (die zentralen 95% der geschätzten Cut Offs) definieren in dieser Anwendung die Zielregion, auf die die Itemauswahl des Kurztests optimiert werden wird: Der Bereich der latenten Eigenschaft, in dem der Cut Off mit großer Plausibilität liegt. Die Items, die für diesen Bereich die meiste Information zur Verfügung stellen, werden zusammen den kleinstmöglichen Standardschätzfehler produzieren.

Die zweite Kurzversion soll für die Erhebung in der klinischen Stichprobe geeignet ist. Eine solche Kurzform ist hilfreich in der Verwendung in Bereichen der Qualitätssicherung, Patient Reported Outcome, Patientenorientierter Versorgungsforschung oder Feedback-Anwendungen. Im Ganzen immer dort, wo eine Veränderung im für die klinischen Fälle typischen Belastungsspektrum stattfinden sollte und dies mit wiederholten Messungen getestet wird. Um Veränderungen in dem Bereich effektiv messen zu können, wird die Verteilung der Personenparameter in zwei Bereiche aufgeteilt: 1) vom 95%-Perzentil der klinischen Fälle bis zum 50%-Perzentil als Bereich der hohen Belastung; 2) vom 50%-Perzentil bis zum 2.5%-Perzentil als Bereich niedriger Belastung – jeweils relativ zu den Fällen der klinischen Stichprobe gesehen. Durch diese Einteilung des latenten

Kontinuums wird sichergestellt, dass die Items nicht alle auf einen Belastungsbereich fallen, denn eine reine Auswahl nach den insgesamt höchsten Informationsfunktionen könnte in Items resultieren, die lediglich in einem Bereich der latenten Beschwerdedimension liegen.

Die Items zu identifizieren, die sich für die Messung in der jeweiligen Zielregion besonders eignen, ist sehr ähnlich zu Fragen in einem diagnostischen Interview: Wenn ein Patient angibt, dass er guter Stimmung ist und keine Probleme innerhalb der letzten Woche hatte, wird die Frage nach Suizidgedanken vermutlich nicht viel neue Information über den Status des Patienten ans Licht fördern, da dies in den meisten Fällen nach einem sehr hohen Belastungsniveau fragt. Eine Frage nach spezifischen, eher niederschwelligen Symptomen wie beispielsweise "Schlafproblemen" dagegen eher schon (Kernargument bei computer-adaptivem Testen; siehe z.B. auch: Fliege et al., 2005; Orlando, Sherbourne, & Thissen, 2000; Wainer, 2000).

Beide Kurzversionen werden kreuzvalidiert (Hawkins, 2004; J. Stevens, 1996). Das Argument für die Auswahl der Items ist, dass die Messqualität einer gerichteten Itemauswahl höher sein soll, als die anderer möglicher Itemkombinationen. Um dies zu überprüfen muss ein Weg gefunden werden, um zwischen der Stichprobenvariation bei der Schätzung von Parametern und dem Messfehler der Items zu unterscheiden. Dazu wird die die Auswahl der Items an der Schätzstichprobe vorgenommen. In der Validierungsstichprobe werden dann wieder $b = 500$ Bootstrap Stichproben gezogen, um zu überprüfen, ob die ausgewählten Items (im Folgenden "Zielitems") tatsächlich eine insgesamt höhere Messqualität aufweisen, als andere Items. Als Vergleichsgruppe werden hier zwei Auswahlen zufälliger Items verwendet:

a) ein Set aus Items, das aus gleich vielen zufällig aus allen Items gezogenen Items besteht; dies ist ein eher strenges Kriterium, da jede der Realisierungen auch eines oder mehrere der Zielitems enthalten kann und damit die Messqualität im Prinzip auch sehr hoch sein kann; dies ist auch das realistischste Kriterium, da es einer zufälligen Auswahl von Items ohne Vorwissen entspricht;

b) ein Set, das aus gleich vielen Items besteht, die aber zufällig nur aus den Items der Skala gezogen werden, die *nicht* Zielitems sind; eher leichteres Kriterium, da so nur getestet wird, ob die Messqualität höher ist als bei den ausgeschlossenen Items.

Ob die Messqualität höher ist, wird in jedem der Samples dadurch bestimmt, dass für alle drei Itemsets (Zielitems, zufällige Items, zufällige nicht-Zielitems) die Fläche unter der Informationsfunktion bestimmt wird. Diese Flächen werden für jeden der Durchgänge durch eine parametrische und eine nicht-parametrische Prozedur verglichen. Parametrisch wird mit einem abhängigen t-Test getestet, ob die die Differenz der Flächen tatsächlich von Null verschieden ist. Nicht-parametrisch wird für jeden Durchgang gespeichert, ob die Fläche unter der Kurve der Zielitems größer war, als die in der Vergleichsgruppe. Eine reliable Kurzversion sollte in beiden Kriterien besser abschneiden als beide Vergleichsgruppen (zufällige Items, zufällige nicht-Zielitems).

3.3. Ergebnisse

Die Ergebnisse der Modellanpassung wie auch der Prüfung der Kurzskala werden in derselben Reihenfolge präsentiert wie im Methodenteil. Zunächst wird das IRT-Modell in der Schätzstichprobe getestet (Eindimensionalität, DIF, DTF, Monotonie). Die Ergebnisse, d.h. die als positiv bewerteten Items werden dann im Validierungssample getestet. Als dritter Teil folgt dann die Auswahl der Items für Kurzversionen, mit den Anwendungsbeispielen des Screenings und der Skala für wiederholte Messungen im Bereich der klinischen Belastung.

3.3.1. Eindimensionalität

Als Ergebnis der Parallelanalyse der polychoren Korrelationen ergab sich ein erster starker Faktor mit einem Eigenwert von 15.87 (zweiter = .97; dritter = .81). Wird dies mit den 95% Quantilen der simulierten Datensätze verglichen, zeigt sich, das prinzipiell drei Faktoren in den Daten identifiziert werden könnten: Der erste simulierte Eigenwert beträgt .79 (zweiter = .68; dritter = .61). Bei dem vierten Faktor liegt das 95%-Perzentil der simulierten Daten (.54) bereits oberhalb des Wertes für die empirischen Daten (.48); daher ist davon auszugehen, dass dieser Faktor nicht mehr relevant Varianz bindet, die über zufällige Variation hinausgeht. Nach einer Varimax-Rotation wurden verschiedene Eigenschaften der Items geprüft. Zunächst war auffällig, dass alle

Items positive Ladungen auf allen Faktoren hatten, was auf eine starke Verbindung zwischen den Faktoren hindeutet. Die Items mit hohen Ladungen auf dem ersten Faktor und zusätzlichen Ladungen auf den Faktoren zwei und drei, waren insbesondere solche Items, die positiv formuliert sind und dementsprechend für die Kodierung umgedreht werden. Daher werden alle Items, die dieses Muster zeigen als hinreichend eindimensional gesehen und in der Skala behalten.

Die einzigen Items, die mit dieser Interpretation nicht abgedeckt werden können, sind die Items 19 ("...war ich selbstbeherrscht") und 27 ("...war ich Teil einer erfüllten und intimen Beziehung"): Beide zeigen keine Ladung auf dem ersten Faktor (.18 bzw. -.01). Für Item 27 wird außerdem eine insgesamt eher niedrige Kommunalität entdeckt (h^2 = .08). Beide Items werden als nicht hinreichend eindimensional mit den anderen Items bzw. überhaupt stark genug mit den anderen Items korreliert angesehen und von der Skala entfernt. Nach der Entfernung der zwei Items liegen die Ladungen der Items auf dem ersten Faktor zwischen .62 (Item 2, "...ging ich vielen Interessen nach") und .91 (Item 38, "...war ich deprimiert und niedergeschlagen") und dieser erste Faktor klärt 60% der Varianz auf. Die Interpretation, dass die zwei weiteren Faktoren v.a. auf die Formulierungen zurückzuführen sein können, wird auch dadurch gestützt, dass nach einer obliquen Rotation (*oblimin*) die drei extrahierten Faktoren zwischen r = .62 und r = .77 miteinander korrelieren.

3.3.2. DIF Analysen

Mit den verbleibenden 26 Items wurden DIF und DTF Analysen durchgeführt. Als Faktoren, die das Beantwortungsverhalten nicht beeinflussen sollten, wurden das Geschlecht und die Gruppenzugehörigkeit zu nicht-klinisch vs. klinisch geprüft. Der Summenscore wurde für die Mantel-Haenszel wie auch die DTF-Statistik als Stratifizierungsvariable verwendet. Er wurde dazu in Scoregruppen von je 5 Skalenpunkten geteilt (mögliche Spannweite des Scores 0 bis 104; empirisch realisierte Spannbreite: 7 bis 96).

Da zwei unabhängige Tests (die beiden DIF-Variablen) und 26 genestete Folgetests (innerhalb jeder Gruppierung alle Items) durchgeführt werden, wurde das α-Fehler-Niveau angepasst: Auf $p < .025$ für die zwei Gruppen und innerhalb dieser dann zu einem Signifikanzniveau von $p < .00097$ ($\chi^2 > 10.88$) für jeden Einzelvergleich (Pallant & Tennant, 2007).

In Bezug auf die beiden Teilungskriterien wurden für das korrigierte Signifikanzniveau keine DIF-Effekte gefunden. Dennoch erwiesen sich einige der Werte als im Vergleich zu den anderen Items unerwartet hoch. Dies waren die Items 15, 21 und 30. Item 15 (χ^2 = 10.02; "...war ich selbstsicher und selbstbewußt") war leichter für die klinische Stichprobe und Item 21 (χ^2 = 10.63; "...empfand ich einen Sinn in meinem Leben") war leichter für die nicht-klinische Stichprobe. Item 30 wurde in beiden Gruppenanalysen mit leicht erhöhten Werten auffällig (leichter für die klinische Stichprobe, χ^2 = 7.00; leichter für männliche Teilnehmer, χ^2 = 5.03; "...war ich voll innerer Ruhe").

Beide Tests für DTF (Geschlecht: v^2 = 1.61; Gruppe: v^2 = 1.78) waren als nicht-signifikant einzustufen (Penfield & Algina, 2006), was bedeutet, dass diese beiden Variablen keinen Effekt auf das Ausfüllen des FEP-26 zu haben scheinen. Daher wurden in diesem Schritt keine weiteren Items ausgeschlossen. Die Items 15, 21 und 30, die in den DIF-Analysen auffällig wurden, werden nicht in Kurzfassungen verwendet werden. Ihre erhöhten DIF-Statistiken weisen darauf hin, dass sie unter Umständen leichte Effekte auf die Bearbeitung der Skala haben (Einzeltests für DIF), diese aber durch die anderen Items ausgeglichen werden (nicht-signifikante Gesamtstatistik für DTF). Da nicht zwangsläufig davon ausgegangen werden kann, dass die ausgleichenden Items auch bei einer Kurzfassung ausgesucht werden und somit diesen ausgleichenden Effekt haben, ist es ratsamer, diese Items nicht zu verwenden (Tennant & Pallant, 2007).

3.3.3. Schätzen des Partial Credit Modells (PCM)

Die Schätzung des PCM wurde in der Schätzstichprobe durchgeführt. Bezogen auf die 26 Items kann die Spannweite Breite des abgedeckten Belastungsbereiches dadurch untersucht werden, dass man betrachtet, wo die niedrigsten Schwellen der Items und wo die höchsten liegen. Die niedrigsten Schwellenparameter der Items befinden sich zwischen -3.84 (Item 5; "...fühlte ich mich unbelastet und zufrieden") und 2.18 (Item 34; "...dachte ich daran, mir das Leben zu nehmen") auf der latenten Beschwerdedimension. Die Itemparameter werden bei der verwendeten Schätzmethode auf den Mittelwert der gesamten Stichprobe genormt und die Skala kann in etwas als Standardabweichungen interpretiert werden (Bond & Fox, 2007). Die unteren Schwellen der Items decken damit allein einen Bereich von knapp 6 *SD* der latenten Eigenschaft ab. Die höchsten Schwellen

befinden sich zwischen .76 (Item 6; "...hatte ich Schlafprobleme") und 3.46 (Item 1; "...fühlte ich mich wohl") auf der latenten Dimension. Damit wird durch die Schwellenparameter, an denen die Messgenauigkeit am höchsten ist, eine Spannweite von -3.84 bis 3.46 erreicht, was als eine sehr breite Abdeckung angesehen werden kann (Abbildung 3-1).

Nach der Anpassung des Modells wurde die Monotonie überprüft. Hierfür wurden die Verläufe der Itemcharakteristiken auf die im Methodenteil (3.2.7) genannten Bedingungen überprüft (Doucette & Wolf, 2009). Von den 26 verbliebenen Items zeigte keines das Muster, dass die Kategorien in der beschriebenen Weise ungeordnet verwendet wurden.

Die grafische Betrachtung der Verteilung der Personenparameter in auf der latenten Dimension (geschätzt per *expected a posteriori*; Rizopoulos, 2006) zeigt, dass die meisten Personen der nicht-klinischen Stichprobe unterhalb eines Wertes von "0" liegen (und die meisten der klinischen Population darüber; siehe Cut Off-Schätzung weiter unten). Das "Rug-Plot" in den beiden Panels der Abbildung 3-1 zeigt die geschätzten Schwellenparameter zwischen den Kategorien 1 und 2 (links) bzw. 4 und 5 (rechts). Dies verdeutlicht, dass die Schwellen im Wesentlichen über dieselbe Spannweite der Dimension verteilt sind, wie die Personenparameter aus den beiden Samples. Weder im unteren noch im oberen Bereich der Beschwerdedimension gibt es also zu wenige oder keine Items, die die Belastung erfassen können. Der Mittelwert der Personenparameter in der klinischen Stichprobe beträgt $M = .79$ ($SD = .92$) und in der nicht-klinischen Stichprobe $M = -.82$ ($SD = .82$). Dieser Abstand entspricht einer Effektstärke von $d = 1.96$. Die korrespondierenden Skalenmittelwerte sind 3.34 ($SD = .71$; klinisch) bzw. 2.12 ($SD = .56$; nicht-klinisch), wofür sich eine Effektstärke von $d = 2.18$ ergibt.

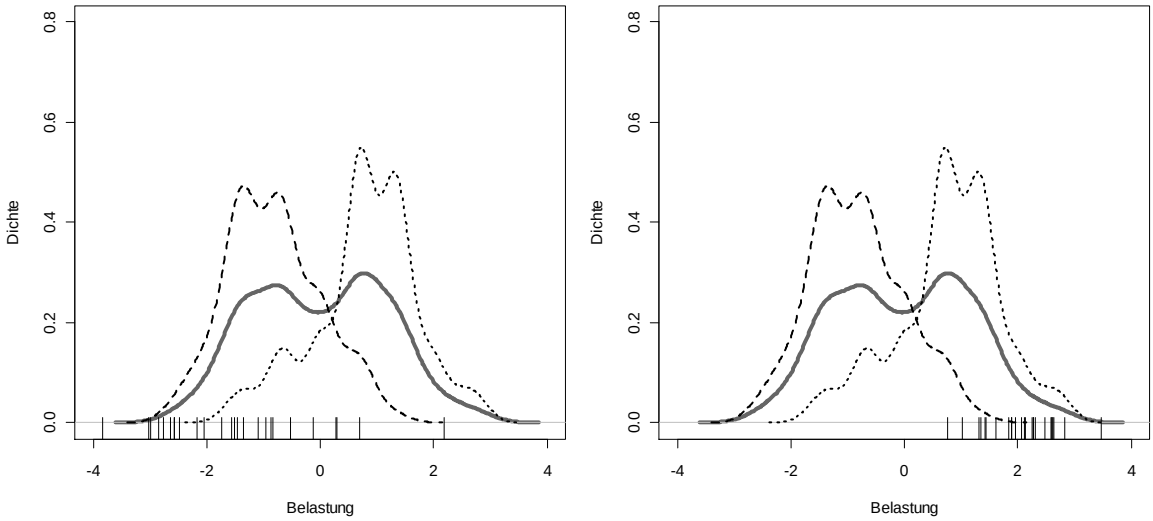

Abbildung 3-1 Verteilung der Personenparameter der klinischen (Punkte) und nicht-klinischen (Striche) Stichproben, sowie die Gesamtverteilung beider Stichproben (graue Linie); das Rug-Plot auf der x-Achse zeigt die Verteilung der ersten Schwellenparameter (zw. Kategorie 1 & 2; links) bzw. der letzten Schwellenparameter (zwischen Kategorie 4 & 5) für alle 26 Items in der Schätzstichprobe.

Eine andere Möglichkeit zu betrachten, wie sich die Items und ihre Mesgenauigkeit über die latente Dimension verteilen, ist, die Iteminformationsfunktion anzugeben (Abbildung 3-2). Hier ist deutlich zu sehen, dass Item 34 stark heraussticht: Das Item zur Frage der Suizidgedanken hat eng beieinander liegende Schwellenparameter und gibt daher sehr viel Information über einen sehr kleinen Bereich der Beschwerdedimension. Das Item kann damit als sehr trennscharfes Item für hohe Belastungsgrade angesehen werden. Es ist in der Abbildung auch zu erkennen, dass die meisten Items das Maximum der Iteminformationsfunktion oberhalb von "0" auf der latenten Dimension haben. Wie in Abbildung 3-1 zu sehen, liegt hier ungefähr eine Trennung zwischen klinischen und nicht-klinischen Fällen vor, d.h. die Items sind insgesamt etwas besser geeignet, die Belastung in der klinischen Stichprobe zu messen. Dieser Eindruck wird auch durch das rechte Panel in Abbildung 3-2 bestätigt: Dort ist die Testinformationsfunktion abgetragen, die alle Item Informationskurven zu einer Angabe über die Messgenauigkeit der Items aggregiert. Das Maximum dieser Kurve liegt oberhalb von "0". Die beste Performanz der Skala liegt etwa zwischen 0 und .5 logits (siehe auch unten).

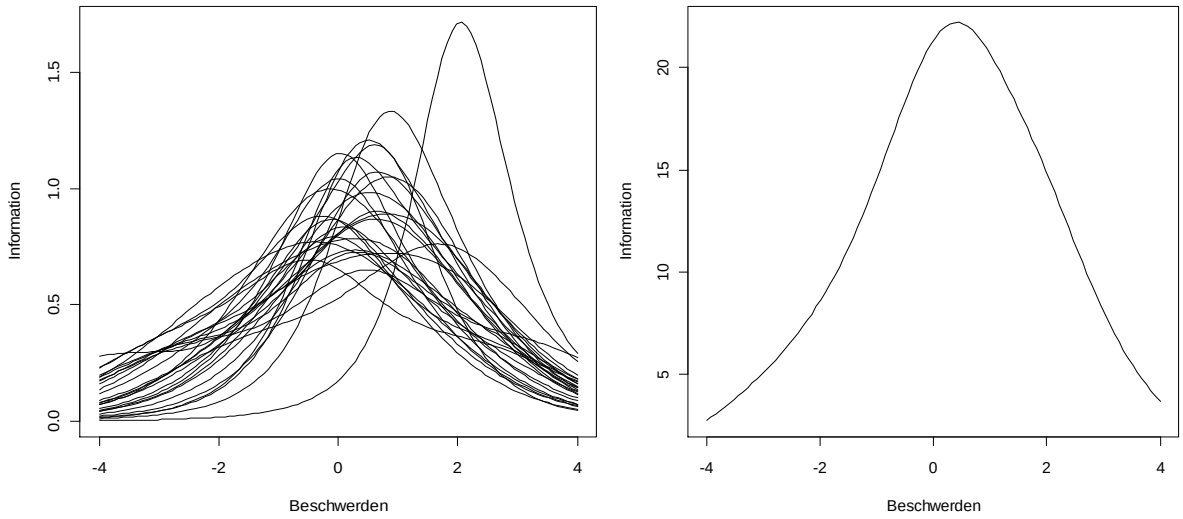

Abbildung 3-2: Iteminformationsfunktionen (links) und die Testinformationsfunktion (rechts) aller 26 Items in der Schätzstichprobe.

3.3.4. Kreuzvalidierung

Die Schätzung des PCMs in der Schätzstichprobe zeigte, dass alle außer zwei Items (19 und 27) als hinreichend eindimensional betrachtet werden konnten (Parallelanalyse). Keines der Items zeigte in relevanter Weise DIF; die Items sind über ein breites Spektrum der psychischen Beschwerden verteilt. Und die Itemschwellen sind hinreichend geordnet (Anpassung des PCM und Monotonie-Bedingung). Das reduzierte Itemset (ohne 19 und 27) hat ein Cronbach-α von .97 (95% Bootstrap-Konfidenzintervall: .96, .97) und das Resultat der Parallelanalyse legt weiterhin einen starken ersten Faktor nahe (Eigenwerte: erster Faktor = 15.59; zweiter Faktor = .87; dritter Faktor = .76). Wenn eine einfaktorielle Lösung an die Matrix der polychoren Korrelationen angepasst wird, zeigt kein Item Ladungen unter .6 auf diesem Faktor (Item 2 mit .62 ist die niedrigste Ladung). Die erklärte Varianz dieses Faktors an den verbliebenen 26 Items beträgt 60%. Bezogen auf diese Kriterien kann die Eindimensionalität als hinreichend angesehen werden.

Für diese Analysen wurde die Schätzstichprobe verwendet. In der Validierung wird geprüft, ob dieses Ergebnis sich auch in der anderen Hälfte der Daten, der Validierungsstichprobe, bestätigen

lässt. Die Parallel Analysis zeigt dass basierend auf polychoren Korrelationen eine Hauptkomponente ausreicht (Eigenwert: 15.07 vs. 1.58 simuliert; Eigenwert der zweiten Hauptkomponente: 1.49 vs. 1.49 simuliert). Ein Faktor erklärt 56% der Varianz und die drei kleinsten Ladungen, die gefunden werden, ergeben sich für Item 2 (.53; "...ging ich vielen Interessen nach"), Item 6 (.59; "...hatte ich Schlafprobleme") und Item 37 (.60; "...fühlte ich mich von anderen im Stich gelassen"). Die Testung mittels logistischer Regressionen auf DIF (lordif; S. W. Choi, Gibbons, & Crane, 2011) zeigt, dass selbst auf dem unkorrigierten α-Niveau von $p < .05$ kein Item mit DIF identifiziert werden konnte. Es wurde im Anschluss eine Monte Carlo Studie durchgeführt um Verteilungswerte zur Relevanzbeurteilung unter Geltung der Null-Hypothese (entspricht "keines der Items zeigt DIF") zu generieren (10000 Stichproben; gewählt wurde Nagelkerkes Pseudo-r^2; S. W. Choi et al., 2011; Crane et al., 2007). Die Ergebnisse zeigen für die Gruppeneinteilung in klinisch vs. nicht-klinisch, dass ein Item knapp oberhalb des simulierten Verteilungswertes liegt (Item 31, "...konnte ich mich für nichts begeistern"), doch ist dieser Effekt in seiner Größe so klein (empirisches Pseudo-$r^2 = 0.0177$; simulierter Verteilungswert Pseudo-$r^2 = 0.0175$), dass er vernachlässigt werden kann. In Bezug auf das Geschlecht war kein Pseudo-r^2-Wert größer als unter der H_0-Verteilung angenommen. Damit kann aufrechterhalten werden, dass in Bezug auf diese beiden Kriterien keine relevanten DIF-Effekte vorliegen.

Als Validierung der Skalierungsstruktur des PCM wurden die Personenparameter in der Validierungsstichprobe zweimal geschätzt:

a) einmal basierend auf den geschätzten Parametern aus der Schätzstichprobe;

b) einmal wurde das PCM in der Validierungsstichprobe neu geschätzt (Item- und Personenparameter).

Diese beiden Schritte wurden in einem Bootstrapverfahren mit $b = 500$ Stichproben durchgeführt, wobei jede Stichprobe aus $n = 300$ Personen (150 klinische, 150 nicht-klinische) bestand. Passt das PCM mit seiner Skalierungsstruktur nicht, werden die Schritte a) und b) zu unterschiedlichen Schätzern der Personenparameter kommen. Es ist zwar richtig, dass beim PCM der Summen-

score die suffiziente Statistik ist und damit prinzipiell eine geringe Variation zu erwarten wäre (der Summenscore ändert sich durch unterschiedliche Modellschätzungen nicht), doch ändert sich die Bedeutung des Summenscores: Der Summenscore ist immer nur eine suffiziente Statistik in Bezug auf die geschätzten Itemparameter, d.h. wenn sich die Reihenfolge der Itemparameter ändert (z.B. ein Item oder die Schwellen eines Items in der einen Stichprobe leichter sind als in der anderen), dann ändert sich auch die Bedeutung des Summenscores (höhere Scores implizieren damit andere Items als höher beantwortet in der einen Stichprobe als der anderen).

Dieser Unterschied würde sich auf der latenten Dimension in unterschiedlichen Personenparametern zeigen: Geringe Korrelationen zwischen diesen beiden Schätzern (keinerlei Übereinstimmung) oder eine breite Streuung der Schätzer (das Modell passt auf unterschiedliche Substichproben unterschiedlich gut) wären Indikatoren dafür, dass die Ergebnisse aus der Schätzstichprobe sich so nicht in der Validierungsstichprobe replizieren lassen und damit die statistische Validität des Modells als eher gering angesehen werden kann: Es fasst die Daten nicht angemessen zusammen. Da beide Schätzungen (a und b) dem Schätzfehler ausgesetzt sind, ist eine direkte Korrespondenz nicht zu erwarten. Als Korrelationsmaße wurden sowohl die Spearman'sche Rang Korrelation als auch die Pearson'sche Produkt-Moment-Korrelation verwendet. Beide kommen zu dem Ergebnis, dass die resultierenden Schätzungen sehr ähnlich sind und die Skalierungsstruktur des PCM damit als replizierbar angesehen werden kann:

- die minimale Rang-Korrelation betrug .9993 (max = .9998);
- die minimale Produkt-Moment-Korrelation betrug .9904 (max = .99995).

3.3.5. Die Konstruktion von Kurzversionen

Die zentrale Frage dieser Arbeit ist wie IRT-Modelle genutzt werden können, um Kurzversionen bereits existierender Fragebögen zu erstellen. Nachdem gezeigt wurde, dass das PCM angemessen die Daten beschreibt, könnte jedes einzelne Item ebenso wie jede beliebige Kombination von Items verwendet werden, um als Kurzversion zu fungieren. Zunächst wird im Folgenden die Selektion von Items für eine Screeningfassung eines Fragebogens vorgenommen und danach für eine Fassung, die zur Veränderungsmessung in der klinischen Stichprobe eingesetzt werden soll.

Abschließend werden andere Möglichkeiten der Definition von Zielbereichen auf dem latenten Kontinuum diskutiert sowie ein Fallbeispiel der Anwendung gegeben.

Beispiel 1

Wie bereits oben in den Effektstärken als Abstand zwischen der klinischen und nicht-klinischen Stichprobe dargestellt, ist der Test mit diesem externen Kriterium hoch korreliert. Die Fläche unter der Receiver Operating Characteristic Curve betrug .90 (DiagnosisMed; Tura, Nicolau, & Oliveira, 2008) was eine zufriedenstellende Diskrimination zwischen den beiden Stichproben durch den Test zeigt. Als Cut Off wurde die im Methodenteil beschriebene iterative Suchprozedur verwendet und der Cut Off wurde geschätzt als -.011. Die Sensitivität für diesen Wert betrug S_e = .82 und die Spezifität S_p = .83, was beides als akzeptabel gewertet werden kann.

Um das Konfidenzintervall um den Cut Off zu schätzen, wurden in der Schätzstichprobe b = 500 Bootstrap Stichproben gezogen (n = 150 aus den beiden Gruppen). In jeder dieser Stichproben wurde der Cut Off nicht-parametrisch geschätzt. Der aus diesen Schätzungen resultierende Mittelwert war -.017 (*Med* = -.011). Die Grenzen des 95%-Konfidenzintervalls waren -.086 und .040. Die Items mit den höchsten Flächenanteilen in diesem Bereich des Konfidenzintervalls waren (Abbildung 3-3): Item 38 ("...war ich deprimiert und niedergeschlagen"; Fläche im Konfidenzintervall unter der Informationsfunktion: .26), Item 35 ("...fühlte ich mich ungenügend und unzureichend"; .24), item 24 ("...fühlte ich mich ohne Wert"; .24), item 29 ("...war ich unabhängig und frei"; .23), und Item 6 ("...hatte ich Schlafprobleme"; .23). Die Angaben der Flächen unter der Informationsfunktion der Items zeigen bereits, warum eine statistische Absicherung gegen Zufallsfehler nötig ist: Die Unterschiede zwischen den einzelnen Items sind eher gering. Dass mit der Auswahl auf zufällige Schwankungen in der Stichprobe optimiert wird, ist also eine mögliche Alternativerklärung.

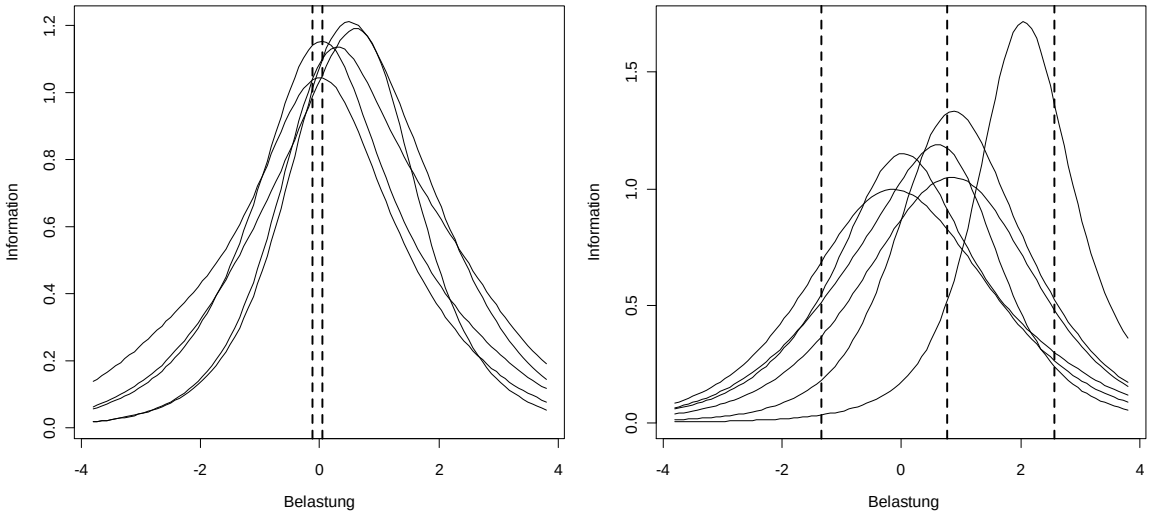

Abbildung 3-3: Zielregionen und ausgewählte Items auf der Beslastungsdimension für die zwei Kurzfassungen in der klinischen Stichprobe; linker Teil der Abbildung zeigt das 95% Bootstrap-Konfidenzintervall für den Cut Off in der Screeninganwendung (Beispiel 1); der rechte Teil zeigt die 2.5%, 50% und 97.5% Perzentile der als Grenzen der beiden Zielregionen für das zweite Beispiel.

Um diese Absicherung zu erreichen, werden die bereits beschriebenen Tests durchgeführt. Die fünf Items werden im Folgenden als "Zielitems" bezeichnet und die anderen 21 Items als "Nicht-Zielitems". Die Auswahl dieser Items wurde in der bereits beschriebenen Weise validiert: In der Validierungsstichprobe wird eine Bootstrap-Stichprobe (150 Personen pro Kriteriumsgruppe) gezogen und in dieser Stichprobe wird das PCM mit dem fünf Items dreimal geschätzt,

a) mit den fünf Zielitems;

b) mit fünf zufällig gezogenen Items;

c) mit fünf zufällig gezogenen Nicht-Zielitems (nur aus den verbleibenden 21 Items).

Das PCM wurde in allen drei Stichproben neu geschätzt, statt die Schätzer aus der Schätzstichprobe zu übernehmen. Damit sollte das Testergebnis noch sensitiver für Unterschiede zwischen Schätz- und Validierungssample sein. Das Kriterium, dass die gewählten fünf Items die höchsten Flächen unter der Informationskurve in dem Bereich des Cut Offs haben, sollte auch über Schätz-

fehler hinaus generalisieren und sich dementsprechend bei Neuschätzungen in einer anderen Stichprobe ebenfalls ergeben. Dass die Schätzer der beiden Stichproben insgesamt hoch korreliert sind und die Ergebnisse sehr ähnlich sind, wurde bereits oben gezeigt. Als Ergebnis zeigte sich, dass in 499 der 500 Durchläufe die fünf Zielitems eine höhere Fläche unter der Informationsfunktion (= geringeren Schätzfehler) erbrachten als die fünf zufällig ausgewählten Nicht-Zielitems (Vergleich a vs. b). Für den Vergleich mit den fünf zufällig gezogenen Nicht-Zielitems (Vergleich a vs. c) ergab sich dieses Ergebnis in allen 500 Durchläufen. Der Vergleich der Flächen unter der Informationsfunktion mittels abhängiger t-Tests (innerhalb aller 500 Durchläufe) zeigte, dass die fünf Zielitems in beiden Vergleichen reliabel größere Flächen unter der Informationsfunktion produzierten (Zielitems vs. zufällige Items: $t = 53.68$; $df = 499$; $p < .001$; Zielitems vs. Nicht-Zielitems: $t = 70.47$; $df = 499$; $p < .001$).

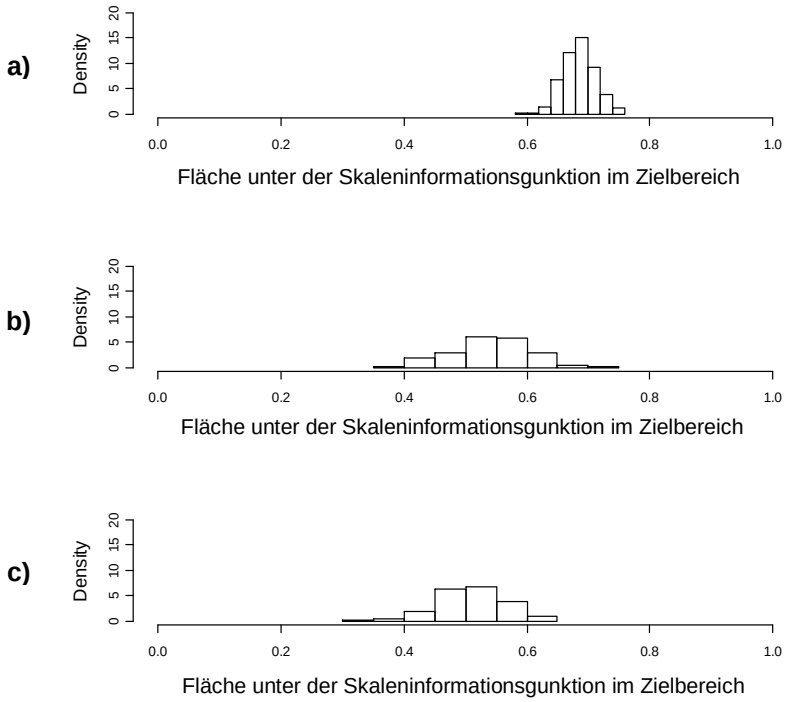

Abbildung 3-4. Dichteverteilungen der geschätzten Flächen unter der Testinformationsfunktion aus den 500 Durchläufen im Zielbereich der Screening-Fassung für die Zielitems (a) und Nicht-Zielitems (b: zufällige Items; c: zufällige Nicht-Zielitems); Beispiel 1.

Abbildung 3-4 zeigt die sich ergebenden geschätzten Flächen unter der Informationsfunktion in den 500 Durchläufen für die Zielitems sowie die beiden Gruppen der Nicht-Zielitems. Die Wirkung des Selektionsprozesses ist deutlich zu erkennen: Die Verteilung der Flächen unter der Testinformationsfunktion für die selegierten Items sind nicht nur im Mittel höher, sondern die Verteilung ist zusätzlich noch deutlich schmaler. Dieses Ergebnis wäre zu erwarten, wenn dieser Test in derselben Stichprobe durchgeführt würde wie die Schätzung der Parameter. Da hier aber eine Kreuzvalidierung vorgenommen wurde, zeigt dies, dass diese Auswahl zumindest für die gewählte zufällige Aufteilung robust gegenüber Zufallsschwankungen ist (Hawkins, 2004).

Beispiel 2

Für wiederholte Messungen im Verlauf der Therapie ist es wichtig, dass der Bereich des latenten Beschwerdekontinuums erfasst wird, in dem sich üblicherweise Veränderungen abspielen. Im Folgenden wird eine Definition dieser Region auf Basis der Verteilung der Population vorgenommen. Da sich die klinischen Fälle verändern sollten, wird ihre Verteilung als Referenz genommen, um den Bereich der möglichen Veränderungen zu erfassen. Dazu werden die 2.5%- und 97.5%-Perzentile bestimmt (alle Perzentile im Folgenden als ihre Mediane aus $b = 500$ Bootstrap-Stichproben). Sie stellen die Maximal- und Minimalwerte der Verteilung dar, die als wahrscheinliche Werte angenommen werden könnten. Innerhalb dieses Bereiches werden zwei Bereiche festgelegt, einmal von 2.5% bis 50% für die niedrige Belastung; und von 50% bis 97.5% für höhere Belastungslagen (2.5%-Perzentil = -1.35; 50%-Perzentil = .76; 97.5%-Perzentil = 2.57; siehe auch Abbildung 3-3). Für beide Bereiche werden die drei jeweils am Besten messenden Items ausgewählt und zu einer Kurzform zusammengestellt, die über das ganze Spektrum der Skala misst. Die Auswahl der Items muss auf diese Weise stratifiziert werden, da eine reine Auswahl nach der Höhe der Informationsfunktionen auch in einer sehr selektiv auf einem Teil des Spektrums liegenden Itemauswahl resultieren kann. Um dies zu verhindern, werden die Bereich definiert, in denen eine Veränderung typischerweise stattfindet und für diese die optimalen Items ausgewählt.

Die Items mit den besten Messeigenschaften im *niedrigen* Belastungsbereich (2.5% bis 50%) der klinischen Stichprobe waren Item 38 ("...war ich deprimiert und niedergeschlagen"; Fläche

unter der Kurve in diesem Bereich 2.02), Item 7 ("...belastete mich meine Zukunftsaussicht"; 1.92), Item 6 ("...hatte ich Schlafprobleme"; 1.90), Item 29 ("...war ich unabhängig und frei"; 1.90) und Item 35 ("...fühlte ich mich ungenügend und unzureichend"; 1.80). Für die Items im *höheren* Belastungsbereich zeigten sich als besonders hoch von der Messqualität Item 34 ("...dachte ich daran, mir das Leben zu nehmen"; 2.32), Item 23 ("...war ich panisch und voller Angst"; 1.79), Item 10 ("...war ich sehr einsam und alleine"; 1.51), Item 24 ("...fühlte ich mich ohne Wert"; 1.48) und Item 3 ("...fühlte ich mich ohnmächtig"; 1.46). Da es keine Überlappung zwischen beiden Listen gab, wurden für die Kurzversion die besten drei des jeweiligen Belastungsbereiches ausgewählt (Item 6, Item 7, Item 23, Item 10, Item 34 und Item 38; siehe auch Abbildung 3-3).

Auch hier ist wie bereits bei der Screeningfassung angesprochen deutlich zu erkennen, dass die Unterschiede in der Messqualität der Items teilweise sehr gering ausfallen und damit eine Absicherung gegen Zufallsfehler nötig ist. Diese sechs Zielitems wurden wie in Beispiel 1 gegen die zwei Referenzverteilungen getestet (nun allerdings mit sechs Items). In 499 der 500 Bootstrap Läufe in der Validierungsstichprobe zeigten die Zielitems eine größere Fläche unter der Informationsfunktion als die sechs zufällig ausgewählten. Im Vergleich mit den sechs zufällig ausgewählten Nicht-Zielitems schnitten die Zielitems in allen 500 Läufen besser ab. Die abhängigen t-Tests zum Vergleich der Verteilungen der Flächen unter den Kurven fielen in beiden Fällen hoch signifikant aus (sechs Zielitems vs. sechs zufällige Items: $t = 58.84$; $df = 499$; $p < .001$; sechs Zielitems vs. sechs zufällige Nicht-Zielitems: $t = 80.17$; $df = 499$; $p < .001$). Auch die auf diese Weise selegierten Items erweisen sich als effizienter als andere Kombinationen der Skala.

Die differenzielle Effizienz der Skalen ist in Abbildung 3-5 zu sehen. Dargestellt sind die unterschiedlichen Testinformationsfunktionen in der Validierungsstichprobe. Die vollständige Skala mit 26 Items erreicht hohe Werte, insbesondere ein wenig oberhalb von dem Wert "0" auf der latenten Dimension, der in etwa dem Cut Off (-.011) entspricht. Dies bedeutet, dass die Items besser geeignet sind, die klinische Stichprobe zu messen. Über einen breiten Bereich (etwa -2 bis +3 auf der latenten Dimension) erreicht die ursprüngliche Skala Reliabilitätswerte von .9 und darüber, die als sehr gut zu bezeichnen sind (Fliege et al., 2005). Außerhalb dieses Bereiches reduziert sich die Messgenauigkeit schnell auf ein nicht akzeptables Maß. Alle Subskalen zeigen niedrigere Informationsfunktionen. Dies liegt daran, dass in der IRT genauso wie in der Klassischen Testtheorie die Messgenauigkeit auch eine Funktion der Anzahl der Items ist (siehe auch Kapitel 2). Im Vergleich der beiden Kurzfassungen zeigt sich deutlich der Effekt der Auswahl: Während das Maximum der Screening Fassung zum Cut Off verschoben ist, zeigt die Monitoring Fassung ein Maximum deutlich rechts der Screening Fassung.

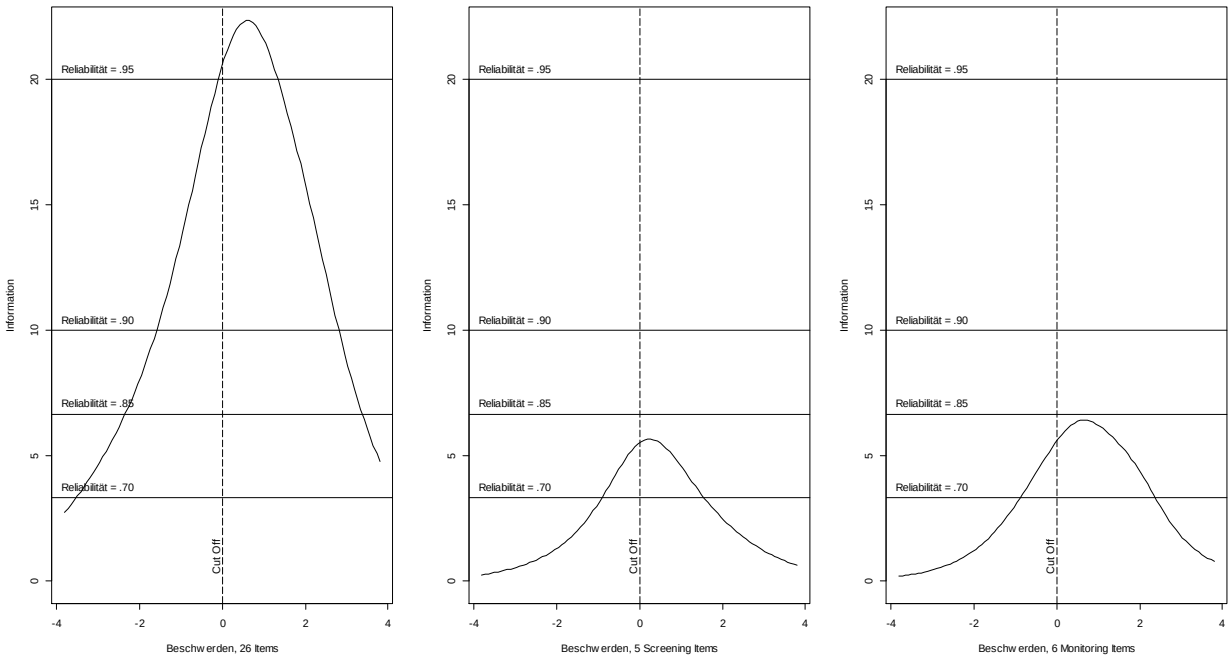

Abbildung 3-5. Vergleich der Testinformationsfunktionen der verschiedenen Testfassungen in der Schätzstichprobe; links die Funktion aller 26 Items; in der Mitte die fünf Items der Screeningfassung; rechts die sechs Items für die Messung von Veränderung in der klinischen Stichprobe; horizontale Linien geben umgerechnete Reliabilitäten als Vergleich (s.a. Babcock & Weiss, 2009); die vertikale Linie in jedem Abbildungsteil gibt den Cut Off zwischen klinischen und nicht-klinischen Fällen.

3.3.6. Anzahl der Optimierungsbereiche, kriteriums- vs. populationsorientierte Itemauswahlen

Die verwendete Anzahl von Optimierungsbereichen und Items stellt einen Kompromiss zwischen verschiedenen Positionen dar. Zunächst sollten Kurzversionen tatsächlich auch substantiell kürzer sein als die Originalfassungen, damit die Zeitersparnis auch deutlich zur Geltung kommt (Lutz, Tholen, et al., 2006). Fünf Items mit fünf Kategorien sind von ihrem Informationsgehalt außerdem 25 dichotomen Items vergleichbar, die durchaus für einen breiten Bereich auf dem latenten Kontinuum in ihrer Reliabilität akzeptable Messungen (s. Kapitel 2). In der aktuellen Debatte zu Kurzfassungen und Screenings in der Versorgung liegen fünf Items außerdem eher am oberen Rand dessen, was als effizient angesehen wird (Cuijpers, Smits, Donker, Ten Have, & De Graaf, 2009; Hart, Werneke, George, & Deutscher, 2012). Das Ergebnis zeigt, dass zumindest für die ausgewählten Bereiche auch akzeptable Reliabilitäten erreicht werden (Abbildung 3-5).

Je nach Ziel der Anwendung kann sich unterscheiden, welche Kriterien für die Itemauswahl sinnvoll sein können. Im zweiten Beispiel wurden sechs (= 2 x 3) Items gewählt, um keinem der beiden Belastungsbereiche ein Übergewicht bei der Skala zu geben. Die Einteilung in eher höheren vs. niedrigeren Belastungsgrad ist eine grobe Einteilung. Eine Möglichkeit wäre daher, das Raster deutlich feiner zu ziehen und die typischen Items für bestimmte Belastungsregionen zu identifizieren. In Abbildung 3-6 wurden die 2.5%, 20%, 40%, 60%, 80% und 97.5% Perzentile genutzt, um das Beschwerdespektrum gemäß der Verteilung der *Patienten*stichprobe in Bereiche zu teilen, in denen jeweils ähnliche Patientenzahlen in Anteilen enthalten sind. Das am Besten messende Item im niedrigsten Quintil ist Item 38 ("...war ich deprimiert und niedergeschlagen"; Fläche: 1.25); das im zweiten Quintil ist Item 24 ("...fühlte ich mich ohne Wert"; 0.77), im dritten Quintil Item 23 ("...war ich panisch und voller Angst"; 0.59), für das vierte Quintil wird Item 10 gewählt ("...war ich einsam und alleine"; 0.21; eigentlich ist es Item 23 (mit .26), das aber schon für das vorherige Quintil verwendet wird und danach Item 34 (mit .22), das im nächsten Quintil verwendet wird); für das fünfte Quintil ist es dann Item 34 ("...dachte ich daran, mir das Leben zu nehmen"; 1.80). Diese Auswahl von Items (10, 23, 24, 34, 38) ist eine populationsbezogene Auswahl. Die Items wurden danach ausgewählt, wo auf dem Kontinuum ähnlich große Gruppen von Patienten liegen und für jede dieser Gruppen wird ein Item gewählt, dass den vorliegenden Beschwerdegrad optimal misst.

Die oben verwendeten Kriterien sind an Stichproben- bzw. Populationskriterien orientiert. Eine solche Definition ist durchaus günstig, da sie es ermöglicht, den verwendeten Items auch eine Bedeutung vor dem Hintergrund der Populationen zu geben (Kamphuis & Noordhof, 2009; Lambert & Ogles, 2009; Lutz, Stulz, et al., 2009; Tingey et al., 1996). Statt einer populationsbezogenen Auswahl kann eine eher kriteriumsbezogene Auswahl vorgenommen werden. Wenn davon ausgegangen wird, dass die durch die Items geschätzten Personenparameter sinnvolle Abschnitte des Beschwerdespektrums operationalisieren, könnte davon ausgegangen werden, dass eine gleichmäßige Einteilung dieses Spektrums in ordinalen Abstufungen der Beschwerdegrade resultieren würde. Hierzu wurden wieder das 2.5%- und das 97.5%-Perzentil als Unter- und Obergrenze verwendet: Dies ist der Bereich der plausiblen Verteilung der *klinischen* Stichprobe. Der Bereich zwischen diesen beiden Punkten wurde aber nun in fünf gleich breite Abschnitte eingeteilt. Ein solches Vorgehen nimmt an, dass die Beschwerdebereiche an sich Informationen über die Belastung der Patienten haben und deshalb jeweils angemessen repräsentiert sein sollten. Im niedrigsten Abschnitt zeigt Item 7 die höchste Fläche unter der Kurve ("...belastete mich meine Zukunftsaussicht"; 0.65), im zweiten Item 38 ("...war ich deprimiert und niedergeschlagen"; 0.86), im dritten Item 23 ("...war ich panisch und voller Angst"; 0.98), im vierten schließlich wieder Item 34 und Item 23 als am Besten messende, die aber für andere Bereiche verwendet werden, daher Item 10 ("...war ich einsam und alleine"; 0.75) und schließlich Item 34 ("...dachte ich daran mir das Leben zu nehmen"; 1.27) für den höchsten Abschnitt.

Vorteil bei diesem Vorgehen ist, dass die Grundlänge auf der latenten Dimension für alle Abschnitte gleich ist. Daher kann auch direkt verglichen werden, für welche Bereiche ein Item besonders repräsentativ ist. Item 34 erreicht bereits im zweithöchsten Abschnitt mit einer Fläche von .91 die Spitzenposition, doch im höchsten Abschnitt erreicht das Item eine Fläche von 1.27 unter der Kurve, womit es als besserer Repräsentant für den höchsten Abschnitt genommen werden sollte (s.a. Abbildung 3-6).

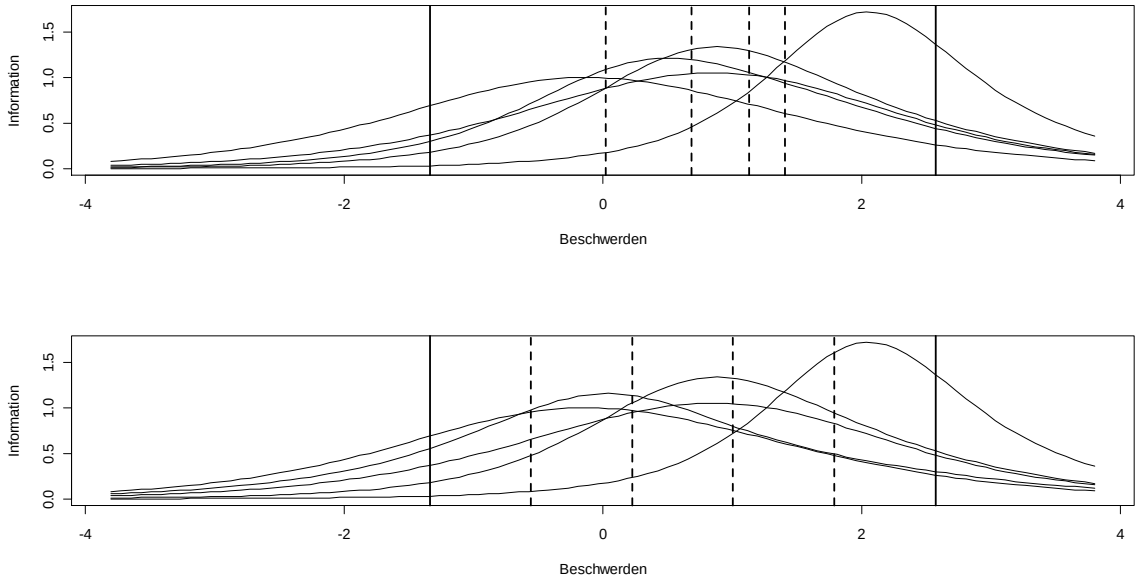

Abbildung 3-6: Iteminformationsfunktionen (oben) für die ausgewählten Items nach der populationsbezogenen Optimierung auf die klinische Stichprobe und (unten) die Auswahl nach gleichmäßigen Abschnitten auf dem Beschwerdespektrum.

Die bisherigen Auswahlen waren alle auf die klinische Stichprobe ausgerichtet. Aus verschiedenen Gründen könnte es aber auch sinnvoll sein, die nicht-klinische Stichprobe als Referenz heranzuziehen (Krause et al., 2011; Krause & Lutz, 2009). Dies könnte dazu dienen, um festzustellen, wie gut die Items auch auf diese Gruppe zugeschnitten sind (Testentwicklung) oder aber, um den Zielbereich therapeutischer Interventionen besonders gut zu messen. Abbildung 3-7 zeigt in ihren Teilen dasselbe Vorgehen noch einmal optimiert für die Stichprobe aus der Normalbevölkerung. Wieder wurde hier zunächst in Quintile eingeteilt und so Abschnitte auf der latenten Variable definiert, die etwa gleichviele Personen enthalten. Im unteren Teil der Abbildung ist die Verteilung der Abschnitte zu sehen, wenn das latente Spektrum zwischen den 2.5% und 97.5% Perzentilen in gleichgroße Abschnitte eingeteilt wird.

Die ausgewählten Items sind entsprechend andere. Für die Quintilsauswahl ist Item 12 ("...war ich unter Anspannung und innerem Druck"; .50) das am besten messende in dem untersten Quintil. Dies gilt auch für das zweite Quintil (dort mit .22), daher wird hier das am zweitbesten messende genommen, Item 7 ("...belastete mich meine Zukunftsaussicht"; .21). Im dritten Quintil misst eben-

falls Item 7 am besten (mit .41); das am zweitbesten messende ist Item 30 ("...war ich voll innerer Ruhe"; .38). Im vierten Quintil misst Item 38 am Besten ("...war ich deprimiert und niedergeschlagen"; .71) und im obersten Quintil Item 24 ("...fühlte ich mich ohne Wert"; .91).

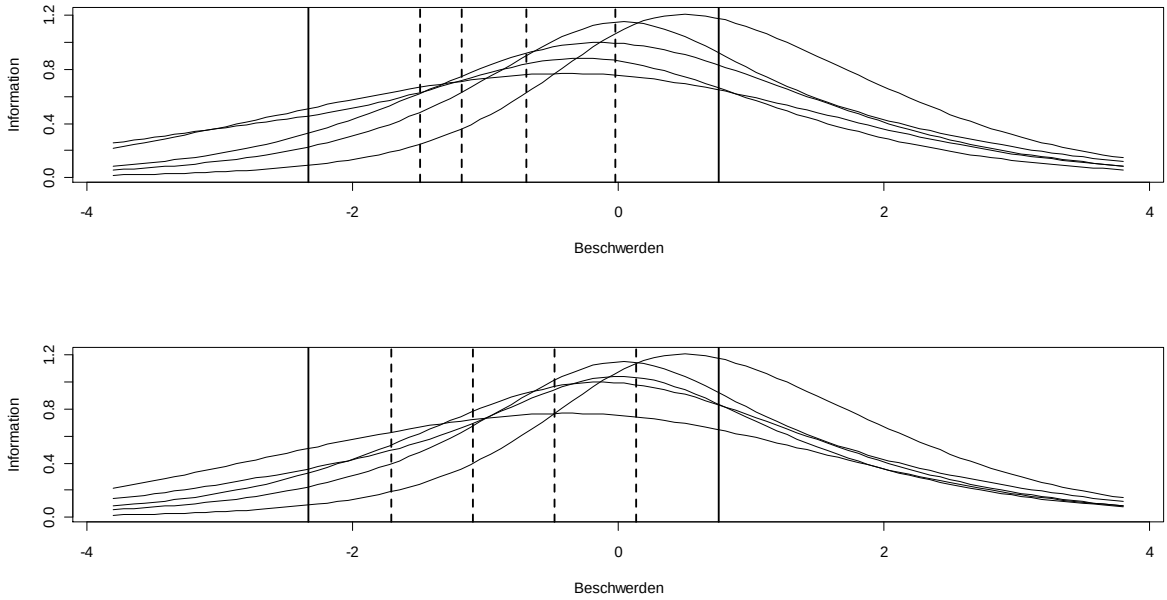

Abbildung 3-7: Iteminformationsfunktionen (oben) für die ausgewählten Items nach der populationsbezogenen Optimierung auf die Normalbevölkerungsstichprobe und (unten) die Auswahl nach gleichmäßigen Abschnitten auf dem Beschwerdespektrum.

Diese Auswahl bestätigt sich im Wesentlichen auch für die kriteriumsbezogene Einteilung des Kontinuums in fünf gleich breite Abschnitte. Für den untersten Abschnitt misst Item 12 ("...war ich unter Anspannung und innerem Druck"; .35) am Besten. Dies gilt auch wieder für den zweiten Abschnitt (dort mit .42), daher wird hier das am zweitbesten messende genommen, Item 7 ("...belastete mich meine Zukunftsaussicht"; .41). Im dritten Abschnitt misst ebenfalls Item 7 am besten (mit .55); das am zweitbesten messende ist hier Item 38 ("...war ich deprimiert und niedergeschlagen"; mit .52). Im vierten Abschnitt misst erneut Item 38 am Besten (hier mit .69), daher wird hier das am zweitbesten messende Item 29 ("...war ich unabhängig und frei"; .63; einziger Unterschied zur Quintilsauswahl) verwendet. Im Abschnitt, der das höchste Beschwerdeniveau der Normalbevölkerung operationalisiert misst Item 24 ("...fühlte ich mich ohne Wert"; .74) am besten.

In Abbildung 3-2 war deutlich zu erkennen, dass die Items des FEP nicht auf die Beschwerden der nicht-belasteten Bevölkerungsstichprobe ausgerichtet sind. Selbst die Items, die am Besten messen, erreichen die Gipfel ihrer Informationsfunktionen erst in einem recht hohen Beschwerdebereich. Wäre eine differenzierte Aussage über die Belastung von Personen in diesen Bereichen nötig, dann müssten neue Items für diesen Bereich des Spektrums entwickelt werden.

3.3.7. Fallbeispiel zur Auswertung im Monitoring

Abbildung 3-8 zeigt ein Fallbeispiel zur Verwendung der Skalen. Abgebildet ist der Verlauf einer 27-jährigen Patientin die sich wegen rezidivierender Majorer Depression in Behandlung befand. Die zwei Verläufe mit 95%-Konfidenzintervallen um die Messzeitpunkte zeigen den gemessenen Verlauf in Personenparametern: erhoben mit allen 26 Items (schwarz) oder den Kurzversionen (fünf Items der Screening Version aus Beispiel 1; sechs Items aus Beispiel 2; grau). Die Verläufe sind ähnlich und liegen innerhalb der Konfidenzintervalle der jeweils anderen Skala.

Die Anwendung in der klinischen Praxis ist recht einfach. Unter den meisten Bedingungen, unter den Therapie durchgeführt wird, ist es möglich, einen Testwert zu erhalten, der aus der Summe der beantworteten Itemkategorien besteht. Bei Geltung des Rasch-Modells ist dies die Information, die zur Bestimmung der latenten Fähigkeitsausprägung nötig ist. Damit ist es möglich, für jede Kurzfassung eine Umrechnungstabelle zu schaffen, in der der erreichte Score in den korrespondierenden Personenparameter (inkl. dessen Standardfehler) umgerechnet werden kann (D. Flora & Thissen, 2002; Thissen & Wainer, 2001). Dies ist vergleichbar mit andern üblichen Testauswertungsmethoden wie der Bestimmung von T-Werten, Stanine oder anderen standardisierten Normen für einen Test. Für jede Kurzform, die ein Test haben soll, müsste im Testmanual letztlich eine zusätzliche Tabelle veröffentlicht werden, die die jeweiligen Umrechnungen der Itemsumme in den Personenparameter umrechnet. Zur Interpretation der Testergebnisse ein Beispiel in Bezug auf den Cut Off zwischen klinischen und nicht-klinischen Fällen: Ein Score von 66 in der 26-Item-Version des FEP liegt gerade unterhalb des Cut Off (θ = -.004; *SE* = .09) und ein Score von 67 darüber (θ = .003; *SE* = .08). Die korrespondierenden Werte auf der Screening-Version sind 13 (θ = -.006; *SE* = .40) und 14 (θ = .15; *SE* = .40).

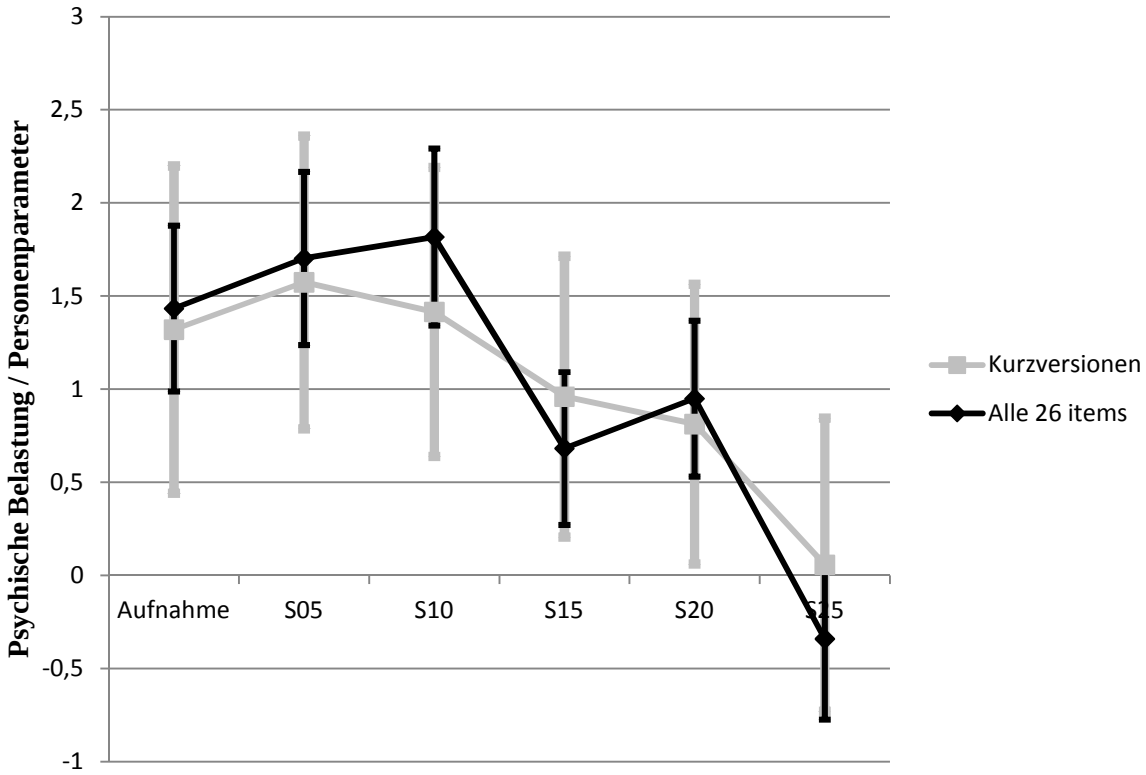

Abbildung 3-8. Beispiel anhand einer Patientin, deren Therapieverlauf über 25 Sitzungen alle 5 Sitzungen dokumentiert wurde; abgebildet sind die Personenparameter geschätzt mit der Information aller 26 Items (schwarz) und der Kurzversionen (Screeningversion bei Aufnahme; alle weiteren mit der Verlaufsversion; grau) inkl. nominaler 95%-Konfidenzintervalle.

Mit einer solchen Umrechnungstabelle ist es möglich, Veränderung in Patienten über Zeit und unterschiedliche Erhebungssituationen zu erfassen und miteinander vergleichbar zu machen (s. Abbildung 3-8). Der Test würde ganz normal als Papier- und Bleistiftversion ausgehändigt, doch anstatt (oder in Ergänzung zu) der Konvertierung der Testscores in eine standardisierte Metrik, könnte dieser Score nun auch in Personenparameter umgerechnet werden und diese über verschiedene Erhebungssituationen miteinander vergleichen. Da zusätzlich der Messfehler für jede Scoregruppe derselbe ist, kann sehr einfach ein Konfidenzintervall um die jeweilige Erhebung konstruiert werden, um zu prüfen, ob sich eine reliable Veränderung zwischen den Messzeitpunkten ergeben hat: Der Postscore muss außerhalb des Konfidenzintervalls des Präscore liegen (Lambert & Ogles, 2009; Reise & Haviland, 2005; Thissen & Wainer, 2001).

Nur Verallgemeinerungen des Rasch-Modells (hier das PCM) bieten diese beiden oben genutzten Eigenschaften. Nur im Rasch-Modell ist der Summenscore eine suffiziente Statistik, die alle

Informationen über die latente Dimension enthält. Und nur aus dieser Eigenschaft entsteht auch die Möglichkeit, für jede Scoregruppe einen eigenen Standardfehler zu bestimmen.

3.4. Diskussion

Das Ziel dieser Arbeit war die Erstellung von zwei Kurzversionen für ein Instrument zur Messung psychologischer Belastung, einmal mit dem Ziel, eine Screening-Version zu erstellen, die im eng umrissenen Belastungsbereich um den Cut Off genau misst; und einmal, die Erstellung einer Version für die Verlaufsmessung in der klinischen Stichprobe. Dazu wurden zunächst Items ausgewählt, die potentiell als eindimensional angesehen werden können und es zeigte sich in diesem Schritt, dass nur zwei Items diese Annahme nicht erfüllten. Diese beiden Items waren auch bereits in einer früheren Studie entsprechend aufgefallen (Schürch et al., 2009) und dementsprechend wurden diese Items in der Folge nicht weiter verwendet.

Aus den verbliebenen Items wurden dann für die beiden Zwecke Items ausgewählt, die in den Zielbereichen die maximalen Informationsfunktionen aufwiesen. Die Verwendung der Bootstrap-Skaleninformationsfunktionen bot die Möglichkeit im Vergleich zu spezifischen Hypothesen zu testen, ob die Items tatsächlich eine Erhöhung der Messeffizienz darstellen. Das Fallbeispiel unterstrich, dass die verschiedenen Fragebogenformen zusammen eingesetzt werden können.

3.4.1. Stärken der Studie und des Vorgehens

Der Kernunterschied zu anderen Arten, Items auszuwählen (z.B. basierend auf Item-Fitstatistiken, Inhaltserwägungen) ist, dass die Ergebnisse aus der Schätzung des IRT-Modells genutzt werden, um eine spezifische Hypothese zu generieren, welche der Items am Besten geeignet sind, einen spezifischen Abschnitt der latenten Dimension zu erfassen. Dadurch wird es auch möglich, spezifische Alternativhypothesen aufzustellen, die die relative Effizienz der Selektion testen können. Die erhöhte Effizienz ist von Bedeutung insbesondere in der Psychotherapieforschung: Bei der Qualitätssicherung und in der Patientenorientierten Versorgungsforschung (Barlow, 2005; Lambert, 2007; Lutz, 2002) werden Fragebögen zum Befinden und zur Erfassung verschiedener relevanter Konstrukte über den gesamten Prozess der Therapie erhoben (Lutz, Böhnke, & Köck,

2011; Lutz, Köck, et al., 2009; Lutz, Mocanu, et al., 2010). Dies stellt insgesamt eine zeitliche Belastung für Therapeuten und Patienten dar, die durch solche Kurzversionen reduziert werden kann.

Über den Vorteil der Zeitersparnis bietet das Rasch-Modell eine andere vorteilhafte Eigenschaft. Erhebungen mit verschiedenen Kurzversionen sowie der Vollversion eines Tests können ohne die Benutzung von Computertechnologie und Softwareprogrammen direkt miteinander verglichen werden. Dies gibt Therapeuten die Möglichkeit, IRT-skalierte Tests auch in einem Papier-und-Bleistift-Setting zu nutzen (D. Flora & Thissen, 2002). Diese Kombination aus IRT und Kurzformen kann also nicht nur dazu beitragen, dass wiederholte und detaillierte Erhebungen im Therapieprozess einfacher werden, sondern so wird auch die Messqualität erhöht. Dies ist nicht eine rein akademische Frage. In den letzten Jahren gab es einen Trend die Ergebnisse aus der Psychotherapieforschung noch näher an den Psychotherapieprozess zu bringen. Die Einrichtung von Forscher-Praktiker-Netzwerken und die Durchführung großer versorgungsnaher Studien (s. Kapitel 1.2) belegen dies. Wenn ein Interesse daran besteht, die Ergebnisse der Forschung sowie auch die Forschung selber in die therapeutische Praxis zu bringen (und dies als ein Beispiel für eine Umgebung nimmt, die arm an den Ressourcen Zeit und Geld ist), dann muss die Forschungspraxis auch an die Gegebenheiten dieser Praxisumgebungen angepasst werden.

Werden als Basis für ein Qualitätsmonitoring oder ein Feedbacksystem beispielsweise Dimensionen wie Wohlbefinden, Symptombelastung und psychosoziale Anpassung verwendet (Ey & Hersen, 2004; Howard et al., 1996; Schulte, 1993) und außerdem noch Prozessvariablen wie die Therapeutische Allianz (Flückiger et al., 2012; Horvath & Luborsky, 1993) und Wirkfaktoren erhoben (Flückiger et al., 2010; Grawe, 2004), bedeutet dies bereits einen erheblichen Aufwand. Der vorgestellte Ansatz versucht dieses Problem zu lösen, indem die Erhebungsinstrumente so kurz gemacht werden, wie es eine möglichst hohe Messqualität erlaubt.

Die Messqualität hängt auch bei einer IRT basierten Auswahl von der Menge der Items ab. Daher muss der Begriff "möglichst hohe Messqualität" immer kontextuell gesehen werden, als ein Kompromiss zwischen dem was möglich und dem was gewünscht ist. Der Vergleich zwischen den Informationsfunktionen, die zur Messung in der klinischen und der nicht-klinischen Stichprobe

ausgewählt wurden, zeigt deutlich, dass für den nicht-klinischen Bereich mehr Items ausgewählt werden müssten, um dieselbe Messqualität zu erreichen. Auch der Vergleich zwischen den Testinformationen in Abbildung 3-5 zeigt, dass zumindest einige Items nötig sind, um eine angemessene Messqualität zu erreichen. Für die praktische Anwendung bedeutet dies, dass vor einer Itemauswahl a) die Zielbereiche, b) eine höchstmögliche Itemzahl und c) die gewünschte Messqualität definiert werden sollten und danach die Items mit dem beschriebenen Vorgehen ausgewählt werden können (für Diskussion dieses Punktes siehe Böhnke & Lutz, submitted; Reckase, 2010). Werden mehr Items benötigt als die als möglich befundene Zahl, dann zeigt dies zunächst, dass die Items für den anvisierten Bereich nicht ausreichend geeignet sind und entweder mehr Items gewählt werden oder neue entwickelt werden müssen.

Eine andere Stärke des gewählten Vorgehens ist, dass alle Schritte in frei verfügbarer Software durchgeführt wurden (Penfield, 2005; R Development Core Team, 2010; siehe auch Studie I). Ein wiederholt genannter Grund für die fehlende Anwendung und Nutzung von IRT-Modellen, sind die hohen Preise und Kosten, die mit der Anschaffung der Programme und der Einarbeitung in die Modelle verbunden sind (Zickar & Broadfoot, 2009). R ist eine Softwareumgebung, die eine ganze Bandbreite an statistischen Analysemethoden bereithält, die alle in Bezug auf Funktionsweise, und Befehlssprache dieselbe Basis haben (Culpepper & Aguinis, 2010). Dadurch ist zusätzliches Training nicht nötig, wenn R an sich bereits gelernt wurde (s. Studie I). Die Anwendung zeigt, dass sowohl die Untersuchung und Skalierung von IRT-Instrumenten wie auch die weitere Anwendung möglich sind.

Das Fallbeispiel und der Exkurs unterstrichen, dass die Kriterien zur Auswahl der Items für den jeweiligen Zweck angepasst werden können, doch auch der Auswahlmechanismus für die Items kann weiter angepasst werden. In der derzeitigen Fassung wurden die Informationsfunktionen der Items genutzt, um festzustellen, welche Items die höchste Messqualität in einer Zielregion haben. Da aber z.B. die Informationsfunktion eines Items nahe der Schwellenparameter höher ist, könnten auch Items für eine Kurzfassung ausgewählt werden, die zumindest eine/einige ihrer Schwellen in der Zielregion haben. Wenn ein Rasch-Modell verwendet wird, ist die Iteminformationsfunktion

immer eingipflig mit dem Maximum an der Lage der Itemschwierigkeit (etwa korrespondierend dem Itemmittelwert; Hambleton et al., 1991), d.h. auch die Lage der Itemschwierigkeit in der Interessenregion könnte als Kriterium genutzt werden. In dieser Arbeit wurde die Informationsfunktion gewählt, da es sich bei ihr um ein quantitatives Maß für die Messgenauigkeit handelt und es so einen guten Vergleich der Leistung der Items ermöglicht.

3.4.2. Kritische Punkte

Ein erster kritischer Punkt ist, wie viel Gewicht psychometrischen Erwägungen zur Konstruktvalidität gegeben wird, die in dieser Arbeit im Mittelpunkt standen. Jede Kurzfassung eines Instrumentes enthält weniger Bedeutungsbereiche als die Originalfassung und könnte daher in ihrer Konstruktvalidität als eingeschränkt gesehen werden (Brod, Tesler, & Christensen, 2009; Hays et al., 2000). Diese Erwägungen werden schon beim ersten Schritt des Vorgehens wichtig: Dem Ausschluss von Items, die nicht als hinreichend eindimensional gesehen werden. Im konkreten Fall sind es nur zwei Items und es bleiben Items für alle Konstruktinhalte erhalten.

Bei der Reduktion auf fünf bzw. sechs Items gehen allerdings zwangsläufig Inhalte verloren. Dem könnte entgegengehalten werden, dass die erfolgreiche Anpassung eines IRT-Modells zeigt, dass alle Items dieselbe Dimension messen und damit jedes beliebige Item als gleichwertiger Repräsentant für die latente Dimension gesehen werden kann und sie sich nur durch ihre relative Lage auf der latenten Dimension von einander unterscheiden. Diesem Argument folgend, wären qualitative Informationen zu den Inhalten der Items nur insofern nötig, als dass *vor* der IRT-Skalierung hinreichend empirisch belegt worden sein sollte, dass die Items inhaltsvalide sind. Dennoch ist gerade von klinischer Seite nicht von der Hand zu weisen, dass es wichtig zu wissen ist, welche Items/Symptome/Beschwerden erhoben werden. Die beschriebenen Itemauswahlen tendieren alle zum depressiven Spektrum, mit keinem expliziten Angstitem in der Screeningauswahl, die den Übergang von typischen zu klinischen Belastungsgraden markieren ("...war ich deprimiert und niedergeschlagen"; "...fühlte ich mich ungenügend und unzureichend"; "...fühlte ich mich ohne Wert"; "...war ich unabhängig und frei"; "...hatte ich Schlafprobleme") und immerhin einem Angstitem in der Monitoring-Fassung ("...war ich panisch und voller Angst"). Die Items 7 ("...belastete

mich meine Zukunftsaussicht") und 38 ("...war ich deprimiert und niedergeschlagen") geben Auskunft über eher niedrige Belastungsgrade in der klinischen Stichprobe, während die Items 23 ("...war ich panisch und voller Angst") und 34 ("...dachte ich daran, mir das Leben zu nehmen") deutlich erhöhte Belastungsgrade anzeigen. Der Vergleich zur nicht-klinischen Stichprobe zeigt weiterhin, dass weder Item 34 (Suizidgedanken) noch Item 23 (panisch und voller Angst) etwas über die typischen Belastungsgrade in der nicht-klinischen Gruppe Aussagen. Hohe Belastungsgrade in der Bevölkerung werden durch Item 24 ("...fühlte ich mich ohne Wert") angezeigt. In der Stichprobe der nicht-klinischen Fälle ist zusätzlich Item 12 ("...war ich unter Anspannung und innerem Druck") ein Indikator für sehr niedrige Belastungen, der in der klinischen Stichprobe nicht zur Differenzierung zwischen den Personen geeignet ist.

Items, die für die Facetten eines Konstrukts als klinisch relevant angesehen werden, können immer Teil eines umfassenden Erhebungsplanes sein. Die Frage z.B. nach den Suizidgedanken kann immer in ein Monitoringsystem aufgenommen werden (zur Debatte, ob so überhaupt sinnvoll erfassbar: Fowler, 2012). Auch die Angstitems des FEP, die eher Indikatoren für niedrige Belastungsgrade sind, könnten trotzdem verwendet werden. Ist für eine spezifische Frage, die aufgenommen werden soll, empirisch nachgewiesen worden, dass sie zu der zu messenden Dimension gehört, dann kann die Kurzversion um ein theoriegeleitetes Item erweitert werden und es ändert sich sonst nichts an dem beschriebenen Vorgehen. Die Erhebung wird lediglich vor dem Hintergrund Anzahl Items vs. Messqualität weniger effizient.

Die Frage der Konstruktinhalte sollte nicht mit der Frage nach Mehrdimensionalität verwechselt werden. Bei der Planung eines Erhebungssystems muss festgestellt werden, wie viele Dimensionen es erfassen soll. Wenn mehrere getrennte Dimensionen in Frage stehen, dann sollten diese auch erhoben werden, ggf. jede mit einer Kurzversion, aber jede möglichst eindimensional (G. T. Smith et al., 2009). Damit dreht sich die Frage also weniger darum, welche Iteminhalte in der Kurzversion auftreten sollten, sondern eher um die Frage, was die relevanten Dimensionen sind (siehe Kapitel 1). Wird bei der Dimensionalitätsprüfung ein Item ausgeschlossen, das als klinische relevant aber vor dem Hintergrund der Konstruktvalidität als zu einer anderen Dimension gehörig

eingestuft wird, so bedeutet dies nicht, dass dieses Item nicht verwendet werden kann. Ist der darin abgefragte Inhalt für den Erhebungsplan relevant, sollte dieser Inhalt mit einer weiteren, aber an sich eindimensionalen Skala, erhoben werden und auf diese Weise in das Monitoring-System integriert werden.

Eine letzte Frage, die kritisch diskutiert werden muss, ist die Frage, wie viele Items zur Messung ausgewählt werden -- bzw. nötig sind. Die umfangreichste Studie zu diesem Thema bisher stellt ein Konferenzbeitrag dar (Babcock & Weiss, 2009), in dem die Autoren zu der abschließenden Beurteilung kommen, dass weniger als 15 dichotome Items nicht genügend sein könnten, mehr als 50 jedoch die Messgenauigkeit nicht weiter erhöhen würden. Emons und Kollegen (Emons et al., 2007) versuchten festzustellen, wie viele Items nötig wären, um einen spezifischen Punkt auf dem latenten Kontinuum zu messen und zu stabilen Klassifikationsurteilen um ihn (wie hier bei dem Cut Off-Beispiel) zu kommen. Ihr Ergebnis war, dass Skalen mit weniger als zwölf Items (und sicher mit sechs oder weniger Items) keine hinreichende Konsistenz der Klassifikation mehr ermöglichen würden. In einer neuen Studie mit einer ähnlichen Fragestellung kommen sie zur Feststellung, dass hochqualitative Tests 20 oder sogar bis zu 40 Items benötigen könnten. Dies würde bezogen auf das Beispiel bedeuten, dass mehr Items nötig wären, als ursprünglich in der Skala vorhanden waren. Inwiefern dies auch auf Anwendungen zutrifft, bei denen durch eine Kurzfassung nicht der gesamte Bereich des latenten Kontinuums, sondern nur spezifische Bereiche gemessen werden sollen, muss derzeit offen bleiben. Mehr Forschung hierzu wird nötig sein, um eine theoretische und empirische Basis für die Verwendung, Erstellung und Länge von Kurzversionen bereitzustellen, insbesondere, da in der letzten Zeit vorgestellte Kurzfassungen von Instrumenten immer weniger Items enthalten und Fassungen mit drei oder weniger Items (pro Fragebogenfacette) diskutiert werden (Cuijpers et al., 2009; DeSalvo et al., 2006; Fang et al., 2011; Hart et al., 2012; Meijer et al., 2011; Yamazaki, Fukuhara, & Green, 2005).

3.4.3. Fazit

Neben diesen offenen Fragen stellt der vorgestellte Ansatz aber Verbindungen zwischen unterschiedlichen Verwendungsmöglichkeiten und Aspekten der IRT-Modelle her. Diese wurden aber

bislang nicht breit in der Forschung und Praxis angewendet. Die Verwendung von IRT-Kalibrierungen um Items von Tests miteinander zu verbinden ist in der Literatur zum "Test-Equating" bekannt (Holman, Lindeboom, et al., 2003). Der vorgeschlagene Konstruktionsweg ist auch nahe verbunden den Itembanking-Ansätzen, da ein Fragebogen nichts weiter ist, als eine sehr kleine Itembank (in diesem Fall) für die Messung von symptomatischer Belastung (Chang & Reeve, 2005). Die Auswahl von Subgruppen von Items ist nah verwandt mit der sog. "Testlet"-IRT, doch IRT-basierte Testlets werden in der Regel nur dazu verwendet, die Messpräzision an bestimmten Stadien im computer-adaptiven Testprozess zu erhöhen (Murphy, Dodd, & Vaughn, 2010). Daher informiert weitere Forschung in diesen Teilgebieten der Psychometrie nicht nur das jeweilige Gebiet alleine: Dieses Manuskript zeigt auf, wie diese Prinzipien in einem angewandten Forschungsfeld implementiert und zur Reduktion der Wissenschaftler-Praktiker-Kluft genutzt werden könnten.

4. Studie III: Vergleich dreier Methoden zur Entwicklung veränderungssensitiver Kurzformen von Beschwerdemaßen[32]

4.1. Einleitung

In der klinischen Routineversorgung wie auch bei epidemiologischen Untersuchungen steht in der Regel die Anforderung einer möglichst ökonomischen Testdurchführung im Vordergrund. Idealerweise sollten die verwendeten Tests schnell und einfach zu erheben sein. Das zu messende Merkmal sollte mit befriedigender Genauigkeit erfasst werden. Der Bedarf für Messinstrumente für wiederholte Messungen in der klinischen Routineversorgung erklärt sich unter anderem durch die gestiegene Kooperation zwischen praktizierenden Therapeuten und Wissenschaftlern in sogenannten Wissenschaftler-Praktiker-Netzwerken (s. Kapitel 1.2.3). Die Nutzung dieser Daten ist wie in der Einleitung festgehalten breit: Bei Qualitätssicherung und Monitoring (Lutz, 1997, 2002), bei der Verwendung von Fragebogen-basierten Feedbacksystemen in der Psychotherapie (Lambert, 2007) und bei adaptiven Vorhersagemodellen (Howard et al., 1996; Lutz et al., 1999; Lutz, Saunders, et al., 2006). Auch in der klinischen Forschung sind wiederholt erhobene Daten von Bedeutung. Bei der Frage nach dem Einfluss von Therapeuten- und Patienteneffekten hat sich die Analyse von Verlaufsdaten mittels Mehrebenenmodellen etabliert (Crits-Christoph et al., 1991; Lutz et al., 1999). Bei der Überprüfung der Phänomenologie und Bedeutung plötzlicher Gewinne und Verluste im Therapieprozess (Lutz et al., in press; Tang & DeRubeis, 1999a; Tschitsaz-Stucki & Lutz, 2009) und der Analyse von Verlaufsmustern im Therapieprozess sind Forschende auf wiederholt erhobene Daten angewiesen (Lutz, Stulz, & Köck, 2009; Stulz et al., 2007).

Der Entwicklung von Fragebogenkurzformen kommt damit eine besondere Bedeutung zu, da sie sowohl Forschung wie auch klinischen Alltag erleichtern. In diesen Kontexten bedeutet eine wiederholte Erhebung in der Regel, dass pro Sitzung (im stationären Setting vielleicht jeden Tag; Newnham & Page, 2010) ein oder mehrere Fragebögen ausgefüllt werden. Forschende wie Prakti-

[32] Dieses Kapitel wurde überarbeitet publiziert (Böhnke & Lutz, 2010c). Ich danke dem Team der Poliklinischen Psychotherapieambulanz in Osnabrück (unter Leitung von Prof. Dr. Henning Schöttke) für die zur Verfügungstellung der Daten sowie Fabian Jung für die Erhebung der Bevölkerungsstichprobe im Rahmen seiner Diplomarbeit.

ker stehen vor dem Problem, dass viele der gängigen Instrumente einen zu großen Zeitaufwand zur Bearbeitung benötigen, der sich schlecht in die Abläufe integrieren lässt und eventuell die Patienten zusätzlich belastet. Generell ist zu empfehlen, dass vollständige Fragebogenversionen zu Aufnahme und Entlassung verwendet werden, bei den Zwischenerhebungen aber auf Kurzformen dieser Instrumente zurückgegriffen wird (Lutz, Tholen, et al., 2006).

Instrumente zur Messung psychischer Belastung operationalisieren meist die empfundene Belastung durch verschiedene Symptome der psychischen Erkrankung. Die Items bzw. Symptome zeigen in der Regel unterschiedliche Veränderungsraten im Verlauf der Therapie und weisen damit auf ein unterschiedliches Ansprechen der Symptome auf die therapeutische Intervention hin. Das Phasenmodell der Veränderung (Howard et al., 1993; Stulz & Lutz, 2007) geht davon aus, dass sich psychische Beschwerden in drei Bereiche gruppieren lassen: Wohlbefinden, Symptome und psychosoziales Funktionieren. Das Modell postuliert, dass voneinander abhängige Veränderungen in den drei Bereichen auftreten. Zunächst sollte eine Besserung im Bereich Wohlbefinden eintreten, dann im Bereich Symptome und schließlich im Bereich psychosozialen Funktionierens (s.a. Kapitel 1.4.3). Entsprechend wären Items, die Wohlbefinden messen, besonders am Anfang der Therapie sensitiv, Items für Symptome eher in mittleren Bereichen und Items für soziales Funktionieren erst bei langfristigen Veränderungsprozessen.

Auch innerhalb dieser drei Bereiche ist ein unterschiedliches Ansprechen auf therapeutische Interventionen belegt. Kopta und Kollegen (1994) untersuchten die unterschiedliche Sensitivität von Symptomen für Veränderung durch Psychotherapie. An Patienten aus der ambulanten Versorgung zeigten sie, dass sich die Items der Symptom Checklist (SCL-90-R) basierend auf unterschiedlichen Graden der Veränderung in drei Gruppen gliedern ließen. Sie fanden eine Itemgruppe, die sie als "akute Veränderungen" bezeichneten, bei denen Remissionsraten von >= 24% nach dem ersten Kontakt und 50% bis zur zehnten Sitzung festgestellt wurden, wobei Remission immer im Sinne klinisch signifikanter Veränderung definiert war (Jacobson & Truax, 1991). Daneben gab es eher "chronische Symptome" mit moderaten Remissionsraten (22% oder weniger nach Erstkontakt; 50% nach 7-27 Sitzungen). Schließlich gab es "charakterologische", die sich durch sehr niedrige Remis-

sionsraten auszeichneten (50% gebesserte Patienten erst nach mehr als 18 Sitzungen; auch nach 52 Sitzungen erst 59% gebesserte). Diese Itemgruppen entsprachen nicht vorigen Ergebnissen durch Faktorenanalysen (Cyr, McKenna-Foley, & Peacock, 1985; Derogatis, 1977), was verdeutlicht, dass die Eigenschaft der Sensitivität für Veränderungen eine eigene Ebene der Konstruktvalidierung ist.

Wie können für die Verlaufsmessung geeignete Items ausgewählt werden, die einerseits eine Vergleichbarkeit mit anderen Messinstrumenten erlauben und andererseits allgemeine Testgütekriterien erfüllen? Aus verschiedenen Publikationen lassen sich drei zentrale Kriterien zusammenfassen (Burlingame et al., 2006; Tyron, 1991; Vermeersch et al., 2000). Die beobachtete Veränderung der Itemmittelwerte sollte in der nach einer Intervention postulierten Richtung feststellbar sein (Rückgang der Belastung). Zusätzlich sollte die gemessene Veränderung bei behandelten Personen größer sein als bei nicht behandelten Personen. Dies ist eine Möglichkeit, eine gewisse Konstruktvalidität herzustellen, da es schwierig ist, Außenkriterien für die gelungene Messung des Veränderungsprozesses zu definieren. Letztlich sollte auch eine multidimensionale Erfassung des Therapieergebnisses ermöglicht werden, damit die gesamte Bandbreite der Dimension "psychische Belastung" erfasst werden kann (Burlingame et al., 2006; Lutz, Tholen, et al., 2006).

Meier (1997) stellte einen breiteren Kriterienkatalog für die Auswahl von Items vor, der weitere Aspekte anspricht (alle Kriterien sind in Tabelle 4-1 zusammengestellt). Zwei rationale Kriterien werden ergänzt, die vor der empirischen Untersuchung und bei der Interpretation der Ergebnisse zur Anwendung kommen. Die Itemauswahl sollte bezogen auf die Interventionstheorie relevante Items beinhalten und es sollte mehr als ein Item verwendet werden, um den Einfluss von Zufallsfehlern zu reduzieren. Die bereits angesprochenen Kriterien der Mittelwertsveränderung und des gezielten Vergleiches mit unbehandelten Gruppen präzisiert Meier (1997) noch in der Hinsicht, dass die Items sich zwischen Gruppen Erkrankter vor einer Intervention nicht unterscheiden sollten. Er schlägt zwei zusätzliche Kriterien für die Untersuchung der Items vor. Es sollten möglichst keine Decken- oder Bodeneffekte auftreten, da dies die möglichen Veränderungen reduziert, die empirisch festgestellt werden können. Außerdem sollten Items nicht berücksichtigt werden, die

sich als anfällig für systematische Verzerrungen oder bestimmte Antwortstile zeigen. Abschließend hält er fest, dass eine solche Itemauswahl auch validiert werden müsse.

Tabelle 4-1: Zusammenstellung von Kriterien zur Identifikation von besonders veränderungssensitiven Items bzw. zur Konstruktion veränderungssensitiver Kurzformen etablierter Instrumente.

Inhaltlich begründete Auswahl der Items*
Mehr als ein Item pro Konstrukt erheben*
Keine Decken- oder Bodeneffekte*
Itemmittelwerte sollten sich in hypothesenkonformer Weise im Verlauf der Intervention ändern (Abnahme der Belastung)*‡ Vergleich mit Bezugsgruppen zur Kriteriumsvalidität der Veränderungserfassung*‡
Bei verschiedenen Trials: die Items sollten vor Durchführung der Interventionen nicht zwischen den Gruppen unterscheiden*
Analyse auf systematische Verzerrungen und Antwortstile*
Möglichst multidimensionale Erfassung‡
Validierung der gefundenen Itemauswahl*

* Kriterien formuliert bei Meier, 1997
‡ Kriterien formuliert bei anderen Autoren; z.B. Burlingame et al., 2006; Lutz, Tholen, et al., 2006; Tyron, 1991; Vermeersch et al., 2000

Empirische Untersuchungen benutzten meist nur einen Teil dieser Kriterienliste, doch zeigen sie, dass sie es ermöglichen, Items gezielt auszuwählen. Vermeersch und Kollegen (2000) untersuchten die Veränderungssensitivität der Items des "Outcome Questionnaire" (OQ; Lambert et al., 1996). Sie verglichen die Veränderungen im OQ bei 1176 behandelten Patienten mit denen bei 284 Unbehandelten mittels eines Mehrebenenmodells. Die meisten der Items zeigten eine statistisch bedeutsame Veränderung in der erwarteten Richtung und veränderten sich zudem stärker in der Gruppe der Behandelten als in der Gruppe der Nicht-Behandelten. Die Autoren verwendeten Effektstärken, um die im Mittel am stärksten veränderten Items zu identifizieren. Burlingame und Kollegen (2006) untersuchten die unterschiedliche Variabilität der Symptome, die durch die Items der Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962) operationalisiert werden. An den erhobenen BPRS-Protokollen von 223 stationären Patienten wurden ebenfalls mittels eines Mehrebenenmodells Items identifiziert, die sich in Richtung Symptomverbesserung veränderten. Außerdem sollten die Items verglichen mit einer Kontrollgruppe (Wiederaufnahmewerte von 63

Patienten) höhere Veränderungen zeigen. Sie fanden 22 Symptome, die das erste Kriterium erfüllten und ebenso viele hielten dem Vergleich mit der Kontrollgruppe stand.

Insgesamt macht dies deutlich, dass für die Auswahl von Items zur Messung von Veränderungen Kriterien verwendet werden müssen, die über die Auswertung von Itemkennwerten an einem Zeitpunkt hinausgehen. In der Literatur diskutierte Ansätze bieten eine ganze Reihe von Kriterien an und zeigen, dass es möglich ist, Items für die Veränderungsmessung zu identifizieren. Allerdings stoßen sie auf Probleme, wenn es beispielsweise um die statistische Auswahl der Items geht. Meier (1997) schlägt Signifikanztests zur Identifikation von deutlich veränderten Prä- und Postwerten bei einzelnen Items vor, was zu einem Problem der α-Fehler-Kumulierung führt. Die Notwendigkeit von nicht-behandelten Kontrollgruppen, idealerweise für dieselbe Dauer wie die behandelten Patienten, ist aus strukturellen wie auch ethischen Gründen nicht zu jeder Gelegenheit umsetzbar (Krause & Lutz, 2009).

In dieser Arbeit sollen zwei Ansätze zur Untersuchung der Veränderungssensitivität vorgestellt werden. Anhand von Prä- und Postmessungen aus zwei universitären Forschungsambulanzen und Bevölkerungsdaten aus einer Quota-Stichprobe werden zunächst deskriptive Methoden zur Bestimmung veränderungssensitver Items präsentiert, bei denen die *beobachteten* Mittelwerte dieser Gruppen zur Schätzung des Veränderungspotenzials herangezogen werden. Danach werden mittels der Latent Profile Analysis die *latenten* Mittelwerte zur Abschätzung des Veränderungspotenzials ebenfalls deskriptiv verglichen. Untersucht werden die Items der Skala "Beschwerden" des "Fragebogens zur Evaluation von Psychotherapieverläufen" (Lutz, Schürch, et al., 2009).

4.2. Methoden

4.2.1. Stichproben

Insgesamt flossen die Daten von $N = 327$ Personen in die Analysen ein. Die Poliklinische Psychotherapieambulanz der Universität Trier steuerte $N = 78$ Ersterhebungen bei, die Poliklinische Psychotherapieambulanz der Universität Osnabrück $N = 129$ Erst- und Abschlusserhebungen und

aus einer Bevölkerungsstichprobe wurden $N = 120$ Personen aufgenommen. Diese Stichproben werden nun im Folgenden genauer beschrieben.

Ambulante Stichprobe

Die verwendeten Daten von $N = 207$ Patienten stammen aus den Standarderhebungen der Therapien in den Poliklinischen Psychotherapieambulanzen an den Universitäten Osnabrück und Trier. Die Ambulanz an der Universität Trier steuerte die Daten von $N = 78$ Ersterhebungen bei. Der Frauenanteil hier betrug 65.3 %. Das Durchschnittsalter lag bei 38.5 Jahren ($SD = 12.5$). Die meisten Patienten suchten die Therapie aufgrund einer affektiven Störung auf (28.2 %), gefolgt von denen aufgrund einer Angststörung 17.9 % und 10.3 % aufgrund einer Reaktion auf eine schwere Belastung oder einer Anpassungsstörung 10.3 %. Die übrigen Patienten litten nach ICD-10 an diversen anderen Störungen und 41.0 % wiesen mindestens eine weitere Diagnose auf.

Aus der Ambulanz der Universität Osnabrück stammen die Daten von $N = 129$ abgeschlossenen Therapien. Der Frauenanteil in dieser Stichprobe betrug 63.8 %. Das Durchschnittsalter lag bei 41.7 Jahren ($SD = 12.2$). Von den Patienten suchten 37.2 % die Behandlung aufgrund einer affektiven Störung auf, 17.8 % aufgrund einer Angststörung und 14.7 % aufgrund einer schweren Belastung oder Anpassungsstörung. Die restlichen Patienten litten nach ICD-10 an diversen anderen Störungen. 60.4 % wiesen mindestens eine weitere Diagnose auf. Die Therapien dauerten zwischen 13 und 103 Behandlungsstunden ($M = 54.0$; $SD = 20.3$).

Nicht-klinische Stichprobe

Die nicht-klinische Stichprobe ($N = 120$) ist eine Quota-Stichprobe, die im Westen der Bundesrepublik Deutschland erhoben wurde (Jung, 2008). Die zu erhebenden Quoten wurden aufgrund des scientific-use Files des Mikrozensus[33] bestimmt (Lüttinger & Riede, 1997). Neben Geschlecht wurde nach Alter in drei Gruppen (18 bis 30 Jahre, 31 bis 50 Jahre und 51 bis 65 Jahre) und dem höchsten allgemeinen Schulabschluss (Haupt-/ Volksschule, Realschule/ Polytechnische Oberschule, Fachhochschulreife und Abitur) erhoben. Diese Stichprobe hatte einen Frauenanteil von 50 %

[33] s. http://www.forschungsdatenzentrum.de/campus-file.asp (heruntergeladen am 31.08.2009)

und 95 % der Befragten waren Deutsche. Das mittlere Alter betrug 43.9 Jahre (*SD* = 13.6). Der Familienstand der meisten war entweder ledig (37.5 %) oder verheiratet (38.3 %). 71.6 % der Befragten lebten mit einem festen Partner zusammen (verheiratet wie unverheiratet; 16.7 % gaben an, längerfristig keinen Partner zu haben). Von den Befragten gaben 40 % an, einen Hauptschulabschluss zu haben, 31.7 % die mittlere Reife und 28.3 % das Abitur. Derzeit arbeiten 53.3 % der Befragten Vollzeit, 33.3 % Teilzeit und 9.2 % gaben an, arbeitslos zu sein.

4.2.2. Der verwendete Fragebogen

Der "Fragebogen zur Evaluation von Therapieverläufen" (FEP; Lutz & Böhnke, 2008) umfasst 40 Items, die auf einer fünfstufigen Skala beantwortet werden (Kategorien nie, selten, manchmal, oft und sehr oft). Aus den Items lassen sich die vier untereinander korrelierten Subskalen Wohlbefinden, Beschwerden, interpersonale Beziehung und Inkongruenz bilden. Ähnlich wie bei vergleichbaren Instrumenten (z. B. der Outcome Questionnaire, Lambert et al., 1996) bildet der Gesamtmittelwert das Ausmaß psychischer Beeinträchtigung ab. Eine eingehende Prüfung der klassischen und probabilistischen Gütekriterien ergab gute bis sehr gute Werte für die (Retest-) Reliabilität so wie konvergente und diskriminante Validität (Lutz & Böhnke, 2008; Lutz, Schürch, et al., 2009; Schürch et al., 2009). Der Gesamtwert des FEP erreicht in der vorliegenden Stichprobe eine Reliabilität von Cronbach-α = 0.94. Zur Bestimmung wurden alle Präwerte der ambulanten Stichproben und die Bevölkerungsstichprobe herangezogen (alle vierzig Items waren von N = 285 Personen bearbeitet; das entsprach 87.2 % der Stichprobe). Die Subskala "Beschwerden", die im Folgenden verwendet wird (Itemformulierungen Tabelle 4-2), erreicht ein Cronbach-α = 0.92 (N = 318; 97.2 % der Stichprobe).

Tabelle 4-2: Die elf Items der Skala "Beschwerden" des "Fragebogens zur Evaluation von Psychotherapieverläufen".

	Formulierung
	In der letzten Woche...
Item 4	war ich nervös
Item 6	hatte ich Schlafprobleme
Item 7	belastete mich meine Zukunftsaussicht
Item 8	war ich ängstlich
Item 10	war ich sehr einsam und alleine
Item 12	war ich unter Anspannung und innerem Druck
Item 23	war ich panisch und voller Angst
Item 24	fühlte ich mich ohne Wert
Item 31	konnte ich mich für nichts begeistern
Item 34	dachte ich daran, mir das Leben zu nehmen
Item 38	war ich deprimiert und niedergeschlagen

4.2.3. Itemuntersuchungen

Zunächst werden die beobachteten Verteilungseigenschaften der Items der Skala untersucht. Die Mittelwerte und Standardabweichungen wurden für die drei Stichprobengruppen (ambulante Präwerte, ambulante Postwerte und Werte der Bevölkerungsstichprobe) getrennt bestimmt. Mittels der Konfidenzintervalle wurde untersucht, welche Items die Kriterien für die Veränderungssensitivität erfüllen. Dieses Vorgehen kann als Standardvorgehen betrachtet werden und liefert Vergleichswerte für die folgenden Untersuchungen. Da in Praxiskontexten oftmals kleine Stichproben anfallen und die Verteilungsannahmen nicht erfüllt sind, wird die Untersuchung der beobachteten Verteilungen der Items abschließend durch Resampling-Methoden ergänzt.

Im zweiten Schritt werden nicht die beobachteten Verteilungen verwendet, sondern es werden mittels der *Latent Profile Analysis* die latenten Verteilungen der Items modelliert. Das Vorgehen ist einerseits darin begründet, dass bei der Betrachtung der Mittelwerte angenommen wird, dass eine Person mit einer klinischen Diagnose vor einer therapeutischen Intervention einer anderen Population angehört als nach der Intervention (z.B. Jacobson & Truax, 1991). Das bedeutet, dass zunächst für ein Item nachgewiesen werden muss, dass es in der Lage ist im Mittel eine Veränderung anzu-

zeigen. Andererseits liegt ein Vorteil darin, dass so größere Stichprobengruppen für die Latent Profile Analysis erreicht werden und diese somit stabilere Schätzer liefern (ist im Rahmen dieser Modelle ein übliches Vorgehen, z.B. Glück & Spiel, 1997, 2007). Ein weiterer Vorteil dieser Analyse liegt darin, dass die latenten Beschwerdegrade modelliert werden, d. h. auch berücksichtigt wird, dass es in der Bevölkerungsstichprobe schwerer belastete Personen gibt. Nachfolgend werden zunächst die Kriterien für die Itemidentifikation vorgestellt und dann das Vorgehen in diesen beiden Analyseschritten erläutert.

Kriterien für die Auswahl der Items

Ziel ist es eine inhaltlich sinnvolle Skala zur Messung der Beschwerden der Patienten zu entwickeln (erste zwei Bedingungen Tabelle 4-1). Eine deskriptive Analyse der empirischen Kriterien von Meier (1997) schließt sich daran an, wobei besonderer Wert auf die hypothesenkonforme Veränderung im Laufe der Intervention gelegt wird: Die Itemmittelwerte der Präwerte sollten höher sein als die Itemmittelwerte der Postwerte. Items, die in diesem Sinne geeignet sind, sollten folgende Bedingung erfüllen:

> *Bedingung 1: Die Konfidenzintervalle der Itemmittelwerte für ambulante Prä- und Postwerte sollten sich nicht überschneiden.*

Da die hier zur Verfügung stehende Referenzgruppe Messungen aus der Bevölkerung sind, sollte nur Veränderung auf klinisch bedeutsamen Symptomen erreicht werden und die ambulanten Postwerte sollten sich im Mittel denen der Bevölkerung annähern:

> *Bedingung 2: Die Konfidenzintervalle der Itemmittelwerte für ambulante Präwerte und die der Bevölkerung sollten sich NICHT überschneiden.*

> *Bedingung 3: Die Konfidenzintervalle der Itemmittelwerte für ambulante Postwerte und die der Bevölkerung sollten sich überschneiden.*

Die letzte Bedingung geht davon aus, dass eine Linderung der Beschwerden auf ein klinisch nicht auffälliges Maß das zentrale Ziel der Therapie ist. Dies kann aus verschiedenen Gründen

unrealistisch sein (Lambert & Ogles, 2009; Tingey et al., 1996). Von Veränderung wird in diesem Fall also dann gesprochen, wenn Bedingungen eins und zwei erfüllt sind.

4.2.4. Analysevorgehen

Die Kriterien werden zunächst direkt an den deskriptiven Verteilungswerten untersucht. Dies kann als Standardvorgehen betrachtet werden. Eine Erweiterung dieses Vorgehens stellt die Nutzung von Resamplingmethoden bzw. des "Bootstraps" dar (Efron & Tibshirani, 1993). Gerade dann, wenn eher kleine Stichproben vorliegen oder stark verzerrte Verteilungen erwartet werden, kann es sinnvoll sein, statt auf der Normalverteilungsannahme aufbauenden Konfidenzintervalle die empirischen Verteilungen wiederholter Stichprobenziehungen zu betrachten. Als Beispiel wurden in den drei Gruppen jeweils 1000 Stichproben von 40 gezogen und die Mittelwerte der Items bestimmt (Analysen durchgeführt mit R: R Development Core Team, 2009, Version 2.9.2). Die 2.5%- und 97.5%-Quantile dieser 1000 Stichprobenmittelwerte geben dann ein Konfidenzintervall ohne Verteilungsannahme an.

Im zweiten Schritt wird der Datensatz mittels *Latent Profile Analyse* (Gibson, 1959; Vermunt & Magidson, 2002) untersucht. Dazu werden die Daten der drei Gruppen so zusammengestellt, dass eine Matrix entsteht, die die Items als Spalten enthält und alle vorhandenen Messungen auf diesen Items als Reihen, d. h. die Patienten der Ambulanz in Osnabrück gehen durch Prä- und Postwerte zweifach in die Analyse ein und die Bevölkerungsstichprobe und die Präwerte der Ambulanz in Trier jeweils einfach (s. z. B. Glück & Spiel, 1997, 2007). Die Latent Profile Analysis ist ein Verfahren, das Klassen unterschiedlicher Mittelwerts- und Varianzmuster intervallskalierter Variablen schätzt. Ergebnis einer Latent Profile Analysis ohne Einschränkungen über die Items der Beschwerde-Skala wären verschiedene Klassen von Personen. Jede dieser Klassen hätte ein eigenes Muster an Mittelwerten über die elf Items (L. K. Muthén & B. O. Muthén, 1998 – 2004).

Bei der vorliegenden Analyse wurde allerdings die Gruppierungsinformation bei der Schätzung des Modells berücksichtigt. Die latenten Mittelwerte wurden so auf die Trennung der drei Gruppen (Prä-, Postwerte, Bevölkerung) optimiert, während zusätzlich Klassen unterschiedlicher Varianzen ("Varianzmuster") geschätzt wurden, um unterschiedliche Variabilität um die Mittelwerte berück-

sichtigen zu können (L. K. Muthén & B. O. Muthén, 1998-2004; Dattatreya, 2002). Ergebnis dieser Analysen sind drei latente Mittelwertverteilungen (eine für jede Stichprobe bzw. Zeitpunkt) die von verschiedenen Varianzmustern umgeben sind.

Vorteil dieses Verfahrens gegenüber der Analyse der beobachteten Verteilungen ist, dass so festgestellt werden kann, ob es unterschiedliche Variabilität um die Items gibt und ob diese beachtenswert ist. Beispielsweise könnte es eine Gruppe von Personen geben, die sich durch besonders große Variabilität um diese Mittelwerte auszeichnet, für die die aufgrund der beobachteten Verteilungen ausgewählten Items vielleicht nicht sensitiv genug wären. Diese Items wären in diesem Falle zu liberal: Es würde für diese Personengruppe eine Veränderung angezeigt, obwohl noch gar keine vorliegt. Diese Analysen wurden mit MPlus (Version 3.11; L. K. Muthén & B. O. Muthén, 1998-2004) durchgeführt. Als Kriterium für das am Besten passende Modell wurden das Bayes Information Criterion (BIC, Schwarz, 1978) und die Stabilität der Lösung verwendet.

Ergebnis der einzelnen Analyseschritte sind verkürzte Skalen. Diese werden mittels Reliabilität und Effektstärke in Beziehung zu den Originalskalen gesetzt. Die Reliabilität wird dabei an allen Präwerten und der Bevölkerungsstichprobe bestimmt, damit jede Person nur ein einziges Mal berücksichtigt wird. Für eine gelungene Verkürzung spricht im Sinne der klassischen Testtheorie eine Reliabilität, die gleich oder oberhalb der zu erwartenden Reliabilität nach der Spearman-Brown-Formel ausfällt (Kempf, 2008). Im Sinne der praktischen Verwendbarkeit der verkürzten Skala sollte sie eine ähnliche Veränderungssensitivität aufweisen wie die Originalskala. Eine mögliche Operationalisierung stellen die Ermittlung und der Vergleich von Prä-Post-Effektstärken der Skalenfassungen dar (Lutz, Tholen, et al., 2006). Diese werden an den vorliegenden Prä- und Postwerten aus der Poliklinischen Psychotherapieambulanz Osnabrück bestimmt.

4.3. Ergebnisse

4.3.1. Untersuchung der beobachteten Mittelwerte

Aufgeteilt nach den drei Stichproben präsentiert Abbildung 4-1 die Mittelwerte und die dazugehörigen Konfidenzintervalle für die elf Items, die die Symptomskala des FEP bilden. Von der

generellen Tendenz zeigen alle Items die erwartete Ordnung der Gruppenmittelwerte: Die Präwerte sind durchweg höher als die Postwerte und die Bevölkerungswerte sind am Niedrigsten. Werden die vorformulierten Bedingungen angewendet, überlappen nur bei zwei Items die Verteilungen der klinischen Postwerte und der Werte der Normalbevölkerung und zeigen gleichzeitig einen Abstand zu den klinischen Präwerten: Item 6 und Item 12 (Itemformulierungen im Anhang). Die Postwerte gehören also mit hoher Wahrscheinlichkeit nicht mehr zu der Verteilung der belasteten Gruppe (den Präwerten) und die Überlappung der Konfidenzintervalle mit den Bevölkerungswerten zeigt an, dass die Belastung der Personen nach der Therapie nicht mehr von der Belastung der Bevölkerungsstichprobe zu unterscheiden ist. Alle drei formulierten Bedingungen sind also erfüllt.

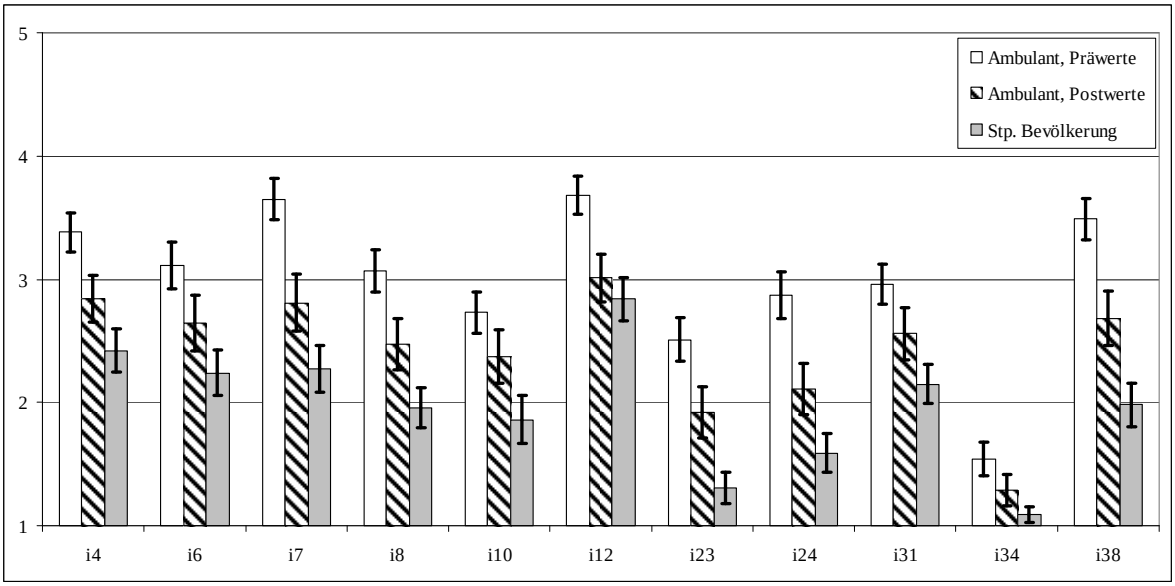

Abbildung 4-1: Mittelwerte und 95 %-Konfidenzintervalle der drei Analysegruppen für die elf Items der Skala "Beschwerden" des FEP (angegeben sind das maximale und minimale *N* für jedes Item; Präwerte *N* = 198-207; Postwerte *N* = 124-129; Stichprobe (Stp.) Bevölkerung *N* = 120; i4 = "Item 4" usf.).

Bei anderen Items (Item 4, Item 7, Item 8, Item 23, Item 24, Item 31 und Item 38) zeigt sich, dass alle drei Mittelwerte nicht-überlappend sind. Diese erfüllen die Bedingung drei nicht, d.h. die Personen sind zwar nicht mehr so belastet wie die Stichprobe vor der Behandlung, die Personen erreichen nach der Therapie in diesen Symptomen aber (noch) nicht das Belastungsniveau der Bevölkerungsstichprobe.

Nach den definierten Kriterien eignen sich die restlichen Items nicht für die Optimierung der Verlaufsmessung. Bei Item 10 ("…war ich sehr einsam und alleine") ist der Abstand zwischen Prä- und Postwerten nicht groß genug, d.h. es ist vielleicht zu insensitiv für den Fortschritt innerhalb der Therapie. Der niedrigere Mittelwert für die Bevölkerungsstichprobe zeigt aber, dass es sich ansonsten für die Messung von Belastung durchaus eignen würde. Bei Item 34 ("…dachte ich daran, mir das Leben zu nehmen") liegt das Problem vor, dass dieses Item einen deutlichen Bodeneffekt zeigt. Dieses Item zeigt einen so hohen Grad an Belastung an, dass es selbst in Stichproben klinisch belasteter nur selten angekreuzt wird. Hier ist nun ein Abwägen zwischen den Kriterien zur Itemauswahl nötig. Zur Veränderungssensitivität leistet dieses Item aufgrund der zu erwartenden Verteilung vielleicht nur einen geringen Beitrag. Nach inhaltlicher Relevanz für die Therapie als Indikator-/ Warnitem wäre eine Beibehaltung unter Umständen trotzdem sinnvoll (s. Tabelle 4-1; Fowler, 2012).

Zur Bestimmung der Reliabilität der verkürzten Skala (ohne Item 10 und Item 34) wurden die Prämessungen der Patienten und die Messungen der Bevölkerungsstichprobe hinzugezogen (s.o.). Es ergibt sich ein Cronbach-α = 0.91 (N = 319, das entspricht 97.6% der Gesamtstichprobe). Die vollständige Skala "Symptome" erreicht in allen vorliegenden abgeschlossenen Therapien eine Effektstärke von ES = 0.71 (N = 129; auf Standardabweichung der Präwerte standardisierte Mittelwertsdifferenz). Mit dieser verkürzten Skala wird eine Effektstärke von ES = 0.76 (N = 129) in der Gruppe der abgeschlossenen Therapien erreicht (s. Tabelle 4-3).

Dass Bootstrap-Vorgehen bietet sich wie beschrieben in Situationen mit kleinen Stichproben und ggf. verletzten Verteilungsannahmen an (s.o.). Es kann dabei von vornherein als eigenständige Analyse verwendet werden oder aber als Ergänzung der eben vorgestellten Betrachtung der Itemmittelwerte. Aus den drei Gruppen wurden jeweils 1000 Stichproben von 40 Personen gezogen und die Mittelwerte der Items bestimmt. In Abbildung 4-2 ist der Itemmittelwert über alle Replikationen eingetragen, und um diesen als Grenzen die Werte, bis zu denen die 95 % zentralsten Replikationen liegen. Es werden nur leicht asymmetrische Konfidenzintervalle gezeichnet. Bei den gewählten Parametern für die Replikationen fällt das Ergebnis für diese Analyse konservativer aus.

Drei der Items (Item 7, Item 12 und Item 24) erfüllen alle drei Bedingungen. Bei diesen drei Items überlappen sich die Konfidenzintervalle der Prä- und Postwerte nicht, es ist also unwahrscheinlich, dass bei Prä- und Postwerten dasselbe Belastungsniveau vorliegt. Dafür überlappen sich aber die Konfidenzintervalle von Postwerten und den Bevölkerungswerten: Nach der Therapie ist also kein Unterschied in den Verteilungen zwischen diesen beiden Gruppen mehr feststellbar. Item 38 erfüllt nur die ersten beiden Bedingungen. Die Reliabilität dieser Kurzversion aus vier Items (Prämessungen der Patienten und die Messungen der Bevölkerungsstichprobe, s.o.) beträgt Cronbach-α = 0.86 (N = 322; entspricht 98.5% der Gesamtstichprobe). Mit dieser Skala wird eine Effektstärke von ES = 0.79 (N = 129) in der Gruppe der abgeschlossenen Therapien erreicht.

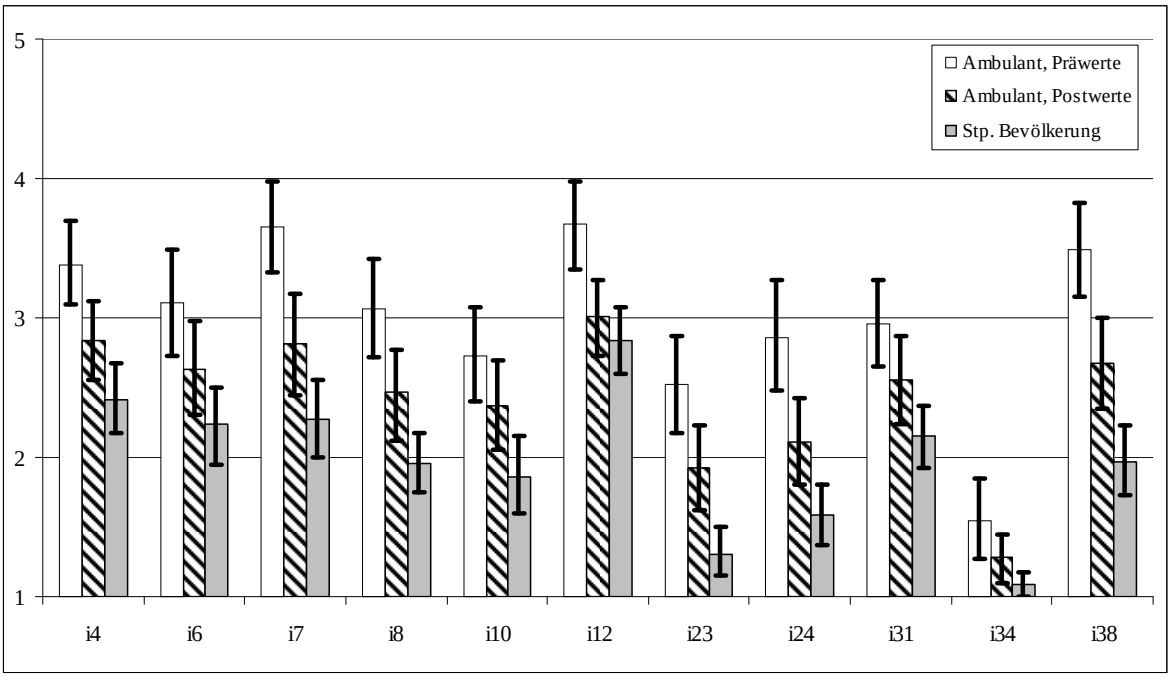

Abbildung 4-2: Mittelwerte und die 95 % zentralsten Mittelwerte aus 1000 Bootstraps mit je n = 40 als Grenzen für alle drei Analysegruppen bei den elf Items der Skala "Beschwerden" des FEP (angegeben sind das maximale und minimale N für jedes Item; Präwerte N = 198-207; Postwerte N = 124-129; Stichprobe (Stp.) Bevölkerung N = 120; i4 = "Item 4" usf.).

4.3.2. Untersuchung mittels Latent Profile Analysis

Zur Untersuchung wurden alle Items außer Item 34 herangezogen, da Items mit geringen Belegungen der meisten ihrer Kategorien zu verzerrten Schätzern führen (Holman, Lindeboom, et al.,

2003; Rost, 2004). Wie bereits oben geschrieben, sprechen genügend inhaltliche Gründe dafür, dieses Item im klinischen Alltag trotz allem zu erheben. In der *Latent Profile Analysis* wurden die Mittelwerte unter Berücksichtigung der Gruppen geschätzt. Zusätzlich wurde angenommen, dass es unterschiedliche Verteilungen gibt, d. h. unterschiedliche Varianzmuster um die Mittelwerte sollten identifiziert werden. Nach dem BIC (Rost, 2004) und der Stabilität der Lösung wurden zwei Varianzmuster als Lösung herangezogen. Abbildung 4-3 präsentiert die Mittelwerte und Konfidenzintervalle der veränderungssensitiven Items.

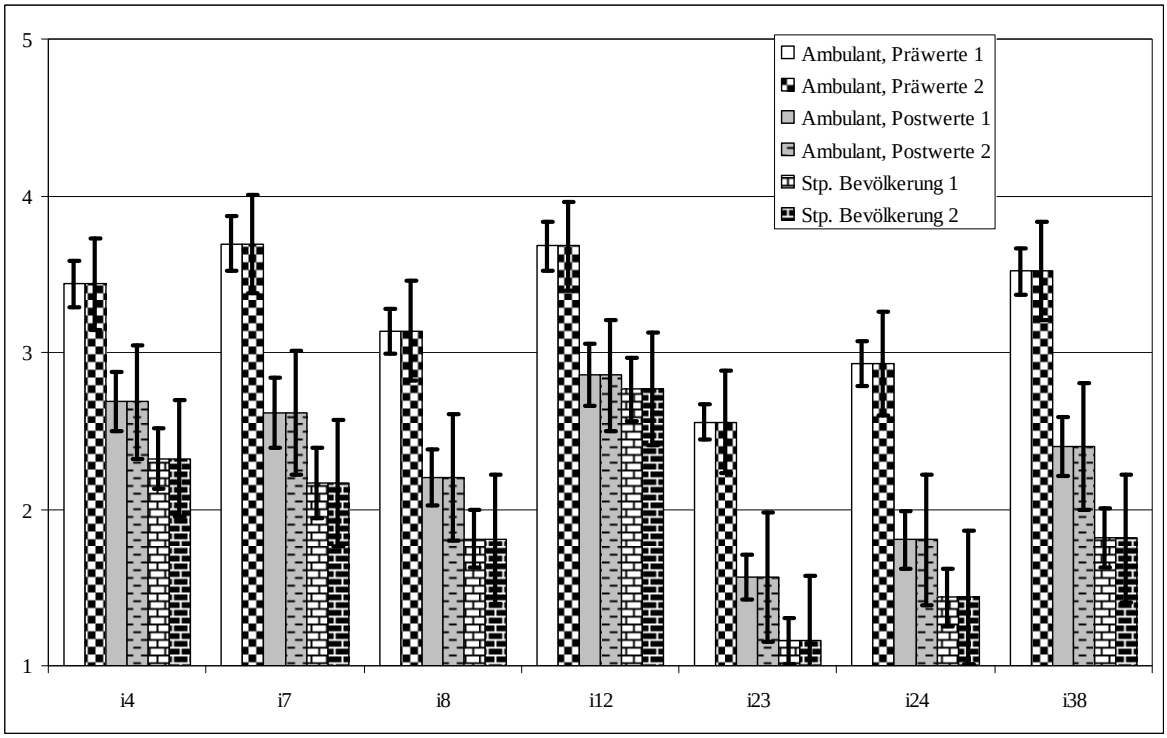

Abbildung 4-3: Veränderungssensitive Items; geschätzte latente Mittelwerte und 95 %-Konfidenzintervalle der Erhebungsgruppen für beide Varianzmusterklassen; erster Balken jeder Gruppe zeigt den latenten Mittelwert mit dem Konfidenzintervall aus dem ersten Varianzmuster (z. B. "Ambulant, Präwerte 1"); der zweite Balken zeigt denselben Mittelwert mit dem Konfidenzintervall aus dem zweiten Varianzmuster (z. B. "Ambulant, Präwerte 2").
Stp. Bevölkerung = Stichprobe Bevölkerung; i4 = "Item 4" usf.
geschätzte Gruppengrößen: Ambulant, Präwerte 1 *N* = 110.70; Ambulant, Präwerte 2 *N* = 88.30; Ambulant, Postwerte 1 *N* = 68.98; Ambulant, Postwerte 2 *N* = 55.02; Stp. Bevölkerung 1 *N* = 66.75; Stp. Bevölkerung 2 *N* = 53.25

Ergebnisse werden genauso dargestellt wie vorher, allerdings finden sich für jede der Gruppen nun zwei Mittelwertsbalken – einer mit dem geschätzten Konfidenzintervall aus dem ersten Varianzmuster (jeweils der erste Balken; erwartete 55.6 % der Stichprobe) und einer mit dem aus dem

zweiten Muster (jeweils der zweite Balken; erwartete 44.4 % der Stichprobe). In Abbildung 4-3 sind also für den einen geschätzten latenten Mittelwert in der ambulanten Stichprobe zwei Balken bei jedem Item zu sehen: Einer mit der Varianzschätzung aus dem ersten Varianzmuster und einer mit der Varianzschätzung aus dem zweiten Muster. Es zeigt sich, dass das zweite Varianzmuster durchwegs konservativer ist (größere Konfidenzintervalle produziert).

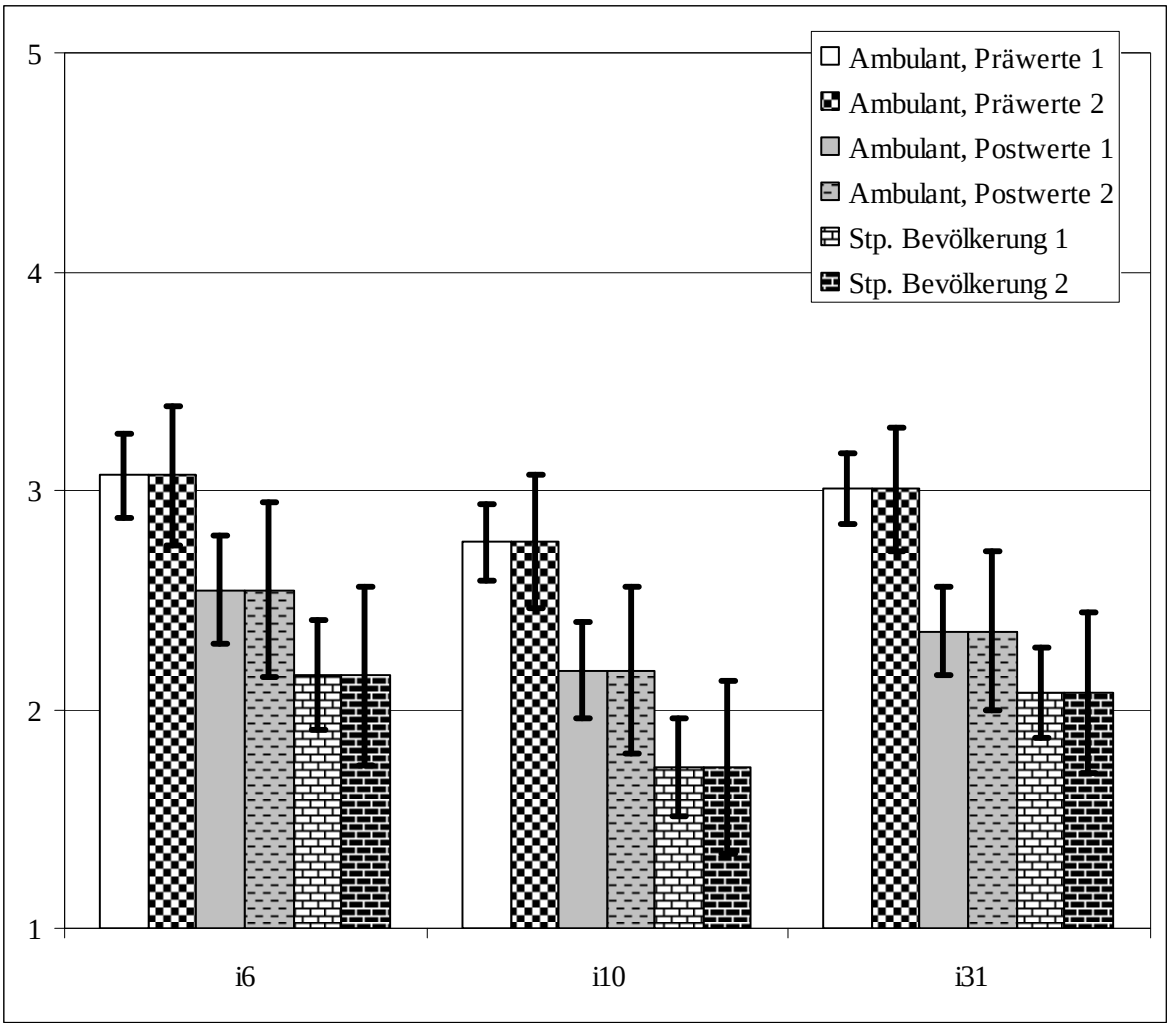

Abbildung 4-4: Nicht veränderungssensitive Items; geschätzte latente Mittelwerte und 95 %-Konfidenzintervalle der Erhebungsgruppen für beide Varianzmusterklassen; erster Balken jeder Gruppe zeigt den latenten Mittelwert mit dem Konfidenzintervall aus dem ersten Varianzmuster (z. B. "Ambulant, Präwerte 1"); der zweite Balken zeigt denselben Mittelwert mit dem Konfidenzintervall aus dem zweiten Varianzmuster (z. B. "Ambulant, Präwerte 2").
Stp. Bevölkerung = Stichprobe Bevölkerung; i4 = "Item 4" usf.
geschätzte Gruppengrößen: Ambulant, Präwerte 1 *N* = 110.70; Ambulant, Präwerte 2 *N* = 88.30; Ambulant, Postwerte 1 *N* = 68.98; Ambulant, Postwerte 2 *N* = 55.02; Stp. Bevölkerung 1 *N* = 66.75; Stp. Bevölkerung 2 *N* = 53.25

Zur Analyse, welche Items zur Veränderungsmessung geeignet sind, gelten dieselben Kriterien, wie im vorigen Analyseschritt, allerdings werden die Regeln nun so angewendet, dass immer das konservativere, größere Intervall herangezogen wird. In Abbildung 3 ist für alle Items zu sehen, dass sich die Konfidenzintervalle der ambulanten Postwerte und der Bevölkerungsgruppe für diese Items überschneiden, nicht aber mit denen der ambulanten Präwerte. Diese Items (Item 4, Item 7, Item 8, Item 12, Item 23, Item 24 und Item 38) erfüllen also alle drei Bedingungen. Wird die Skala "Beschwerden" auf diese sieben Items gekürzt, hat sie eine Reliabilität von Cronbach-α = 0.91 (N = 320; entspricht 97.9% der Gesamtstichprobe) und mit ihr wird in der ambulanten Stichprobe eine Effektstärke von ES = 0.76 (N = 129; alle erhobenen Prä- und Postwerte) erreicht.

Zum Vergleich sind in Abbildung 4-4 die Items präsentiert, die gemäß Kriterien nicht veränderungssensitiv sind. Auch hier sind für jeden Gruppenmittelwert zwei Balken eingetragen, einer mit dem Konfidenzintervall aus dem ersten Varianzmuster und einer mit dem Konfidenzintervall aus dem zweiten. Mindestens eines der Konfidenzintervalle der Postwerte überschneidet sich mit einem der Präwerte, womit nicht von einer Veränderung in der Interventionszeit ausgegangen wird.

Tabelle 4-3 präsentiert einen Überblick über die Qualitätsmerkmale der erstellten Kurzformen. Alle Kurzformen zeichnen sich durch eine gute Reliabilität aus, die über den bei Verkürzung auf die Anzahl essenziell paralleler Items erwarteten Reliabilitäten liegt (Kempf, 2008). Die Effektstärken fallen sogar (deskriptiv) höher aus. Die kürzeste Skala mit der höchsten Veränderungssensitivität wird durch die Verwendung des empirischen Bootstraps erreicht. Die anderen beiden Methoden erreichen ähnliche Werte.

Tabelle 4-3: Effektstärken (ES), Reliabilitäten und erwartete Reliabilitäten bezogen auf die vollständige Skala "Beschwerden" des FEP für alle verwendeten Kurzformen; eine akzeptable Kurzform ist dann erstellt, wenn sie in ähnlicher Weise veränderungssensitiv ist (ähnlich hohe Effektstärke) und ihre Reliabilität nicht unter das durch die Skalenverkürzung zu erwartende Niveau fällt.

Skala	*ES*	Cronbach-α	Erwartete Reliabilität
Vollständige Skala "Beschwerden" (11 Items)	0.71	0.92	--
Skala "Beschwerden" nach deskriptiver Analyse (9 Items)	0.76	0.91	.90
Skala "Beschwerden" nach deskriptiver Analyse mit Bootstrap (4 Items)	0.79	0.86	.81
Skala "Beschwerden" nach Latent Profile Analysis mit Gruppierungsinformation (7 Items)	0.76	0.91	.88

4.4. Diskussion

Ziel der Untersuchung war es, Items zu identifizieren, die besonders für die Verlaufsmessung geeignet sind. Dabei sollte eine Bevölkerungsstichprobe als Referenz herangezogen werden. Angewendet wurden deskriptive Vergleiche über Konfidenzintervalle der beobachtbaren Verteilungen und der resultierenden Schätzer aus einer Latent Profile Analysis bei gruppierten Daten. Als besonders veränderungssensitiv wurden bei allen Vorgehensweisen die Items bezeichnet, die sich dadurch auszeichneten, dass die Verteilungen der Mittelwerte sich gemäß mindestens zweier von drei formulierten Kriterien veränderten: Die Prä- und Postwerteverteilungen von Patienten sollten sich nicht überlappen (d.h. es gibt eine Symptomreduktion während der Intervention), die Verteilungen von Bevölkerungswerten und Präwerten sollten sich nicht überlappen (d.h. klinisch bedeutsame Belastung sollte gegeben sein) und im günstigsten Fall sollten sich die Verteilungen der Post- und Bevölkerungswerte überlappen (d.h. Symptomreduktion auf ein akzeptables Maß).

Die Vorgehensweisen zielten auf unterschiedliche Aspekte ab, resultierten aber in ähnlichen Ergebnissen. Am konservativsten war die Itemauswahl über das Bootstrapverfahren bei dem nur vier von elf Items beibehalten wurden. Die Ergebnisse der Latent Profile Analysis deckten zwei Varianzmuster auf, von denen eines eine breitere Streuung um die Mittelwerte der Analysegruppen hatte. Wurden die breiteren Konfidenzintervalle des Varianzmusters mit den breiteren Streuungen herangezogen, resultierte dies in einer Skala von sieben Items. Diese enthielt auch die vier durch den Bootstrap identifizierten Items.

Die erstellten Kurzformen zeigen die erwünschten Eigenschaften einer hohen Reliabilität bei gleichzeitig hohen Effektstärken (Tabelle 4-3). Während das Ergebnis des Bootstraps einen Kern von Fragen lieferte, der in ähnlichen ambulanten Settings relativ sicher herangezogen werden könnte, machte die Latent Profile Analysis darauf aufmerksam, dass es eine Personengruppe gibt, die die Items mit einer breiteren Streuung beantwortet. Für diese Personen wäre eine Auswahl von Items nach den beobachteten Streuungen ohne Bootstrapping zu liberal, d.h. es würde zu früh eine Veränderung angezeigt.

Nach den eingangs diskutierten Kriterien würden aufgrund dieser Ergebnisse die Versionen mit vier (Bootstrap) oder sieben Items (Latent Profile Analyse) zur Auswahl stehen. Das Kriterium der beobachteten Veränderung in Effektstärken fällt befriedigend, d.h. zumindest auf dem Niveau der Originalskala mit elf Items, aus und die Reliabilitäten bleiben oberhalb der erwarteten Reliabilität bei Skalenverkürzung (s. Tabelle 4-3). Werden die anwendbaren Kriterien aus Tabelle 4-1 auf diese Analyse angewendet, zeigen die verbleibenden Items keine Decken-/Bodeneffekte und es bleibt mehr als ein Item zur Messung von "Beschwerden" übrig. Ein kritischer Punkt ist die Entscheidung, die bei dem Item 34 ("…dachte ich daran, mir das Leben zu nehmen") getroffen werden muss. Vermutlich würde das Kriterium der inhaltlich begründeten Auswahl (Meier, 1997; Fowler, 2012) bei diesem Item so hoch gewertet, dass es trotz Bodeneffekt in der Skala verbleiben würde. Eine endgültige Entscheidung über die Güte der entwickelten Skala könnte vermutlich nur mittels einer Kreuzvalidierung getroffen werden.

Mit den vorliegenden Ergebnissen der Latent Profile Analysis kann allerdings sehr wahrscheinlich ein Vorliegen von systematischen Verzerrungen oder Antwortstilen (s. Tabelle 4-1) in größerem Maße ausgeschlossen werden. Gäbe es Personengruppen in dem Sample, die z. B. lediglich eine Kategorie verwenden, würde sich dies in einem auffälligen Varianzmuster niederschlagen. Ein solches (z. B. keine Varianz oder gegen sehr große Werte strebend) wurde nicht identifiziert. Dies zeigt insgesamt, dass mit den verwendeten Stichproben und methodischen Kriterien Items identifiziert werden können, die deutliche systematische Trends über die Zeit zeigen.

Methodisch sind zwei Punkte zu diskutieren. Die Größe der ermittelten Konfidenzintervalle hängt von der Stichprobengröße ab und kann dementsprechend in Ergebnissen münden, in denen alle Konfidenzintervalle überlappen oder aber keines. Dies führt indirekt zu dem eingangs erwähnten Problem des multiplen Testens und der α-Fehler Kumulierung. Eine Orientierung für die angemessene Stichprobengröße (oder Größe der Stichprobengrößen beim Bootstrap) können a priori durch Überlegungen zur Stichprobenumfangsplanung geben. Je nach theoretisch erwarteter Veränderungsspanne und der damit verbundenen Effektstärke (J. Cohen, 1988) können Stichproben so gewählt werden, dass beispielsweise mittlere Effekte entdeckt werden können.

Ein zweiter Punkt ist, dass die Verwendung anspruchsvoller statistischer Modelle wie der Latent Profile Analysis eine Unsicherheit der Anwendbarkeit und Interpretierbarkeit der Befunde mit sich bringt. Ein Problem, das sich bei der Analyse der beobachteten Verteilungen (mit oder ohne Bootstrapping) nicht stellt. Die in diesem Beispiel verwendete Stichprobengröße könnte als zu gering eingeschätzt werden, um das Modell hinreichend gut schätzen zu können. Die Debatte um angemessene Stichprobenumfänge zur Identifikation der Modelle hält derzeit an (Davier, 1997; Nylund, Asparouhov, & Muthén, 2007; Rost, 2004; Tollenaar & Mooijaart, 2003).

Eine praktische Frage bezieht sich auf die Verwendung der Referenzgruppen. Die vorliegende Auswahl beantwortet nur die Frage, ob im Zeitraum einer Therapie eine Reduktion der Symptomatik beobachtet wird. Diese Reduktion ist nicht mit "Genesung" gleichzusetzen, sondern sie zeigt nur an, dass sich in diesem Symptom in dem gewählten Zeitraum überhaupt eine Veränderung feststellen lässt.

Mit dem gewählten Vorgehen kann wie eingangs bereits erläutert kein Nachweis erbracht werden, dass sich Personen in Therapie stärker in diesen Symptomen verändern als Personen, die keine Therapie erhalten. Die vorliegende Untersuchung arbeitet mit Daten aus der Routineversorgung, in der es oftmals nicht möglich oder nicht vertretbar ist, Daten Unbehandelter zu erheben (Howard et al., 1996; Krause & Lutz, 2009). Da hier in der Regel als wirksam beurteilte Behandlungsmethoden zum Einsatz kommen, ist der Vergleich mit einer solchen Gruppe auch weniger zwingend, da über die Ermittlung der Veränderungen in den Items kein Rückschluss auf die generelle Effektivität der

Behandlung vorgenommen werden soll. Dennoch könnte als weitere Gruppe Daten von Unbehandelten aufgenommen werden und somit als vierte Gruppe eine weitere Referenz abgeben. Auch eine zweite Erhebung in der Bevölkerungsstichprobe nach einer Zeit, die dem Mittelwert oder Median der Behandlungsdauer in den ambulanten Vergleichsdaten entspricht könnte das vorgeschlagene Vorgehen sinnvoll ergänzen. So könnte die Variabilität der Symptome über diesen Zeitraum in einer nicht belasteten und nicht in Therapie befindlichen Gruppe bestimmt werden.

Insgesamt erbringen die vorgeschlagenen Vorgehensweisen in der Praxis anwendbare Analyseergebnisse, die bei der Optimierung von Verlaufsmessungen angewendet werden können. Die Erweiterung um Bedingungen für den Vergleich mit Bevölkerungsstichproben stellt für den Bereich der Routineversorgung eine nützliche Erweiterung dar, da sie nicht nur einen negativen Standard in die Auswahl der Items einbringt ("Welches Belastungsniveau sollte nach einer Intervention nicht mehr gegeben sein?"), sondern ein positives Ziel definiert, welches (Un-)Belastungsniveau erreicht werden könnte (Krause & Lutz, 2009).

In dieser Arbeit wurde zu illustrativen Zwecken und der Übersichtlichkeit halber eine Skala mit nur elf Items verwendet. Gewinne in Bezug auf eine verbesserte Erhebungsökonomie fallen bei längeren Originalinstrumenten größer aus als in diesem Fall. Offen bleibt die Erweiterung der Methode auf multidimensionale Ergebniskriterien. Die Messung eines einzigen Bereiches als Ergebniskriterium reicht sicher nicht, um den komplexen Prozess, der innerhalb einer Psychotherapie abläuft, angemessen abzubilden (Lutz & Böhnke, 2010; Schulte, 1993). Dementsprechend müssten veränderungssensitive Kurzformen für verschiedene Ziel- oder Veränderungsbereiche der Psychotherapie entwickelt werden.

5. Diskussion

Zum Abschluss der Arbeit sollen die drei Studien zusammengefasst und gemeinsam diskutiert werden. Dafür werden die Ergebnisse der drei Studien kurz dargestellt und offene Forschungsaspekte angerissen. Für eine detaillierte Diskussion der einzelnen Studien sei auf die jeweiligen Kapitel verwiesen. Daran schließt sich eine übergreifende Diskussion der Grundannahmen der Studien an. Dies sind die Bedeutung von Fragebogenerhebungen im Monitoring (5.2.1), die faktorielle Validität des FEP (5.2.2) und die Bedeutung für Messwiederholungen (5.2.3), die Auswahl des spezifischen IRT Modells (5.2.4) sowie die Angemessenheit von IRT Modellen zur Messung psychischer Belastungen (5.2.5) und abschließend die Notwendigkeit des Einsatzes von Fragebogenkurzformen (5.2.6). Die Arbeit endet danach mit einem Ausblick möglicher Anwendungen der vorliegenden Ergebnisse (5.3).

5.1. Zusammenfassung der Arbeiten

5.1.1. Studie I

In der ersten Studie wurde untersucht, ob das Rasch-Modell zur Modellierung dichotomer Fragebogendaten in der Patientenorientierten Versorgungsforschung geeignet ist. Augenmerk wurde dabei auf a) die Verwendung von "open-source" Software (Culpepper & Aguinis, 2010), b) die Stichprobengröße und c) nicht-normale Verteilungen der Stichproben gelegt. Es wurden drei R-Pakete genutzt (eRm, Mair & Hatzinger, 2007; ltm, Rizopoulos, 2006; mixRasch, Willse, 2011), die jeweils andere Schätzer für die Modelle verwenden (Details siehe Kapitel 2.1.2). Die Ergebnisse zeigten, dass die Reproduktion der Itemparameter bei allen Stichprobengrößen gut gelingt (Korrelationen > .97), aber dennoch signifikante Zuwächse in der Schätzgenauigkeit der Itemparameter bis $N = 1000$ zu verzeichnen waren. Die Verteilungsform hatte keinen differentiellen Einfluss, womit IRT-Modelle auch dann zuverlässig schätzbar sind, wenn nicht-normale Verteilungsdaten vorliegen. Insgesamt konnte allen Paketen eine Anwendungsempfehlung ausgesprochen werden, auch wenn eRm am besten abschnitt (Tabelle 2-34).

Aus Sicht der Patientenorientierten Versorgungsforschung war in dieser Studie zu prüfen, ob die verwendeten R-Pakete IRT-Analysen in angemessener Qualität möglich machen, ohne die Kos-

ten kommerzieller Softwareprodukte zu verursachen. Alle drei Programme führten unter den gewählten Bedingungen insgesamt zu akzeptablen Schätzern, und auch der Fallstudienvergleich mit kommerzieller Software (Tabelle 2-10 bis Tabelle 2-15) zeigte, dass die Ergebnisse im Vergleich mit Programmen mit denselben Schätzern identisch waren. Damit kann zumindest für den Bereich der IRT-Modelle festgehalten werden, dass R (R Development Core Team, 2010) eine Möglichkeit zur weiteren Reduktion der Distanz zwischen Wissenschaftlern und Praktikern darstellt, da es verglichen mit kommerziellen Produkten bei weitaus geringeren Kosten die Implementierung hochqualitativer und moderner Analysemethoden ermöglicht.

Für zukünftige Studien stellt neben der Untersuchung von noch höheren Itemzahlen und polytomer Antwortformate die Untersuchung von fehlenden Werten auf Itemebene die nächste Herausforderung dar. Fehlende Werte beziehen sich nicht nur auf die auftretende Auslassung von Items durch die Patienten bei der Bearbeitung eines Fragebogens, sondern in heutigen Testanwendungen besonders auf die Verwendung von Teilen von Fragebögen, z.B. in Kalibrierungsstudien (Holman, Lindeboom, et al., 2003; Walker et al., 2010), bei computer-adaptiven Tests (Wainer, 2000) und der systematischen Verwendung von Testteilen wie in Kapitel 3 vorgeschlagen. Es weisen erste Studienergebnisse aus Kompetenzerhebungen im Bildungsbereich hier auf deutlichere Vorteile der CML-Methode gegenüber der MML-Methode hin, wenn die für die MML-Methode verwendete Verteilungsannahme empirisch nicht gegeben ist (Ullrich et al., 2012).

5.1.2. Studie II

In der zweiten Studie wurden zwei Pakete aus Studie I genutzt (eRm, Mair & Hatzinger, 2007; ltm, Rizopoulos, 2006), um das Rasch-Modell für die Konstruktion verschiedener Subskalen eines bestehenden Instrumentes zu schätzen. Nach dem Rasch-Modell konstruierte Skalen haben den Vorteil, dass der Summenscore eine direkte Ermittlung des Personenparameters ermöglicht und dieser über verschiedene Testformen auf der latenten Dimension verglichen werden kann (siehe Abbildung 3-8 für eine Anwendung). Es wurden zwei Kurzfassungen in der Studie konstruiert und mittels des Vergleiches der Testinformationsfunktionen für die gewählten Optimierungsbereiche der Subskalen getestet. Die Verwendung des Bootstrap-Tests zum Vergleich der Messqualität in

den Zielbereichen erwies sich als praktikabel und aussagekräftig: Die optimierten Testfassungen erreichten zuverlässig eine bessere Messqualität als zufällige Auswahlen von Items aus der Skala.

Die Vorgehensweise der Optimierung der Testkurzformen ist ähnlich der Vorgehensweise für die Definition unterschiedlicher Cut Offs für bestimmte Populationen (Lambert & Ogles, 2009; Tingey et al., 1996) und wurde für mehrere Kriterien demonstriert. Die resultierenden Testformen zeigen, dass es wichtig ist, anwendungsbezogen über die Optimierungsbereiche nachzudenken. In einer Praxisanwendung wie dem TK-Modellvorhaben (Lutz, Böhnke, Köck, et al., 2011; Wittmann et al., 2011) hätte dieses Vorgehen in der folgenden Weise verwendet werden können: Bei der Erhebung zur Aufnahme war ein Kriterium die nachgewiesene Belastung in einem psychometrischen Instrument. Da zumindest für das Brief Symptom Inventory diskutiert wird, ob es ein eindimensionales Instrument ist (Cyr et al., 1985; Derogatis, 1993; Meijer et al., 2011; Thomas, 2012; Tran et al., 2012), könnte dieser diagnostische Schritt verkürzt werden, indem nur die Items verwendet werden, die optimal entscheiden lassen, ob der Patient ober- oder unterhalb des Cut Off für klinische relevante Belastung liegt. In vergleichbarer Weise können Mittel der Qualitätssicherung und Diagnostik weiter für die jeweiligen Anwendungen optimiert werden.

5.1.3. Studie III

In der dritten Studie wurden am Beispiel der Beschwerden-Skala des FEP (Lutz et al., 2009; Lutz & Böhnke, 2008) drei Methoden der Bestimmung der Sensitivität von Items für Veränderung verglichen. Die erste Methode verwendete die beobachteten Mittelwerte und Streuungen der Items und führte vermutlich zu einer Überschätzung der Anzahl sensitiver Items. Das Bootstrapverfahren mit gleich gewichteten Ziehungen aus den drei Stichproben zeigte durch die geringere Stichprobengröße der Ziehungen größere Standardfehler gegenüber den Konfidenzintervallen aus den beobachteten Stichprobenstatistiken und fiel konservativer aus. Dieses Vorgehen ermöglichte aber gerade auch in allen Stichproben das Konfidenzintervall unabhängig von der jeweiligen Stichprobengröße zu bestimmen. Dieses Vorgehen kann optimiert werden (z.B. N in den Ziehungen entspricht dem Gesamt-N der kleinsten Stichprobe), doch bildeten die Scores der vier verbleibenden Items eine Skala, die sich am Veränderungssensitivsten und hoch reliabel zeigte.

Die Latent Profile Analyse identifizierte zwei verschiedene Varianzmuster und legte damit nahe, dass es zwei Klassen von Personen in den Stichproben gab, die sich durch eine unterschiedliche Variabilität zu den verschiedenen Messzeitpunkten auszeichneten. Die Klassen waren in Hinsicht auf dieses Kriterium geordnet, d.h. die eine Klasse zeigte in allen Items eine höhere Variabilität als die andere. Die Verwendung derselben Kriterien zur Beurteilung des Fortschrittes hätte für die eine Klasse eine Über-, für die andere eine Unterschätzung des Therapieffektes zur Folge. Wenn die Latent Profile Analysis auch ein akzeptabler Mittelweg zwischen den vorigen beiden Methoden zu sein scheint, ist sie stärker als das Bootstrap-Verfahren von den Annahmen der Eindimensionalität der Items und der Annahme von Normalverteilungen der Items abhängig.

Insgesamt stellen Bootstrap und Latent Profile Analysis mögliche Erweiterungen der bisherigen Vorgehensweise zur Identifikation von veränderungssensitiven Items dar, doch sollte weitere Forschung untersuchen, wie die Stichprobengrößen für den Bootstrap sinnvoll vorausgewählt werden können. Inhaltlich sollte außerdem bei der Latent Profile Analysis untersucht werden, ob unterschiedliche Arten der Variabilität vorhersagbar sind und damit Rückmeldesysteme an das Nutzungsverhalten des Fragebogens angepasst werden können.

5.2. Diskussion weiterführender Aspekte

5.2.1. Verwendung von Fragebögen im Monitoring

Diese Arbeit setzt sich mit dem Fragebogen als Mittel in der Diagnostik auseinander, daher muss zunächst betont werden, dass Fragebogeninstrumente nicht allein Kriterien für die Bewertung von Psychotherapie sein können. Evans (2012) vergleicht allgemeine Belastungsmaße mit dem Fiebermessen (siehe für ähnliche Metapher Lambert, Hansen, et al., 2001). Das Fieber ist ein relevanter, aber unspezifischer Marker von einer körperlichen Entzündungsreaktion. Es ist sicher Ziel, das Fieber zu senken, aber es muss auch beachtet werden, dass ein Fieberrückgang keine Beseitigung der Ursachen ist. Somit ersetzen Rückmeldungssysteme nicht die klinische Evaluation von Testergebnissen, sondern diese sind nur eine Hilfe und Unterstützung im Prozess der Therapie (Lutz, 2011). Auf Fragebögen basierende Vorhersagen und Monitorings können für relevante diagnostische Bereich blind sein, daher können sie nur ein Baustein in einer multimodalen und multi-

dimensionalen Strategie sein, um das Augenmerk der Therapeuten systematisch auf bestimmte Aspekte zu lenken (Lutz, Böhnke, Köck, et al., 2011). Aktuelle Publikationen betonen besonders die kommunikative Funktion, die Fragebogenerhebungen als Vermittler von kondensierter diagnostischer Information im Therapieprozess übernehmen, die so nicht langwierig in der Sitzung durch den Therapeuten erhoben werden muss (Duncan, 2012; Evans, 2012; Lambert, 2012).

Dies verdeutlicht, dass kurze, effiziente Erhebungen keinen Selbstzweck darstellen, sondern immer im Kontext der jeweiligen Anwendung zu beurteilen sind. Die Ergebnisse zur Verbesserung von Service und Ergebnis aus der Qualitätssicherung und Patientenorientierten Versorgungsforschung begründen die Verwendung von Fragebogeninstrumenten (Carlier et al., 2012; Lambert, 2007; Poston & Hanson, 2010; und detaillierter Überblick in Kapitel 1.4.2). Auch die gut dokumentierten Unterschiede zwischen mechanischer und klinischer Prädiktion unterstützen, dass die alleinige Verwendung klinischer Urteile der Therapeuten und Mediziner nicht ausreichend zur Bewertung und Unterstützung klinischer Verläufe ist (Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hannan et al., 2005; Katsikopoulos et al., 2008; Meehl, 1954).

Zusammengenommen wurden in der vorliegenden Arbeit mehrere Vorgehensweisen zur Optimierung der Messungen im Rahmen von Feedbacksystemen in der ambulanten Psychotherapie geprüft. Es zeigte sich, dass die Messqualität und Konstruktvalidität mit geringem Aufwand angemessen geprüft werden können (Studie I) und dass die Verbesserung der Effizienz eines Monitoringsystems durch verschiedene Methoden zur Skalenverkürzung möglich ist (Studien II + III). Mit dem FEP (Lutz & Böhnke, 2008) wurde dabei ein junges Instrument geprüft, das aber bereits in verschiedenen Ambulanzen und niedergelassenen Praxen im Einsatz befindet (z.B. Schöttke, Sembill, Eversmann, Waldorf, & Lange, 2011). Studien II + III wurden an Daten aus der Routineversorgung durchgeführt, also an Stichproben, für die die Verwendung des FEP intendiert ist. Darüberhinaus ähneln sich die Ergebnisse einer Reihe von Studien mit dem Instrument über qualitativ unterschiedliche Stichproben, so dass die Ergebnisse zur Konstruktvalidität als stabil angesehen werden können (Böhnke & Lutz, submitted; Lutz, Tholen, et al., 2006; Schürch et al., 2009). Dieses Argument wird auch noch einmal dadurch gestärkt, dass der FEP ursprünglich basierend auf der Klassischen Testtheorie entwickelt wurde, sich die Erkenntnisse zur Konstruktvalidität

mit Methoden der probabilistischen Testtheorie aber bestätigen. Dies unterstreicht, dass beide Perspektiven sinnvoll miteinander verbunden werden können (Bechger et al., 2003; Blanchin et al., 2011; Hays, Brown, Brown, Spritzer, & Crall, 2006; Prieto, Alonso, & Lamarca, 2003; Schürch et al., 2009).

5.2.2. Faktorielle Validität

Es zeigte sich, dass zumindest für zwei das Phasenmodell operationalisierende Skalen (Wohlbefinden, Symptome; Howard et al., 1993) eine Dimension angenommen werden kann. Ein ebenfalls repliziertes Ergebnis ist, dass die angestrebte Erhebung von Störungsursachen (Schulte, 1993) durch die Berücksichtigung der Skala Inkongruenz im Fragebogen (Grawe, 1998, 2004) konzeptuell insofern zu glücken scheint, als dass Wohlbefinden, Symptome und Inkongruenz im Wesentlichen eine Dimension bilden, was für eine Konvergenz der Perspektiven auf die Belastungslage spricht. Damit hat sich in Bezug auf die Dimensionalität ein weiteres Mal bestätigt, dass bei der Messung psychischer Belastungslagen in der Regel ein Generalfaktor dominiert.

In Studie II wurden alle Items außer jenen der Interpersonellen Skala ausgewertet, da sich bereits in vorigen Studien zeigte, dass die Interpersonelle Dimension einen eigenen Faktor formt (Lutz & Böhnke, 2008; Schürch et al., 2009). Zur Überprüfung dieses Vorbefundes wurden alle 40 Items in einem Bifaktor-Modell (g + vier Domänenfaktoren) zusammen analysiert. Es ergaben sich für die Korrelationen des Generalfaktors mit den ersten drei Domänenfaktoren Werte zwischen .67 und .74. Dies deutet für einen Großteil der Items immer noch auf eine starke erste Dimension hin. Der vierte Faktor, auf dem ausschließlich Items mit interpersonellem Inhalt laden[34], zeigt dagegen nur eine Korrelation von .27 mit dem Generalfaktor. Somit reihen sich die Ergebnisse in die bekannte Literatur ein, die z.B. für die als sehr heterogen konstruierte "Symptom Checklist" (Derogatis, 1977) mit ihren neun Subskalen eher eine Dimension annimmt (Halstead et al., 2007; Meijer et al., 2011; Thomas, 2012; Tran et al., 2012). Dabei muss beachtet werden, dass die Wohl-

[34] Dies sind die Items (Itemnummer in Klammern): "...war ich leicht von anderen zu überreden" (13), "...hatte ich Probleme, Aggressionen zu zeigen, wenn nötig" (18), "...war ich selbstbeherrscht" (19), "...war ich leicht von anderen auszunutzen" (20), "...hatte ich Probleme, vertrauten Personen gegenüber Ärger zu äussern, wenn nötig" (25), "...gingen mir die Probleme anderer Menschen schnell zu nahe" (26), und "...war ich Teil einer erfüllten und intimen Beziehung" (27).

befindens-Symptom-Dimension im Sinne eines multimodalen und multidimensionalen Erhebungssystems nur einer von mehreren Erhebungsaspekten ist (Seidenstücker & Baumann, 1987). Der Nachweis, dass diese Facetten des FEP eindimensional sind und die Entwicklung von Kurzversionen für diesen Aspekt bedeuten natürlich nicht, dass nur dies erhoben werden sollte. Das Vorgehen ist prinzipiell anwendbar auf jede Art von Erhebungsinstrument und es sollten mehrere eng umrissene, klar definierte eindimensionale Messungen erfolgen (G. T. Smith et al., 2009). Wenn die Verwendung computer-adaptiver Tests nicht möglich ist, sollten für die notwendigen Dimensionen statische Kurzversionen mittels IRT entwickelt, und diese gemeinsam genutzt werden. Die Zahl der berücksichtigten Dimensionen sollte sich dabei einerseits nach den diagnostischen Erfordernissen und andererseits den Konstrukten, für die tatsächlich ein inkrementeller Wert festgestellt werden kann, richten (Hunsley & Mash, 2005; Meyer et al., 2001).

Hier liegt eine Schwäche der Studien: Es wird kein Nachweis erbracht, dass sich die Kurzformen des FEP als "evidenzbasierte Erhebungen" qualifizieren lassen (Hunsley & Mash, 2005). Dies war auch nicht das Ziel der Arbeit, auch wenn diese Perspektive auf den diagnostischen Prozess eine wesentliche Begründung für die vorliegenden Untersuchungen darstellen. Studien zur Entwicklung von Instrumenten befinden sich immer in einem Netzwerk und die hier vorgestellten Möglichkeiten der Verkürzung der Erhebungen machen es wahrscheinlicher, dass die Fragebögen auch in der Praxis eingesetzt werden. Damit kann dann auch wiederum leichter unter Praxisbedingungen geprüft werden, ob die Fragebögen inkrementelle Validität besitzen bzw. als evidenzbasiert eingeschätzt werden können. In dieser Weise können sich die Feedbackforschung und die Fragebogenentwicklung gegenseitig befruchten und entwickeln. Der Review zur aktuellen Lage der Psychotherapieforschung von (Padberg, 2012), der den Forschungsstrang mit keinem Wort erwähnt, verdeutlicht, wie dringlich eine weitere Verbreitung dieser seit bald knapp 15 Jahren vorhandenen praxisorientierten Forschung ist (Howard et al., 1996; Lambert, 2001; Lutz, 2002).

Die Frage der Dimensionalität spielt auch aus dem Grund eine Rolle, da eine übliche Forderung an Kurzversionen von Messinstrumenten ist, dass sie die intendierte Struktur mit allen Faktoren-/ Domänenfaktoren widergibt und diese auch reliabel misst (G. T. Smith, McCarthy, & Anderson, 2000). Ob dieses geforderte Ziel überhaupt sinnvoll erfüllbar ist, ist derzeit offen.

Sinharay und Kollegen (2010) zeigten mittels realer wie simulierter Daten, dass der Bericht von Subscores auch bei hohen Itemzahlen und leicht mehrdimensionalen Konstrukten oft keinen zusätzlichen Wert über den Bericht eines Gesamtscores hinaus hat. Vor dem vorliegenden Hintergrund kann daher festgehalten werden, dass eine reliable Messung mit Kurzfassungen des FEP für die Hauptdimension gelingt, dass eine fehlende weitere Ausdifferenzierung der Skaleninhalte einerseits an dem Instrument, andererseits aber auch an einer noch notwendigen weiteren Ausdifferenzierung der klinischen Modelle über Dimensionen und Veränderungen im Psychotherapieprozess liegen kann.

5.2.3. Messwiederholungen

Neben der Forderung von drei Dimensionen macht das für den FEP maßgebliche Phasenmodell (Howard et al., 1993) auch eine klare Aussage darüber über die sequentielle Abnahme der Belastungswerte auf den drei Dimensionen Wohlbefinden, Symptomatische Belastung, und psychosoziales Funktionieren. Die externe Validität des Phasenmodells könnte mit Untersuchungen gestützt werden, die zeigen, a) dass, selbst wenn die Items nahe an eindimensional sind, sich die Belastungen für die drei Einzeldimensionen in der vorhergesagten Weise verändern, oder b) dass die Belastungsgrade ähnlich bleiben, sich die Itemparameter aber systematisch für die drei Dimensionen in ihrer Schwierigkeit ändern. Befund a) wäre trotz der bislang weitgehend bestätigten geringeren Dimensionalität der relevanten Fragebögen möglich, da diese lediglich ein Zustandsbild der erhobenen Personen abbildeten, nicht jedoch die Veränderung über die Zeit, die trotzdem gestaffelt erfolgen könnte. Hier lassen sich die Grundlagen der IRT-Modelle in inhaltliche Hypothesen überführen, die getestet werden können. Erste Untersuchungen mit IRT-Bifaktor-Modellen (Reise et al., 2010, 2007) mit Daten des COMPASS-Systems (Grissom et al., 2002; Sperry, Brill, Howard, & Grissom, 1996) sowie dem FEP zeigten allerdings, dass nach Extraktion einer gemeinsamen Hauptdimension kaum Information für die Subdimensionen übrigblieb und die Effektstärken auf den Subdimensionen sehr klein waren (Böhnke & Lutz, 2011a).

Je nach Verwendungszweck müsste zur Nutzung des FEP und seiner Kurzversionen in der Veränderungsmessung der Nachweis erfolgen, dass der Fragebogen über die Erhebungszeiten hinweg in seiner Struktur invariant ist. Erweiterungen des Modells ermöglichen prinzipiell die An-

wendung wiederholter Messdaten (Blanchin et al., 2011; Glück & Spiel, 1997, 2007; Hatzinger & Rusch, 2009; Meiser, Stern, & Langeheine, 1998) und auch eine Anwendung für die Verfolgung der Veränderung von Einzelfalldaten sind möglich (Van Rijn & Molenaar, 2005). In dieser Arbeit lag der Schwerpunkt auf der Evaluation der Anwendbarkeit dieser Modelle und einer verstärkt nomothetischen, populationsbasierten Perspektive: Wie verändern sich Fälle vor dem Hintergrund zweier definierter Populationen (Patienten und Nicht-Patienten), die die Interpretation der Testwerte ermöglichen (Lambert & Ogles, 2009). Die Möglichkeiten, die Testwerte auch relativ zu anderen Patienten oder aber intraindividuell im dynamischen Prozess zu vergleichen (Bergmann-Warnecke, 2011; Molenaar & Campbell, 2009; Schmitz, 2000; van Rijn & Molenaar, 2005) stellen weitere Forschungsfragen dar, die über den Rahmen der Arbeit hinausgehen.

5.2.4. Auswahl des IRT-Modells

In Bezug auf die Modellwahl können zwei Aspekte kritisiert werden. Zum Einen die Wahl eines eindimensionalen Modells zur Beschreibung von drei Aspekten des Fragebogens (Wohlbefinden, Symptome, Inkongruenz). Zum Anderen die Wahl des spezifischen IRT-Modells (Partial Credit Modell). Zur Wahl der Zahl der Dimensionen wurde bereits dargelegt, dass sich dieser Befund in die derzeitige Literaturlage einfügt. In der vorliegenden Arbeit wurde das Partial Credit Modell verwendet, da es für die anvisierte Anwendung deutliche Vorteile bot (Doucette & Wolf, 2009). Dennoch ist es als Mitglied der Rasch-Familie ein sehr restriktives IRT-Modell (Rost, 2001). Insbesondere vor dem Hintergrund, dass flexiblere Modelle vielleicht weniger anfällig für Verletzungen der Eindimensionalität wären (Drasgow & Parsons, 1983; Reise et al., 2010), stellt sich also die Frage, ob die Wahl eines anderen Modells nicht angemessener gewesen wäre. Reise und Waller (2003) untersuchten dichotome Minnesota Multiphasic Personality Inventory-Daten (Butcher et al., 1992) an nicht-klinischen Stichproben Heranwachsender und kamen zu dem Schluss, dass das 2PL-/ Birnbaum-Modell eine bessere Passung zeigte als das 1PL-/Rasch-Modell. Dumenci und Achenbach (2008) zeigten, dass die Unterschiede in den Ergebnissen für die Schätzung der Personenparameter allerdings eher gering sind.

Zur Illustration der wesentlichen Unterschiede in der vorliegenden Untersuchung sind in Abbildung 5-1a die Personenparameter-Schätzungen für zwei Modelle abgetragen, das Generalized

Partial Credit Modell (Entsprechung des weniger restriktiven 2PL-Modells für dichotome Daten; Muraki, 1992) und das in der Arbeit verwendete Partial Credit Model (Masters, 1982). Auf der x-Achse sind die aus dem Partial Credit-Modell resultierenden Schätzer abgetragen, auf der y-Achse die aus dem Generalized Partial Credit Modell, beide mittels Empirical Bayes Schätzern geschätzt. Die resultierenden Schätzer für die Personen sind nahezu gleich, was auch durch die hoher Pearson-Korrelation unterstrichen wird. Wie aber der untere Teil der Abbildung 5-1 zeigt, verändert sich die Schätzgenauigkeit. Hier ist links die Informationsfunktion für die Partial Credit Ergebnisse und rechts die für die Generalized Partial Credit Ergebnisse abgebildet. Insgesamt verändert dies die Interpretation der Ergebnisse nicht substantiell, was die Spannweite des gut gemessenen Bereiches auf der latenten Dimension angeht. Dennoch wird deutlich, dass die Messgenauigkeit bei der Auswertung mittels des Generalized Partial Credit Modells insgesamt höher ist und besonders im mittleren Teil im Vergleich noch deutlicher ansteigt.

Für die vorliegende Anwendung sprechen die Auswertbarkeit mittels des Summenscores und die vorteilhaften Messeigenschaften des Rasch-Modells (Doucette & Wolf, 2009) für die Verwendung des Partial Credit Modells. Die Abweichungen von der Annahme der Eindimensionalität (Faktorenanalyse) und der Skalierbarkeit unter dem Modell (Fit-Statistiken; Böhnke & Lutz, submitted) wurden als vernachlässigbar eingestuft. Und der obige Vergleich mit dem Alternativmodell zeigt ebenfalls, dass die resultierenden Nachteile eher gering sind. Dies ist auch eine wichtige Grundlage für Studie III in der von Vorneherein ein eindimensionales Modell zur Auswertung der Belastungsitems des FEP verwendet wurde (s. Kapitel 1.5.4): Die vorliegenden Untersuchungen belegen, dass lediglich ein geringer Informationsverlust durch dieses Vorgehen entsteht.

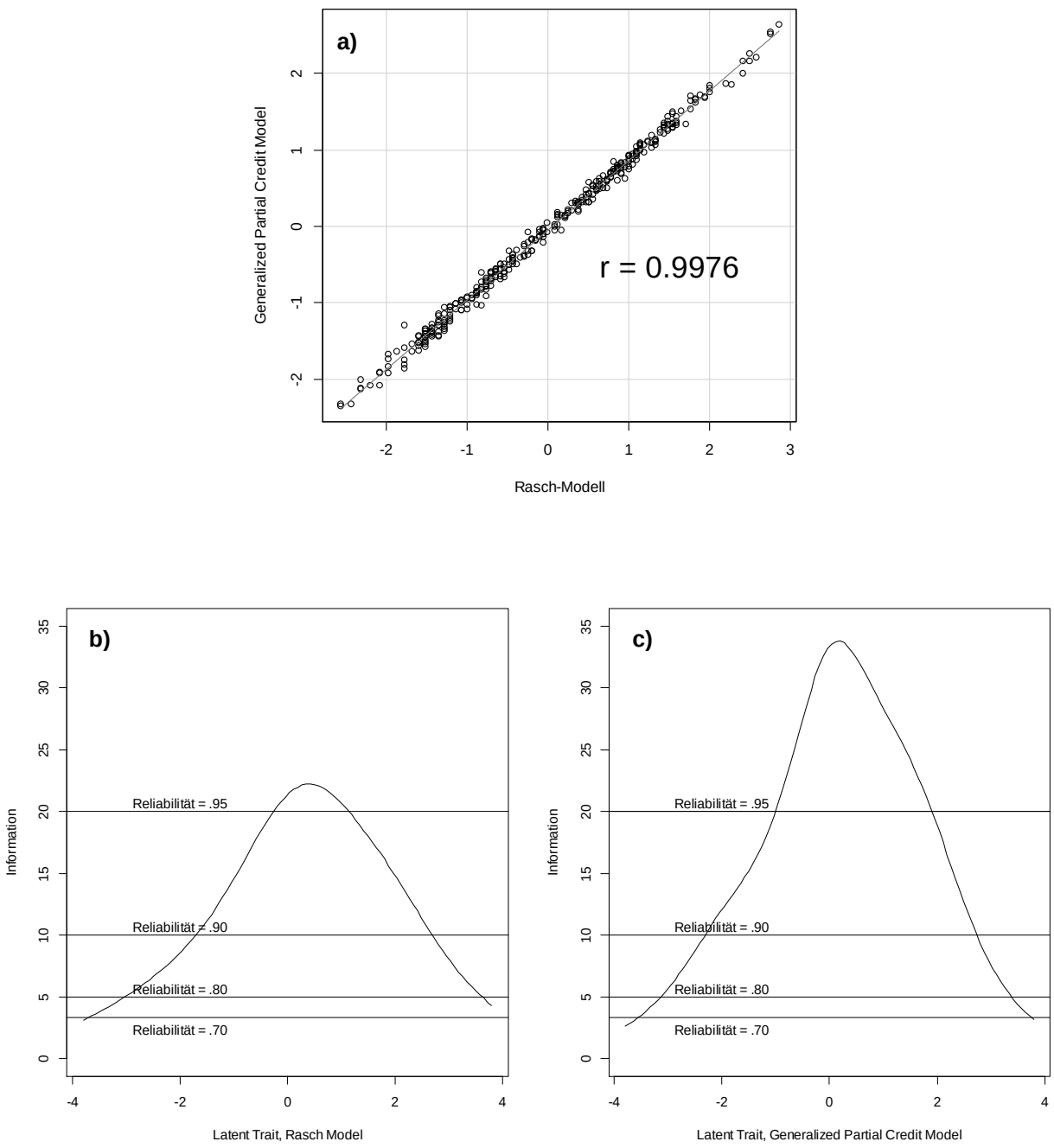

Abbildung 5-1: Mitte oben (a) enthält das Scatterplot für die Personenparameter geschätzt mit dem PCM und dem Generalized PCM; unten links (b) zeigt die Informationsfunktion für die Gesamtskala nach Schätzung des PCM; unten rechts (c) die Informationsfunktion der Gesamtskala nach Schätzung des Generalized PCM.

5.2.5. Angemessenheit von IRT Modellen für die Messung psychischer Belastung

Eine letzte zu diskutierende Frage ist, ob IRT-Modelle überhaupt eine angemessene Repräsentation psychischer Belastung leisten können. Diese Frage gliedert sich in zwei Aspekte: Zum Einen die Verwendung und Bedeutung des statistischen Modells, zum Anderen die Möglichkeit, dass

"psychische Belastung" ein sog. "Quasi-Trait" ist (Reise & Waller, 2009). Diese beiden Aspekte sollen im Folgenden nacheinander behandelt werden.

Für den ersten Aspekt ist es wichtig, sich auf der klinisch-theoretischen Modellebene klar zu werden, ob latente Variablen die Symptome (Items in den Fragebögen) verursachen oder ob diese Symptome eigentlich Indikatoren sind, die gemeinsam ein latentes Konstrukt "Belastung" verursachen. Cohen und Kollegen (Cohen, Cohen, Teresi, Marchi, & Velez, 1990) vertreten diesen Punkt schon früh in der Debatte um latente Variablen (s.a. Bollen & Lennox, 1991; Borsboom et al., 2003; J. R. Edwards, 2011; Reise & Henson, 2003; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Sie unterscheiden zwischen drei Typen sog. "operativer Variablen" in Strukturgleichungsmodellen. *Operative Variablen* sind ein zusammenfassender Begriff für alle Variablen, die in Strukturgleichungsmodellen verwendet werden. Dies können zum Einen manifeste Variablen sein. Demgegenüber unterscheiden sie auf der Seite der nicht gemessenen Variablen die Typen latente Variablen und emergente Variablen. *Latente Variablen* werden als ein verursachender Faktor für die Ausprägungen einer Reihe von Indikatoren gesehen, was in solchen Modellen (d.h. auch bei IRT-Modellen, s. Kapitel 1.5.3) darüber rekonstruiert wird, dass die Interkorrelationen zwischen Items verschwinden, wenn die latente Variable konstant gehalten wird (lokal stochastische Unabhängigkeit). Die verursachende latente Variable kann bei solchen Modellen also durch die Interkorrelationen (nahezu) vollständig wieder identifiziert werden. Bei *Emergenten Variablen* dagegen zeigt der kausale Pfeil in die andere Richtung, d.h. auch sie hängen kausal mit den Indikatoren zusammen, aber in diesem Fall wird die emergente Variable durch die Indikatoren verursacht.

Cohen (P. Cohen et al., 1990) erläutern dies an dem Beispiel, dass die Variablen a) Anzahl Kinder in einer Familie, b) eine Erkrankung der Mutter und c) Anzahl der Stunden, die eine Mutter Arbeiten geht, alles valide Indikatoren dafür sind, welche Möglichkeiten die Mutter hat, mit einem Kind zu interagieren. Aber diese drei Indikatoren werden *nicht* durch die latente Variable "Möglichkeit mit einem Kind zu interagieren" verursacht. Hier wird auch deutlich, dass in solch einem Fall die Indikatoren nicht einmal miteinander korreliert sein müssen. Ein zweites Beispiel ist "schlechte Gesundheit" ("ill health", Cohen et al., 1990: 186), das oft über die Anzahl und Schwere verschiedener Erkrankungen (koronare Herzerkrankung, Diabetes, o.ä.) genmessen wird, und so

keine latente Variable darstellen kann: Auch hier ist nicht anzunehmen, dass die Kovarianz vieler Erkrankungen in einer Population durch einen latenten Faktor erklärt werden kann. Dies ist schon vor dem Hintergrund unterschiedlicher physiologischer Genesen der Erkrankungen unplausibel. Stattdessen schlagen die Autoren zur Messung von Belastungen durch Erkrankungen vor, die funktionalen Konsequenzen (Stärke von Schmerzen, Müdigkeit, Einschränkungen im Alltag,...) zu erfassen (P. Cohen et al., 1990), eine Perspektive, der sich die Forschung zur störungs- und erkrankungsübergreifenden Erfassung von Lebensqualität angeschlossen hat (Cella, Yount, et al., 2007; Fayers & Machin, 2007; Huppert & Whittington, 2003; Lai, Crane, & Cella, 2006; Rose, Bjorner, Becker, Fries, & Ware, 2008; Willke et al., 2004; Wood-Dauphinee, 1999).

Solche Überlegungen haben direkten Bezug zur Konstruktion von Messinstrumenten für psychische Belastung. IRT Modelle nehmen die Existenz einer latenten Variable an, die die Antworten auf die Items verursacht (Fayers & Machin, 2007). Ein solches Erhebungsinstrument umfasst als Items eine Sammlung an Einschätzungen von symptomatischen Aspekten, Wohlbefinden, etc. von denen angenommen wird, dass diese als belastender bewertet werden, je höher die psychische Belastung (latente Variable) der Patienten ist (Abbildung 5-2a). Wird der allgemeine Belastungsgrad aus den Itemantworten herausgerechnet, besteht keine Korrelation mehr zwischen den Itemantworten (Kempf, 2010; Moosbrugger & Kelava, 2007). Die psychische Belastung variiert wiederum abhängig von biologischen, psycho-sozialen und weiteren Aspekten. Direkte Verbindungen zwischen diesen Ursachen psychischer Belastung und den einzelnen Items sollten eigentlich nicht vorhanden sein: Dies würde sonst Multidimensionalität der Items implizieren, da Variabilität durch die latente Dimension unaufgeklärt bleibt, die noch mit externen Kriterien korreliert ist (Jöreskog & Goldberger, 1975; Woods, 2009).

Dies ist anders für die emergenten Variablen. Auch hier könnte in einem Fragebogen eine Reihe von symptomatischen Belastungen, Wohlbefinden, etc. abgefragt werden, doch würde bei diesen nicht davon ausgegangen, dass sie durch eine gemeinsame Variable verursacht werden (Abbildung 5-2b). Höhere Werte in diesen Items würden zwar anzeigen, dass die emergente Variable Belastung eine ebenfalls höhere Ausprägung hat, doch würde dies nicht bedeuten, dass eine stärker belastete Person auch automatisch höhere Werte in den anderen Items haben müsste. Da

nicht die Varianz in den Items durch die latente Variable erklärt wird, wäre die Beziehung zu Außenkriterien weniger klar: So wäre schwer unterscheidbar, ob psycho-soziale Faktoren, biologische Faktoren etc. die Items beeinflussen oder aber ebenfalls nur einfach weitere Indikatoren für die Belastung sind (in Abbildung 5-2b für das erste Item verdeutlicht).

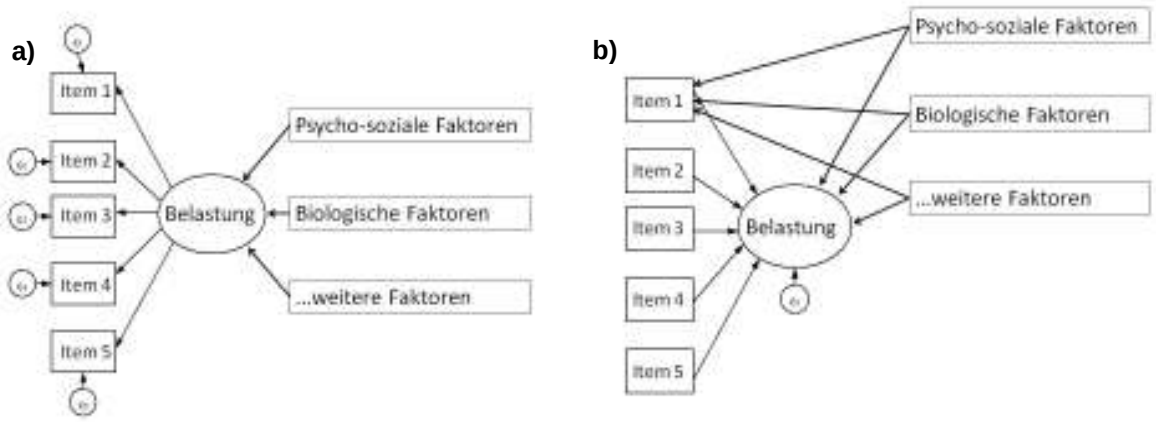

Abbildung 5-2: Schematische Darstellung von (a) einem Modell mit latenter Belastungsvariable als kausaler Faktor der Items; (b) einem Modell mit emergenter Belastungsvariable als Ergebnis der Messung mit einer Reihe von Indikatoren.

Derzeit werden Faktorenanalysen (formative/ latente Variablen) und Hauptkomponentenanalysen (emergente Variablen) zur Untersuchung von Skalenstrukturen verwendet. Eine empirische Entscheidung über die Angemessenheit des einen oder des anderen Vorgehens ist kaum möglich, sondern hier müssen in jedem Fall auch theoretische Überlegungen des Forschungsfeldes herangezogen werden (J. R. Edwards, 2011; Fayers & Machin, 2007). Die Relevanz dieser Frage kann an einem Grenzfall der derzeitigen klinischen Forschung zur Psychopathologie verdeutlicht werden. IRT Modelle werden in letzter Zeit immer wieder herangezogen, um zu Prüfen, ob das Spektrum bestimmter Diagnosen als eindimensional aufgefasst werden kann (Goldberg & Goodyer, 2005; Krueger & Finger, 2001; McGlinchey & Zimmerman, 2007). Krueger und Finger (2001) benutzten in ihrer Untersuchung die 354 Personen, die im National Comorbidity Survey (Kessler, McGonagle, Zhao, Nelson, Hughes, Eshleman, et al., 1994) die Screening-Frage, ob sie sich derzeit wegen psychischer Probleme in Behandlung befänden, bejahten. Für diese Personen untersuchten die Autoren mittels eines IRT Modells, ob die vergebenen Diagnosen sieben psychische Störungen

(dichotom kodiert; Majore Depression, Spezifische Phobie, Agoraphobie, Sozialphobie, Panikstörung, Generalisierte Angststörung, Dysthymie; Krueger & Finger, 2001: 143) eindimensional sind.

Dies erscheint dem Erkrankungsbeispiel von Cohen und Kollegen oben ähnlich (P. Cohen et al., 1990): Werden Symptome psychischer Erkrankungen als durch einen latenten Faktor "psychische Belastung" verursacht angesehen, dann wäre dieses Modell also nicht zulässig – Störungen würden Einfluss auf die Schwere der Belastung nehmen, aber ihre gemeinsame Kovarianz würde nicht durch einen latenten Faktor erklärt (Abbildung 5-2a). Ausgehend von einem solchen theoretischen Modell wäre es also angemessener eine emergente Variable anzunehmen (Abbildung 5-2b). Krueger und Finger (2001) gehen allerdings von einer anderen theoretischen Annahme aus. Sie nehmen die Existenz einer latenten Variable mit den Polen "Internalisierungs- vs. Externalisierungsstörungen" an, die das Auftreten der Störungen kausal beeinflusst. Dieses klinische Modell rechtfertigt die Verwendung einer latenten Variable – doch muss in diesem Modell nun geklärt werden, wie die Symptome, das Wohlbefinden etc. von der latenten Dimension "Internalisierungs- vs. Externalisierungsstörungen" abhängt. Vor diesem Hintergrund wird klar, dass die Angemessenheit von IRT Modellen für die Messung psychischer Belastung auch von den Annahmen der Klinischen Psychologie und Psychopathologie abhängt, welchen Status Symptome, Syndrome und Störungen haben. Psychometrische Ansätze können bei der Untersuchung der empirischen Gegebenheiten helfen (Keller & Kempf, 1997; Lubke & Neale, 2008; Meehl, 1954; Ruscio, Brown, & Meron Ruscio, 2009), aber dennoch wird deutlich, dass hier auch die inhaltliche Theoriebildung noch weiter fortschreiten muss.

Der zweite angesprochene Aspekt, der bei der Angemessenheit von IRT-Modellen zur Messung psychischer Belastung berücksichtigt werden muss, ist das Vorliegen der Eigenschaften eines "Quasi-Trait" (Reise & Waller, 2009): Oft wird bei der Erstellung von Skalen zur Erhebung psychischer Belastungen festgestellt, dass sie lediglich den Belastungsbereich gut erfassen, den Bereich der Abwesenheit von Belastung allerdings eher schlecht. Eine Erklärung dafür ist, dass dies ein Problem der Itemkonstruktion ist, d.h. zu vermuten, dass für den unteren Belastungsbereich noch Items entwickelt werden müssten (s. Kapitel 3 zur Kurzform hohe Belastung; Doucette & Wolf, 2009; Sharp, Goodyer, & Croudace, 2006). Da dieser Befund aber oft repliziert wird, kommt

eine andere Hypothese mehr in den Vordergrund: Statt fehlender Items könnte es an der Natur der gemessenen latenten Eigenschaft liegen. Reise und Waller (2009: S. 31) bezeichnen solche Traits als "Quasi-Traits", bei denen nicht beide Pole der Dimension empirisch (und ggf. theoretisch) vorhanden sind, sondern nur der eine. Im Fall von psychischer Belastung hieße dies, dass es am oberen Ende immer mehr psychische Belastung gibt, die stünde aber nur "keiner Belastung" gegenüber, nicht einem inhaltlich definierten Gegenpol. Es gibt derzeit keine Hinweise, dass die Verwendung von IRT-Modellen unter diesen Umständen unangebracht wäre (Doucette & Wolf, 2009; Reise & Waller, 2009). Aber nach Reise und Waller (2009) ist nicht nur die Variation auf diesem Trait aus theoretischen Gründen am unteren Ende weniger informativ. Es gäbe auch kaum eine Möglichkeit, diese zu messen, da es faktisch keine Indikatoren für das untere Ende gibt. Aus ihrer Sicht wären Versuche zur Entwicklung solcher Items nur Zeitverschwendung.

Ob diese Position haltbar ist, ist sowohl eine theoretische wie eine empirische Frage. Sicher trifft auch hier zu, dass die Klinische Theoriebildung mehr Arbeit leisten müsste, Indikatoren des Pols "Unbelastetheit" zu identifizieren. Aus klinischer Sicht ist allerdings nicht abzusprechen, dass die Items am unteren Ende des Belastungsbereiches inhaltlich sehr interessant wären. Krause und Lutz (2009) verdeutlichen, dass gerade positive Baselines, also Entwicklungsziele für Patienten, in der Untersuchung der Wirksamkeit klinischer Interventionen eine hohe Bedeutung haben. Diese müssten sich von der reinen Messung psychischer Belastung unterscheiden. Intuitiv wäre zu vermuten, dass gerade die Messung von Lebensqualität das Problem von "Quasi-Trait"-Messungen beheben müsste, doch stellt sich nach bisherigen empirischen Befunden dieselbe Frage (S. Choi, Reise, Pilkonis, Hays, & Cella, 2010; Fayers & Machin, 2007). Die Ergebnisse zum FEP zeigen ebenfalls, dass die Wohlbefindens-Items nicht systematisch am positiven unteren Ende der Skala liegen. Daher stellt sich damit die Frage, ob diese Baselines nicht einen Übergang zu anderen Ergebnis-Kriterien nötig machen: Nach Betrachtung der Symptomreduktion und einer Abnahme negativer Aspekte des Wohlbefindens ist vielleicht ein Schwenk auf andere Konstrukte wie Kongruenz (statt Inkongruenz; Grawe, 1998, 2004), Selbstwirksamkeit (Schwarzer, 1994; Überblick über Befunde: Ruholl, 2007; Jung, 2008) und Konstrukte aus der Positiven Psychologie (Lee Duckworth, Steen & Seligman, 2005; Seligman & Csikszentmihalyi, 2000) nötig, die zwar nur

bedingt relevant für die Entwicklung in akuten Belastungsphasen sind (Fokus aus Dimensionen von Schulte oder dem Phasenmodell, s. Kapitel 1.4.3), aber danach eine Beschreibung des Gesundungsprozesses ermöglichen.

5.2.6. Ziel und Zweck von Kurzskalen

Am Schluss der Diskussion steht die Frage, was Kurzskalen an sich leisten können. Wie beschrieben sprechen verschiedene Gründe für die Entwicklung von Kurzskalen. Aus praktischer Sicht sprechen die Erwägungen der Länge der Erhebungen für eine Kürzung der Skalen, da schon wenige Minuten Zeitersparnis über die Möglichkeit der Umsetzung entscheiden können (Locke et al., 2012). Und die Ergebnisse der Feedbackforschung verdeutlichen, dass eine Umsetzung von Erhebungen im Praxiskontext wünschenswert ist (Carlier et al., 2012; Shimokawa et al., 2010). Außerdem müsste nachgewiesen werden, dass viele Items gegenüber kürzeren Fassungen auch inkrementelle Validität besäßen (Hunsley & Mash, 2005; Meyer et al., 2001), sonst wären kürzere Erhebungen zu bevorzugen.

Zwei Aspekte fallen bei der Bewertung der Möglichkeiten von Kurzskalen ins Auge, die gegen eine Verwendung solcher Instrumente sprechen könnten: Eine Reduktion der Itemanzahl führt notwendigerweise zu einer Verringerung der Reliabilität der Messung und es besteht die Gefahr, dass die Auslassung von Items dazu führt, dass die Skala an Inhaltsvalidität verliert (Brod et al., 2009). Verschiedene Strategien können dazu eingesetzt werden, diese Gefahren zu verringern. Der Verlust an Reliabilität kann dadurch begrenzt werden, dass wie in Studie II eine Mindestreliabilität für die Kurzfassung festgelegt wird. In Studie III wurde, wie in einem Review zur Erstellung von Kurzskalen gefordert (G. T. Smith et al., 2000), die erreichte Reliabilität mit der nach Spearman-Brown heruntergerechneten Reliabilität verglichen (Kempf, 2003, 2008): Die erreichten Reliabilitäten der Kurzfassungen lagen sogar oberhalb der erwarteten Werte (Tabelle 4-3). Damit kann das Problem des Reliabilitätsverlustes eingegrenzt werden. Die vorliegende Arbeit zeigt aber auch, dass eine zu deutliche Reduzierung der Items auf wenige Items, u.U. sogar nur eines (Cuijpers et al., 2009; Hart et al., 2012; Yamazaki et al., 2005) zu optimistisch ist und hinter den Ansprüchen, die an reliable Messungen gestellt werden, zurückbleibt (Böhnke & Lutz, submitted).

Die Frage, ob die Inhaltsvalidität erhalten bleibt, kann nicht allgemein beantwortet werden. Wenn ein IRT-Modell angepasst wurde, ist damit belegt, dass alle Items dieselbe Dimension erfassen und damit alle Items auch adäquate Repräsentanten der latenten Dimension sind. Dennoch ist die Auswahl von Items aus einer "Item Bank", wie eine Vollversion sie darstellt, immer eine Gradwanderung zwischen dem Argument, dass jedes beliebige Subset aus der Bank genommen werden kann (Cella, Gershon, et al., 2007) und der Einschränkung dass jedes Auslassen von Items zu schmaleren Definitionen des Konstrukts führt (Hays et al., 2000). In beiden Studien (II + III) findet sich ein Übergewicht der depressiven Symptomatik in den ausgewählten Subskalen. Dies liegt daran, dass in den Studien der Konstruktvalidität deutlicher Vorrang gegenüber der Inhaltsvalidität gegeben wurde. Eine Anpassung der Vorgehensweisen ist dahingehend möglich, dass dieselbe Anzahl von den verschiedenen Facetten des Fragebogens ausgewählt wird. Dies hätte in der Regel eine Verlängerung der Fragebögen zur Folge (wenn dieselbe Reliabilität erreicht werden soll), wäre aber eine Möglichkeit, der Breite des Skaleninhalts stärker Rechnung zu tragen. Außerdem muss aus der Praxis betont werden, dass die Kurzfassungen nur ein Teil eines Erhebungssystems sein können und erstens Erhebungen mit unterschiedlichen, aber äquivalenten Kurzfassungen gemischt werden könnten (Forkmann et al., 2012), und zweitens Erhebungspläne Kurz- und Vollversionen der Skalen miteinander abwechseln, um so ein breites Bild der Veränderung zu erhalten (Grosse Holtforth et al., 2010; Lutz, Mocanu, et al., 2010). Dies trägt dem Faktum Rechnung, dass eine Kurzversion eines guten, aber langen Instrumentes grundsätzlich nicht dieselbe Güte wie das Original erreichen kann.

Die Frage der Inhaltsvalidität wird mittlerweile in ähnlicher Form auch an computer-adaptive Testsysteme gestellt (Brod et al., 2009; Cella, Gershon, et al., 2007). Im Praxiskontext kommt zusätzlich die Frage der Augenscheinvalidität hinzu, da es intuitiv schwer verständlich sein kann, warum zwei Messungen, die vielleicht nicht ein einziges Item gemeinsam haben, vergleichbar sein sollen (Höhler, Hartig, & Ullrich, 2012). Ein statistisches Prinzip wie die lokal-stochastische Unabhängigkeit reicht nicht unbedingt, um inhaltliche Bedenken auszuräumen. Eine Frage ist z.B. ob für den Kliniker als relevant angesehene Items im Prozess erhoben wurden oder nicht. Während diese Fragen für computer-adaptive Tests durch Restriktionen in der Itemauswahl gelöst werden

könnten (z.B. Sets an Items, die immer erhoben werden; Sets an Items, die bei der nächsten Messung wieder erhoben werden sollen; usf.), führt dies für statische Testformen zu einem Konsensprodukt aus inhaltlichen Erwägungen der Anwender und empirischen Ergebnissen der Testkonstruktion.

Bezogen auf die Zeitersparnis der Kurzversionen (G. T. Smith et al., 2000) kann hinterfragt werden, als wie groß die höhere Belastung durch längere Erhebungen tatsächlich ist. Die allgemeinere Frage des erhöhten Aufwandes wird immer wieder von Therapeuten gegen den Gebrauch von Instrumenten eingewendet (Gilbody et al., 2002b; Hagemeister et al., 2010; D. Hatfield & Ogles, 2007). Nach vorliegenden Arbeiten muss festgehalten werden, dass die Patienten solchen Erhebungen aber eher positiv gegenüberstehen (Lutz, Böhnke, Köck, et al., 2011; Steffanowski et al., 2011). Die Argumentation, dass die Patienten es in der Breite nicht wollen, muss entsprechend der Studien aus der ambulanten Versorgung entgegengehalten werden, dass die Erhebungen durchaus als machbar und auch als sinnvoll erlebt werden. Oft führen also nur Plausibilitätserwägungen dazu, dass bei besonders hervorgehobenen Patientengruppen angenommen wird, dass sie durch solche Erhebungen zu stark belastet werden, wie z.B. im Extremfall in der Palliativmedizin (Walker et al., 2010).

Darüber hinaus muss festgestellt werden, dass es eigentlich keine empirische Evidenz dazu gibt, ab wann Erhebungen zu lang für Patienten werden. Rücklaufraten des CORE in der Routineversorgung in Großbritannien zeigen, dass es vor allem eine Frage der Einbettung der Erhebungen in den systemischen Kontext eines Behandlungszetrums ist, der über die Höhe Rücklaufraten entscheidet (Bewick, Trusler, Mullin, Grant, & Mothersole, 2006). In der TK-Studie mit ihrem wenig eng gestrickten Messplan (Lutz, Böhnke, & Köck, 2011; Wittmann et al., 2011) waren über 91% der Patienten der Auffassung, dass Qualitätsmonitoringprojekte gut seien und über 95% fanden den Aufwand für die Erhebungen vertretbar (Lutz, Böhnke, Köck, et al., 2011: Tabelle 4). In der PSY-BAY-Studie schätzten 74.5% der Befragten mit abgeschlossenen Therapien den zusätzlichen Aufwand durch die Erhebungen als "gering" ein; 75.3% fanden den regelmäßigen Einsatz der Fragebögen und Rückmeldungen sinnvoll und 63.2% würden das in der Studie umgesetzte Verfahren als Standard in der ambulanten Versorgung empfehlen (Steffanowski et al., 2011: S. 280). Spätestens

vor diesem Hintergrund sollte eine grundsätzlich abwehrende Haltung aufgrund des Patientenschutzes hinterfragt werden.

5.3. Fazit & Ausblick

Die vorliegende Arbeit unterstreicht, dass es möglich ist, zu praxisorientierten Verbesserungen in den Methoden der Patientenorientierten Versorgungsforschung zu kommen. Eine sowohl inhaltlich wie psychometrisch gesteuerte Verkürzung der für Monitoringzwecke eingesetzten Instrumente ermöglicht kürzere Erhebungen und sichert (im Fall der IRT Modelle) die Vergleichbarkeit der Messungen innerhalb einer Institution wie auch mit der Forschungsliteratur. Die Verfügbarkeit von valider open source Software ermöglicht es darüber hinaus mit einfachen Mitteln auf die jeweiligen Praxissettings angepasste Lösungen zu entwickeln. Dies ermöglicht eine engere Verzahnung von Wissenschaft und Forschung und damit mittelfristig auch eine bessere Versorgung der Patienten.

Durch die Verwendung von IRT-Modellen werden die Daten auch grundsätzlich angemessener ausgewertet (Dumenci & Achenbach, 2008). Mit Fragebogeninstrumenten erhobene Daten sind in den meisten Fällen kategorialer Natur (Kempf, 2010; Sharp et al., 2006; Wirth & Edwards, 2007) und es gibt keine guten Gründe (mehr), sie nicht auch so auszuwerten. Neben der in Studie I verwendeten reinen IRT-Strategie, können durch die Verwendung mehrerer Methoden zur Untersuchung kategorialer Daten ergänzende Informationen über das verwendete Instrument gewonnen werden. In der zweiten Studie (siehe auch Böhnke & Lutz, submitted) wurde zur Untersuchung der Dimensionalität der empirischen Daten eine Faktorenanalyse basierend auf polychoren Korrelationen (D. B. Flora & Curran, 2004; McDonald, 1999) verwendet. Während typische faktorenanalytische Vorgehensweisen auf Kovarianzen oder Pearson-Korrelationen beruhen, die lediglich bei linearen Verbindungen intervallskalierter und idealerweise normalverteilter Variablen angemessene Informationen über die Beziehung zwischen zwei Variablen geben, ermöglichen polychore Korrelationen es, unter der Beibehaltung der Annahme der Normalverteilung die Annahmen der Linearität und Intervallskalierung aufzugeben, und so auch ohne eine explizite IRT-Strategie Zusammenhänge zwischen ordinalen Daten zu untersuchen (Wirth & Edwards, 2007; zum Zusammenhang zwischen IRT-Parametern und polychoren Korrelationen: McDonald, 1999).

Praktische Anwendungsmöglichkeiten der Ergebnisse sind in den Kapiteln 3 und 4 bereits erläutert worden. Bezogen auf die weitere Entwicklung der Patientenorientierten Versorgungsforschung könnte mit kürzeren Erhebungen z.B. leichter in Wissenschaftler-Praktiker-Netzwerken beforscht werden, wie stark der Einfluss von Rückmeldungen auf Therapieergebnisse und –prozessmerkmale wirklich ist, da sich die Erhebungen leichter umsetzen lassen. Wie in Kapitel 1.4.2 bereits dargestellt, ist die Evidenz zu dieser Frage weit weniger gesichert und die qualitativ hochwertigsten Studien stammen aus einer Arbeitsgruppe (Shimokawa et al., 2010). Hier besteht weiterer Forschungsbedarf, der die Evidenzbasierung der Psychotherapie verbessern kann.

Kazdin und Blase (2011) machen in ihrem Überblick die Vorhersage, dass sich computergestützte Anwendungen in der Psychotherapie (und damit auch in der Forschung) verbreiten werden. Ihr Einsatz in der Versorgung ist sicher noch nicht ausgereizt (Chang, 2007). Damit werden vielleicht in der Zukunft statische Fragebogenformen weniger notwendig, aber es ergäben sich neue Forschungsbereiche in der Patientenorientierten Versorgungsforschung, in denen IRT-Modelle eingesetzt werden könnten. In Kapitel 1.4.1 wurden Studien und Ergebnisse zur Vorhersage von Psychotherapieverläufen berichtet. Sollten computergestützte Begleitungen von Therapieverläufen in der ambulanten Psychotherapie Standard werden (z.B. sensu Lutz et al., 1999; Lutz, Leach, et al., 2005), dann ergäbe sich hier die Möglichkeit computer-adaptive Tests zu integrieren. So könnten die Testerhebungen mit computer-adaptiven Systemen dadurch effektiver werden, indem die vorhergesagten Verlaufswerte als Startwerte für die Belastung bei der jeweiligen Testadministration des spezifischen Patienten genutzt werden. So wäre beispielsweise der Bereich zwischen dem letzten Messwert und dem aktuell vorhergesagten Messwert ein inhaltlich gut begründeter Schätzer für das Belastungsniveau, welches für die Auswahl von Items bei einer anstehenden Folgemessung verwendet wird. Die vielen zur Verfügung stehenden Methoden zur Bewertung der Güte einer Messung im IRT-Modell (Raîche et al., 2007) könnten weiterhin eine Möglichkeit sein, plötzliche Veränderungen in Prozess- wie Ergebnismaßen direkt festzustellen und ebenfalls als Rückmeldung an die Therapeuten bereitzustellen (Cahill et al., 2011; Coutinho, Ribeiro, Hill, & Safran, 2011; Lutz et al., in press; Tang & DeRubeis, 1999b; Tschitsaz-Stucki & Lutz, 2009). Umgekehrt könnten IRT Modelle für eine bessere Messqualität der jeweiligen Konstrukte sorgen und so zur Ver-

besserung der Vorhersagen beitragen. Ein erstes Ergebnis in dieser Richtung zeigte, dass Personen, deren individueller Messfehler aus einem IRT Modell zu Beginn der Therapie höher war, auch im Verlauf schlechter vorhersagbar waren (Köhler, 2012).

Diese Ergebnisse der Arbeit zusammengenommen mit den weiterführenden Überlegungen zeigen, dass sich die Befunde dieser Arbeit gut in die Praxis und Forschung verschiedener Bereiche der Patientenorientierten Versorgungsforschung integrieren lassen.

6. Referenzen

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch model and the disordered threshold controversy. *Educational and Psychological Measurement*, online first.

Ader, D. N. (2007). Developing the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*, S1–S2.

Ahn, H. -n., & Wampold, B. E. (2001). Where oh where are the specific ingredients? A meta-analysis of component studies in counseling and psychotherapy. *Journal of Counseling Psychology*, *48*, 251–257.

Albani, C., Blaser, G., Geyer, M., Schmutzer, G., & Brähler, E. (2010). Ambulante Psychotherapie in Deutschland aus Sicht der Patienten. *Psychotherapeut*, *55*, 503–514.

Albani, C., Blaser, G., Geyer, M., Schmutzer, G., & Brähler, E. (2011). Ambulante Psychotherapie in Deutschland aus Sicht der Patienten. *Psychotherapeut*, *56*, 51–60.

Allaire, J. J., Cheng, J., Paulson, J., & DiCristina, P. (n.d.). RStudio. Retrieved October 29, 2011, from http://www.rstudio.org/

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060–1073.

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*, 123–140.

Andersen, E. B. (1997). The rating scale model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 67–84). New York: Springer.

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, *77*, 693–704.

APA Presidential Task Force on Evidence-Based Practice. (2006). APA Presidential Task Force on Evidence-Based Practice. *American Psychologist*, *61*, 271–285.

Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Presented at the 2009 GMAC (R) Conference on CAT. Retrieved December 13, 2012, from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09babcock.pdf

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose–effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology*, *77*, 203–211.

Barkham, M., Connell, J., Stiles, W. B., Miles, J. N. V., Margison, F., Evans, C., & Mellor-Clark, J. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology*, *74*, 160–167.

Barkham, M., Evans, C., Margison, F., Mcgrath, G., Mellor-Clark, J., Milne, D., & Connell, J. (1998). The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Journal of Mental Health*, *7*, 35–47.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., et al. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*, 184–196.

Barkham, M., Stiles, W. B., Connell, J., Twigg, E., Leach, C., Lucock, M., et al. (2008). Effects of psychological therapies in randomized trials and practice-based studies. *British Journal of Clinical Psychology*, *47*, 397–415.

Barlow, D. H. (2005). What's new about evidence-based assessment? *Psychological Assessment*, *17*, 308–311.

Barlow, D. H. (2010). Negative effects from psychological treatments: A perspective. *American Psychologist*, *65*, 13–20.

Bassler, M., Potratz, B., & Krauthauser, H. (1995). Der "Helping Alliance Questionnaire" (HAQ) von Luborsky. *Psychotherapeut*, *40*, 23–32.

Beaton, D. E. (2003). Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Medical Care*, *41*, 593-596.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*, 319 –334.

Berger, M., & Linden, M. (2009). Brauchen wir noch Psychotherapieschulen bzw. -verfahren? *Verhaltenstherapie*, *19*, 263–271.

Bergmann-Warnecke, K. (2011). *Individuelle Kontrolle und Vorhersage von Therapieverläufen: Anwendung der Zeitreihenanalyse*. Trier: Unveröffentlichte Diplomarbeit.

Berking, M., Orth, U., & Lutz, W. (2006). Wie effektiv sind systematische Rückmeldungen des Therapieverlaufs an den Therapeuten? *Zeitschrift für Klinische Psychologie und Psychotherapie*, *35*, 21–29.

Beutler, L. E. (1998). Identifying empirically supported treatments: What if we didn't? *Journal of Consulting and Clinical Psychology*, *66*, 113–120.

Beutler, L. E. (1999). Manualizing flexibility: The training of eclectic therapists. *Journal of Clinical Psychology*, *55*, 399–404.

Beutler, L. E. (2001). Comparisons among quality assurance systems: From outcome assessment to clinical utility. *Journal of Consulting and Clinical Psychology*, *69*, 197–204.

Beutler, L. E., & Harwood, M. (2000). *Prescriptive psychotherapy: A practical guide to systematic treatment selection*. New York: Oxford Univ. Press.

Beutler, L. E., Moleiro, C., & Talebi, H. (2002). How practitioners can systematically use empirical evidence in treatment selection. *Journal of Clinical Psychology*, *58*, 1199–1212.

Bewick, B. M., Trusler, K., Mullin, T., Grant, S., & Mothersole, G. (2006). Routine outcome measurement completion rates of CORE-OM in primary care psychological therapies and counselling. *Counselling and Psychotherapy Research*, *6*, 33–40.

Bickman, L., & Hoagwood, K. E. (2010). Introduction to special issue. *Administration and Policy in Mental Health and Mental Health Services Research*, *37*, 4–6.

Bickman, L., Kelley, S. D., Breda, C., De Andrade, A. R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services*, *62*, 1423–1429.

Bishop, M. J., Bybee, T. S., Lambert, M. J., Burlingame, G. M., Wells, M. G., & Poppleton, L. E. (2005). Accuracy of a rationally derived method for identifying treatment failure in children and adolescents. *Journal of Child and Family Studies*, *14*, 207–222.

Blanchin, M., Hardouin, J.-B., Neel, T. L., Kubis, G., Blanchard, C., Mirallié, E., & Sébille, V. (2011). Comparison of CTT and Rasch-based approaches for the analysis of longitudinal patient reported outcomes. *Statistics in Medicine*, *30*, 825–838.

Böhnke, J. R., & Lutz, W. (submitted). Using item information to optimize targeted assessments of psychological distress.

Böhnke, J. R., & Lutz, W. (2008). Response patterns in questionnaires: From theory to the application in clinical practice to identify aberrant responses. Presented at the 39th International Meeting of the Society for Psychotherapy Research, Barcelona, Spain.

Böhnke, J. R., & Lutz, W. (2010a). Item response theory in clinical practice: Results from simulation studies and applications in psychotherapy settings. Presented at the 41st International Meeting of the Society for Psychotherapy Research, Asilomar, CA, USA.

Böhnke, J. R., & Lutz, W. (2010b). Konstruktion von Skalen & Kennwerten für das Monitoring von Veränderungen in der Psychotherapie. (W. Hiller & M. Witthöft, Eds.) *Zeitschrift für Klinische Psychologie und Psychotherapie*, *39*(S1), 6.

Böhnke, J. R., & Lutz, W. (2010c). War da was – oder doch nicht? Methoden zur Entwicklung veränderungssensitiver Kurzformen für Verlaufsmessung und Qualitätsmonitoring. *Klinische Diagnostik und Evaluation*, *3*, 38–58.

Böhnke, J. R., & Lutz, W. (2011a). Re-evaluating the fit of inherently multi-dimensional measures: A new look on the phase model of psychotherapy outcome. Presented at the 42nd International Meeting, Bern, Switzerland.

Böhnke, J. R., & Lutz, W. (2011b). Epidemiologie und Versorgungsforschung. In *Klinische Psychologie - Grundlagen* (Band 1, pp. 269–294). Göttingen: Hogrefe.

Böhnke, J. R., & Lutz, W. (2011c). Estimating the Rasch model in R: A simulation study comparing eRm, ltm, and mixRasch. Presented at the 5th Rasch User Day, Durham, UK.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314.

Bolt, D. M. (2005). Limited- and full-information estimation of item response theory models. In *Contemporary psychometrics* (pp. 27–71). Mahwah, NJ: Lawrence Erlbaum.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Bootzin, R. R., & Bailey, E. T. (2005). Understanding placebo, nocebo, and iatrogenic treatment effects. *Journal of Clinical Psychology*, *61*, 871–880.

Borkovec, T. D., Echemendia, R. J., Ragusea, S. A., & Ruiz, M. (2001). The Pennsylvania Practice Research Network and future possibilities for clinically meaningful and scientifically rigorous psychotherapy effectiveness research. *Clinical Psychology: Science and Practice*, *8*, 155–167.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203–219.

Bottler, B. (2009). *Flexibilität der Zielanpassung als Prädiktor der psychopathologischen Belastung: Zusammenhänge mit dem Fragebogen zur Evaluation von Psychotherapieverläufen (FEP)*. Trier: Unveröffentlichte Diplomarbeit.

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, *74*, 1–4.

Brabec, B., & Meister, R. (2001). A nearest-neighbor model for regional avalanche forecasting. *Annals of Glaciology*, *32*, 130–134.

Bram, A. D. (2010). The Relevance of the Rorschach and patient–examiner relationship in treatment planning and outcome assessment. *Journal of Personality Assessment*, *92*, 91 – 115.

Brod, M., Tesler, L. E., & Christensen, T. L. (2009). Qualitative research and content validity: Developing best practices based on science and experience. *Quality of Life Research*, *18*, 1263–1278.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2012). On the factor structure of the Beck Depression Inventory–II: G is the key. *Psychological Assessment*, online first.

Brown, T. A., & Barlow, D. H. (2009). A proposal for a dimensional classification system based on the shared features of the DSM-IV anxiety and mood disorders: Implications for assessment and treatment. *Psychological Assessment*, *21*, 256–271.

Bullinger, M., & Kirchberger, I. (1998). *SF-36. Fragebogen zum Gesundheitszustand*. Göttingen: Hogrefe.

Burlingame, G. M., Seaman, S., Johnson, J. E., Whipple, J., Richardson, E., Rees, F., et al. (2006). Sensitivity to change of the Brief Psychiatric Rating Scale-Extended (BPRS-E): An item and subscale analysis. *Psychological Services*, *3*, 77–87.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *MMPI–A (Minnesota Multiphasic Personality Inventory-Adolescent): Manual for administration, scoring, and interpretation.* Minneapolis, MN: University of Minnesota Press.

Cahill, J., Barkham, M., Stiles, W. B., Twigg, E., Rees, A., Hardy, G. E., & Evans, C. (2006). Convergent validity of the CORE measures with measures of depression for clients in brief cognitive therapy for depression. *Journal of Counseling Psychology*, *53*, 253–259.

Cahill, J., Stiles, W. B., Barkham, M., Hardy, G. E., Stone, G., Agnew-Davies, R., & Unsworth, G. (2011). Two short forms of the Agnew Relationship Measure: The ARM-5 and ARM-12. *Psychotherapy Research*, *22*, 241–255.

Cannon, J. A. N., Warren, J. S., Nelson, P. L., & Burlingame, G. M. (2010). Change trajectories for the Youth Outcome Questionnaire Self-Report: Identifying youth at risk for treatment failure. *Journal of Clinical Child & Adolescent Psychology*, *39*, 289–301.

Carlier, I. V. E., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van der Wee, N. J. A., & Zitman, F. G. (2012). Routine outcome monitoring and feedback on physical or mental health status: Evidence and theory. *Journal of Evaluation in Clinical Practice*, *18*, 104–110.

Castonguay, L. G. (2011). Psychotherapy, psychopathology, research and practice: Pathways of connections and integration. *Psychotherapy Research*, *21*, 125–140.

Castonguay, L. G., Boswell, J. F., Constantino, M. J., Goldfried, M. R., & Hill, C. E. (2010). Training implications of harmful effects of psychological treatments. *American Psychologist*, *65*, 34–49.

Cella, D., Gershon, R., Lai, J.-S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*, 133–141.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. on behalf of the PROMIS Cooperative Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, *45*, S3–S11.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Chang, C.-H. (2007). Patient-reported outcomes measurement and management with innovative methodologies and technologies. *Quality of Life Research*, *16*, 157–166.

Chang, C.-H., & Reeve, B. B. (2005). Item Response Theory and its applications to patient-reported outcomes measurement. *Evaluation & the Health Professions*, *28*, 264 –282.

Chiles, J. A., Lambert, M. J., & Hatch, A. L. (1999). The impact of psychological interventions on medical cost offset: A meta-analytic review. *Clinical Psychology: Science and Practice*, *6*, 204–220.

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Applied Psychological Measurement*, *34*, 483 –504.

Choi, S., Reise, S. P., Pilkonis, P., Hays, R., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*, 125–136.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/Item Response Theory and Monte Carlo Simulations. *Journal of Statistical Software*, *39*(8).

Claiborn, C. D., & Goodyear, R. K. (2005). Feedback in psychotherapy. *Journal of Clinical Psychology*, *61*, 209–217.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., Chan, T., Jiang, T., & Seburn, M. (2008). Consistent estimation of Rasch item parameters and their standard errors under complex sample designs. *Applied Psychological Measurement, 32,* 289–310.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement, 14,* 183 –196.

Colder, C. R., Campbell, R. T., Ruel, E., Richardson, J. L., & Flay, B. R. (2002). A finite mixture model of growth trajectories of adolescent alcohol use: Predictors and consequences. *Journal of Consulting and Clinical Psychology, 70,* 976–985.

Connell, J., Barkham, M., Stiles, W. B., Twigg, E., Singleton, N., Evans, O., & Miles, J. N. V. (2007). Distribution of CORE-OM scores in a general population, clinical cut-off points and comparison with the CIS-R. *The British Journal of Psychiatry, 190,* 69–74.

Contopoulos-Ioannidis, D. G., Alexiou, G. A., Gouvias, T. C., & Ioannidis, J. P. A. (2008). Life cycle of translational research for medical interventions. *Science, 321,* 1298 –1299.

Cook, K. F., Choi, S. W., Crane, P. K., Deyo, R. A., Johnson, K. L., & Amtmann, D. (2008). Letting the CAT out of the bag: Comparing computer adaptive tests and an 11-item short form of the Roland-Morris Disability Questionnaire. *Spine, 33,* 1378–1383.

Coutinho, J., Ribeiro, E., Hill, C., & Safran, J. (2011). Therapists' and clients' experiences of alliance ruptures: A qualitative study. *Psychotherapy Research, 21,* 525–540.

Crane, P., Gibbons, L., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., et al. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research, 16,* 69–84.

Crits-Christoph, P. (1997). Limitations of the dodo bird verdict and the role of clinical trials in psychotherapy research: Comment on Wampold et al. (1997). *Psychological Bulletin, 122,* 216–220.

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A., Carroll, K., Perry, K., et al. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1,* 81–91.

Crits-Christoph, P., Gibbons, M.B., & Hearon, B. (2006). Does the alliance cause good outcome? Recommendations for future research on the alliance. *Psychotherapy: Theory, Research, Practice, Training, 43,* 280–285.

Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review, 30,* 768–778.

Cuijpers, P., Smits, N., Donker, T., Ten Have, M., & De Graaf, R. (2009). Screening for mood and anxiety disorders with the five-item, the three-item, and the two-item Mental Health Inventory. *Psychiatry Research*, *168*, 250–255.

Cuijpers, P., Van Straten, A., Andersson, G., & Van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology*, *76*, 909–922.

Cuijpers, P., Van Straten, A., Warmerdam, L., & Smits, N. (2008). Characteristics of effective psychological treatments of depression: A metaregression analysis. *Psychotherapy Research*, *18*, 225–236.

Culpepper, S. A., & Aguinis, H. (2010). R is for Revolution: A cutting-edge, free, open source statistical package. *Organizational Research Methods*, *14*, 735-740.

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*, 583–619.

Cyr, J. J., McKenna-Foley, J. M., & Peacock, E. (1985). Factor structure of the SCL-90-R: Is there one? *Journal of Personality Assessment*, *49*, 571–578.

Dattatreya, G. R. (2002). Gaussian mixture parameter estimation with known means and unknown class-dependent variances. *Pattern Recognition*, *35*, 1611–1616.

Davier, M. von. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a monte carlo study. *Methods of Psychological Research - online*, *2*, 29–48.

Davier, M. von. (2000). *WINMIRA 2001 user's guide*. Kiel: IPN.

De Jong, K., Van Sluis, P., Nugter, M. A., Heiser, W. J., & Spinhoven, P. (2012). Understanding the differential impact of outcome monitoring: Therapist variables that moderate feedback effects in a randomized clinical trial. *Psychotherapy Research*, *22*, 464–474.

De Los Reyes, A., & Kazdin, A. E. (2006). Conceptualizing changes in behavior in intervention research: The range of possible changes model. *Psychological Review*, *113*, 554–583.

De Los Reyes, A., Kundey, S. M. A., & Wang, M. (2011). The end of the primary outcome measure: A research agenda for constructing its replacement. *Clinical Psychology Review*, *31*, 829–838.

De Vet, H. C. W., Terluin, B., Knol, D. L., Roorda, L. D., Mokkink, L. B., Ostelo, R. W. J. G., et al. (2010). Three ways to quantify uncertainty in individually applied "minimally important change" values. *Journal of clinical epidemiology*, *63*, 37–45.

DeMars, C. (2010). *Item response theory*. Oxford: Oford Univ. Press.

Derogatis, L. R. (1977). *SCL-90, administration, scoring, and procedures. Manual 1 for the R(evised) version and other instruments of the Psychopathology Rating Scale Series.* Baltimore: Johns Hopkins University School of Medicine.

Derogatis, L. R. (1993). *BSI. Brief Symptom Inventory: Administration, scoring, and procedures manual* (4th edition). Minneapolis, MN: National Computer Systems.

Derogatis, L. R. (2001). *Brief Symptom Inventory (BS)-18: Administration, scoring, and procedures manual.* Minneapolis, MN: National Computer Systems Pearson.

DeRubeis, R. J., Brotman, M. A., & Gibbons, C. J. (2005). A conceptual and methodological analysis of the nonspecifics argument. *Clinical Psychology: Science and Practice*, *12*, 174–183.

DeSalvo, K., Fisher, W., Tran, K., Bloser, N., Merrill, W., & Peabody, J. (2006). Assessing measurement properties of two single-item general health measures. *Quality of Life Research, 15*, 191–201.

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*, S50–S59.

Dimidjian, S., & Hollon, S. D. (2010). How would we know if psychotherapy were harmful? *American Psychologist*, *65*, 21–33.

Dold, M., Lenz, G., Demal, U., & Aigner, M. (2010). Monitoring- und Feedback-Systeme in der Psychotherapie. *Psychotherapie Forum*, (4), 208–214.

Donabedian, A. (2005). Evaluating the quality of medical care. *The Milbank Quarterly*, *85*, 691–729.

Doucette, A., & Wolf, A. W. (2009). Questioning the measurement precision of psychotherapy research. *Psychotherapy Research*, *19*, 374–389.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*, 363–373.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189–199.

Duhem, P. (1998). Physical theory and experiment. In M. Curd & J. A. Cover (Eds.), *Philosophy of science: The central issues* (pp. 257–279). New York: W.W. Norton & Company.

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment*, *20*, 55–62.

Duncan, B. L. (2012). The Partners for Change Outcome Management System (PCOMS): The Heart and Soul of Change Project. *Canadian Psychology/Psychologie canadienne*, *53*, 93–104.

Durham, C. J., McGrath, L. D., Burlingame, G. M., Schaalje, G. B., Lambert, M. J., & Davies, D. R. (2002). The effects of repeated administrations on self-report and parent-report scales. *Journal of Psychoeducational Assessment, 20*, 240–257.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*, 370–388.

Edwards, M. C., Cheavens, J. S., Heiy, J. E., & Cukrowicz, K. C. (2010). A reexamination of the factor structure of the Center for Epidemiologic Studies Depression Scale: Is a one-factor model plausible? *Psychological Assessment, 22*, 711–715.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ehlers, A., & Margraf, J. (2001). *AKV: Fragebogen zu körperbezogenen Ängsten, Kognitionen und Vermeidung* (2. überarb. & neunormierte Auflage). Göttingen: Beltz.

Eid, M., Gollwitzer, M., & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Beltz.

Eisen, S. V., & Dickey, B. (1996). Mental health outcome assessment: The new agenda. *Psychotherapy: Theory, Research, Practice, Training, 33*, 181–189.

Elliott, R. (2010). Psychotherapy change process research: Realizing the promise. *Psychotherapy Research, 20*, 123–135.

Ellwood, P. M. (1988). Outcomes management. *New England Journal of Medicine, 318*, 1549–1556.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12*, 105–120.

Emsley, R., Dunn, G., & White, I. R. (2010). Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research, 19*, 237 –270.

Evans, C. (2012). Cautionary notes on power steering for psychotherapy. *Canadian Psychology/Psychologie canadienne, 53*, 131–139.

Everitt, B. S. (2005). *An R and S-Plus companion to multivariate analysis*. London: Springer.

Everitt, B. S., & Hothorn, T. (2010). *A handbook of statistical analyses using R* (2nd edition). Boca Raton, FL: Chapman & Hall.

Everitt, B. S., & Wessely, S. (2008). *Clinical trials in psychiatry* (2nd edition). Chichester: Wiley.

Ey, S., & Hersen, M. (2004). Pragmatic issues of assessment in clinical practice. In M. Hersen (Ed.), *Psychological assessment in clinical practice: A pragmatic guide* (pp. 3–20). New York: Brunner-Routledge.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*, 319–324.

Fang, J., Power, M., Lin, Y., Zhang, J., Hao, Y., & Chatterji, S. (2011). Development of short versions for the WHOQOL-OLD module. *The Gerontologist*, *52*, 66–78.

Farin, E., & Bengel, J. (2003). Qualitätssicherung, Evaluationsforschung und Psychotherapieforschung: Abgrenzung und Zusammenwirken. In M. Härter, H. W. Linster, & R.-D. Stieglitz (Eds.), *Qualitätsmanagement in der Psychotherapie: Grundlagen, Methoden und Anwendung* (pp. 47–68). Göttingen: Hogrefe.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fayers, P. M., & Machin, D. (2007). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes* (2nd edition). Chichester: Wiley.

Fellows, I. (2012). *Deducer*. CRAN. Retrieved December 13, 2012, from http://www.deducer.org/manual.html

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1996). *Structured clinical interview for DSM-IV axis I disorders, clinician version (SCID-CV).* Washington, DC: American Psychiatric Press.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*, 59–77.

Fisseni, H.-J. (2004). *Lehrbuch der psychologischen Diagnostik* (3., überarb. und erw. Auflage). Göttingen: Hogrefe.

Fliege, H., Becker, J., Walter, O., Bjorner, J., Klapp, B., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, *14*, 2277–2291.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Flora, D., & Thissen, D. (2002). *User's guide for IRTSCORE: Item response theory score approximation software* (No. L.L. Thurstone Psychometric Laboratory Electronic Research Memorandum 2002-1). Chapel Hill, NC: L.L. Thurstone Psychometric Laboratory. Retrieved December 13, 2012, from http://www.unc.edu/depts/psychology/dthissen/840F10/IRTScore.pdf

Flückiger, C., Del Re, A. C., Wampold, B. E., Symonds, D., & Horvath, A. O. (2012). How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. *Journal of Counseling Psychology*, *59*, 10–17.

Flückiger, C., Regli, D., Zwahlen, D., Hostettler, S., & Caspar, F. (2010). Der Berner Patienten- und Therapeutenstundenbogen 2000. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *39*, 71–79.

Food and Drug Administration. (2006). Draft guidance for industry or patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register*, *71*, 5862–5863.

Forkmann, T., Böcker, M., Wirtz, M., Norra, C., & Gauggel, S. (2012). Entwicklung, Validierung und Normierung des Rasch-basierten Depressionsscreenings. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *41*, 19–29.

Forkmann, T., Boecker, M., Wirtz, M., Glaesmer, H., Brahler, E., Norra, C., & Gauggel, S. (2010). Validation of the Rasch-based depression screening in a large scale German general population sample. *Health and Quality of Life Outcomes*, *8*, 105.

Fowler, J. C. (2012). Suicide risk assessment in clinical practice: Pragmatic guidelines for imperfect assessments. *Psychotherapy*, *49*, 81–90.

Fowler, J. C., Ackerman, S. J., Speanburg, S., Bailey, A., Blagys, M., & Conklin, A. C. (2004). Personality and symptom change in treatment-refractory inpatients: Evaluation of the phase model of change using Rorschach, TAT, and DSM-IV Axis V. *Journal of Personality Assessment*, *83*, 306–322.

Fox, J. (2005). The R commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software*, *14*(9).

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd edition). Los Angeles, CA: Sage.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–631.

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Franke, G. (2000). *BSI. Brief Symptom Inventory – Deutsche Version. Manual*. Göttingen: Beltz.

Fung, C. H., & Hays, R. D. (2008). Prospects and challenges in using patient-reported outcomes in clinical practice. *Quality of Life Research*, *17*, 1297–1302.

Fydrich, T., Nagel, A., Lutz, W., & Richter, R. (2003). Qualitätsmonitoring in der ambulanten Psychotherapie: Modellprojekt der Techniker Krankenkasse. *Verhaltenstherapie*, *13*, 291–295.

Garland, A. F., Bickman, L., & Chorpita, B. F. (2010). Change what? Identifying quality improvement targets by investigating usual mental health care. *Administration and Policy in Mental Health and Mental Health Services Research*, *37*, 15–26.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gibson, W. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, *24*, 229–252.

Gilbody, S. M., House, A. O., & Sheldon, T. A. (2002a). Outcomes research in mental health: Systematic review. *British Journal of Psychiatry*, *181*, 8–16.

Gilbody, S. M., House, A. O., & Sheldon, T. A. (2002b). Psychiatrists in the UK do not use outcomes measures. *British Journal of Psychiatry*, *180*, 101–103.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitations of four different approaches. *Methods of Psychological Research - online*, *2*(1). Retrieved December 13, 2012, from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art6/article.html

Glück, J., & Spiel, C. (2007). Studying development via item response models: A wide range of potential uses. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 281–292). New York: Springer.

Gmür, W., & Straus, F. (1998). Partizipatives Qualitätsmanagement in psychosozialen Beratungsstellen: Hintergründe, Anforderungen und Möglichkeiten von Qualitätssicherung nach dem "Münchener Modell". In A.-R. Laireiter & H. Vogel (Eds.), *Qualitätssicherung in der Psychotherapie und psychosozialen Versorgung: Ein Werkstattbuch* (pp. 75–99). Tübingen: DGVT Verlag.

Goldberg, D., & Goodyer, I. (2005). *The origins and course of common mental disorders*. London: Routledge.

Gollwitzer, M. (2007). Latent-Class-Analysis. In H. Moosbugger & A. Kelava, *Testtheorie und Fragebogenkonstruktion* (pp. 279–306). Heidelberg: Springer.

Gonçalves, M. M., & Stiles, W. B. (2011). Narrative and psychotherapy: Introduction to the special section. *Psychotherapy Research*, *21*, 1–3.

Grawe, K. (1982). Psychotherapieforschung. In R. Bastine, P. Fiedler, S. Schmidtchen, & G. Sommer (Eds.), *Grundbegriffe der Psychotherapie* (pp. 323–331). Weinheim: Edition Psychologie.

Grawe, K. (1997). Research-informed psychotherapy. *Psychotherapy Research*, *7*, 1–19.

Grawe, K. (1998). *Psychologische Therapie*. Göttingen: Hogrefe.

Grawe, K. (2004). *Neuropsychotherapie*. Göttingen: Hogrefe.

Grawe, K., & Baltensberger, C. (1998). Figurationsanalyse - Ein Konzept und Computerprogramm für die Prozess- und Ergebnisevaluation in der Therapiepraxis. In A.-R. Laireiter & H. Vogel (Eds.), *Qualitätssicherung in der Psychotherapie und psychosozialen Versorgung: Ein Werkstattbuch* (pp. 179–207). Tübingen: DGVT Verlag.

Grawe, K., Donati, R., & Bernauer, F. (1994). *Psychotherapie im Wandel: Von der Konfession zur Profession*. Göttingen: Hogrefe.

Grissom, G. R., Lyons, J. S., & Lutz, W. (2002). Standing on the shoulders of a giant: Development of an outcome management system based on the dose model and phase model of psychotherapy. *Psychotherapy Research*, *12*, 397–412.

Grosse Holtforth, M., Lutz, W., & Egenolf, Y. (2010). Diagnostik und Therapieplanung in der Psychotherapie. In W. Lutz (Ed.), *Lehrbuch Psychotherapie* (pp. 71–87). Bern: Huber.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30.

Haas, E., Hill, R. D., Lambert, M. J., & Morrell, B. (2002). Do early responders to psychotherapy maintain treatment gains? *Journal of Clinical Psychology*, *58*, 1157–1172.

Hagemeister, C., Lang, F., & Kersting, M. (2010). Einstellung von Psychologinnen und Psychologen in Deutschland zu Tests. *Report Psychologie*, *35*, 428–439.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

Halstead, J. E., Leach, C., & Rust, J. (2007). The development of a brief distress measure for the evaluation of psychotherapy and counseling (sPaCE). *Psychotherapy Research*, *17*, 656–672.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*, 155–163.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, *9*, 329–343.

Hänsgen, K.-D. (2006). *Hogrefe Test System*. Göttingen: Hogrefe.

Hanson, W. E., & Poston, J. M. (2011). Building confidence in psychological assessment as a therapeutic intervention: An empirically based reply to Lilienfeld, Garb, and Wood (2011). *Psychological Assessment, 23*, 1056–1062.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E. J., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist–client feedback and clinical support tools. *Psychotherapy Research, 17*, 379–392.

Hart, D. L., Werneke, M. W., George, S. Z., & Deutscher, D. (2012). Single-item screens identified patients with elevated levels of depressive and somatization symptoms in outpatient physical therapy. *Quality of Life Research, 21*, 257–268.

Härter, M., Linster, H. W., & Stieglitz, R.-D. (2003). Grundlagen und Konzepte von Qualitätsmanagement in der Psychotherapie. In M. Härter, H. W. Linster, & R.-D. Stieglitz (Eds.), *Qualitätsmanagement in der Psychotherapie: Grundlagen, Methoden und Anwendung* (pp. 17–46). Göttingen: Hogrefe.

Harwell, M. R. (1997). Analyzing the results of monte carlo studies in item response theory. *Educational and Psychological Measurement, 57*, 266 –279.

Harwell, M. R., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101 –125.

Hatfield, D., McCullough, L., Frantz, S. H. B., & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy, 17*, 25–32.

Hatfield, D., & Ogles, B. (2007). Why some clinicians use outcome measures and others do not. *Administration and Policy in Mental Health and Mental Health Services Research, 34*, 283–291.

Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice, 35*, 485–491.

Hatzinger, R., & Rusch, T. (2009). IRT models with relaxed assumptions in eRm: A manual-like instruction. *Psychology Science Quarterly, 51*, 87–120.

Haughton, D., Legrand, P., & Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent GOLD, poLCA, and MCLUST. *The American Statistician, 63*, 81–91.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44*, 1–12.

Hays, R. D., Brown, J., Brown, L. U., Spritzer, K. L., & Crall, J. J. (2006). Classical test theory and item response theory analyses of multi-item scales assessing parents' perceptions of their children's dental care. *Medical Care*, *44*, S60–S68.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *39*, II28–II42.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191 –205.

Helbig, M., Theus, M., & Urbanek, S. (2005). JGR: Java GUI for R. *Statistical Computing & Graphics (Newsletter of ASA)*, *16*, 9–12.

Helmreich, I., Wagner, S., Mergl, R., Allgaier, A.-K., Hautzinger, M., Henkel, V., et al. (2011). The Inventory of Depressive Symptomatology (IDS-C28) is more sensitive to changes in depressive symptomatology than the Hamilton Depression Rating Scale (HAMD17) in patients with mild major, minor or subsyndromal depression. *European Archives of Psychiatry and Clinical Neuroscience*, *261*, 357–367.

Hennig, C. (2010). Mathematical models and reality: A constructivist perspective. *Foundations of Science*, *15*, 29–48.

Henretty, J. R., Levitt, H. M., & Mathews, S. S. (2008). Clients' experiences of moments of sadness in psychotherapy: A grounded theory analysis. *Psychotherapy Research*, *18*, 243–255.

Hentschel, U. (2005a). Die therapeutische Allianz - Teil 1: Die Entwicklungsgeschichte des Konzepts und moderne Forschungsansätze. *Psychotherapeut*, *50*, 305–317.

Hentschel, U. (2005b). Die therapeutische Allianz - Teil 2: Ergänzende Betrachtungen über Verbindungen und Abgrenzungsmöglichkeiten zu ähnlichen Konstrukten. *Psychotherapeut*, *50*, 385–393.

Hersen, M. (Ed.). (2004). *Psychological assessment in clinical practice: A pragmatic guide*. New York: Brunner-Routledge.

Hidalgo, M., & López-Pina, J. (2011). Item-fit evaluation in biased tests: A study under Rasch model. *Quality & Quantity*, *45*, 715–734.

Hill, C. E., & Knox, S. (2009). Processing the therapeutic relationship. *Psychotherapy Research*, *19*, 13–29.

Hiller, W., Zaudig, M., & Mombour, W. (2004). *IDCL - International Diagnostic Checklists for ICD-10 and DSM-IV*. Bern: Huber.

Höhler, J., Hartig, J., & Ullrich, M. (2012). Arbeitsgruppe: Modellbasierte Diagnostik von Schülerkompetenzen. Presented at the 48. Kongress der Deutschen Gesellschaft für Psychologie, Bielefeld.

Holman, R., Glas, C. A. W., & De Haan, R. J. (2003). Power analysis in randomized clinical trials based on item response theory. *Controlled Clinical Trials, 24*, 390–410.

Holman, R., Lindeboom, R., Glas, C. A. W., Vermeulen, M., & De Haan, R. J. (2003). Constructing an item bank using item response theory: The AMC linear disability score project. *Health Services & Outcomes Research Methodology, 4*, 19–33.

Holman, R., Weisscher, N., Glas, C. A. W., Dijkgraaf, M., Vermeulen, M., De Haan, R., & Lindeboom, R. (2005). The Academic Medical Center Linear Disability Score (ALDS) item bank: Item response theory analysis in a mixed patient population. *Health and Quality of Life Outcomes, 3*, 83.

Horowitz, L. M., Strauss, B., & Kordy, H. (2000). Inventar zur Erfassung interpersonaler Probleme (IIP-D). Handanweisung (2. Auflage). Weinheim: Beltz.

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy, 48*, 9–16.

Horvath, A. O., & Luborsky, L. (1993). The role of the therapeutic alliance in psychotherapy. *Journal of Consulting and Clinical Psychology, 61*, 561–573.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology, 38*, 139–149.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist, 41*, 159–164.

Howard, K. I., Krause, M. S., Saunders, S. M., & Kopta, S. M. (1997). Trials and tribulations in the meta-analysis of treatment differences: Comment on Wampold et al. (1997). *Psychological Bulletin, 122*, 221–225.

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of Consulting and Clinical Psychology, 61*, 678–685.

Howard, K. I., Moras, K., Brill, B. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness and patient progress. *American Psychologist, 51*, 1059–1064.

Howard, K. I., & Orlinsky, D. E. (1972). Psychotherapeutic processes. *Annual Review of Psychology, 23*, 615–668.

Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 131–137.

Hubert, L., & Wainer, H. (2011). A statistical guide for the ethically perplexed. In A. T. Panter & S. K. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 61–124). New York: Routledge.

Hunsley, J., & Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychological Assessment*, *17*, 251–255.

Hunter, A., M., Muthén, B. O., Cook, I. A., & Leuchter, A. F. (2010). Antidepressant response trajectories and quantitative electroencephalography (QEEG) biomarkers in major depressive disorder. *Journal of Psychiatric Research*, *44*, 90–98.

Huppert, F. A., & Whittington, J. E. (2003). Evidence for the independence of positive and negative well-being: Implications for quality of life assessment. *British Journal of Health Psychology*, *8*, 107–122.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika*, *59*, 111–119.

Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. *Psychometrika*, *61*, 31–39.

ICH E9. (1998). *Statistical principles for clinical trials* (No. CPMP/ICH/291/95; adopted by CPMP March 1998). London: International Conference on Harmonisation.

Ilardi, S. S., & Craighead, W. E. (1994). The role of nonspecific factors in cognitive-behavior therapy for depression. *Clinical Psychology: Science and Practice*, *1*, 138–155.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester: Wiley.

Jacob, G., & Bengel, J. (2003). Die Perspektive der Patienten. In M. Härter, H. W. Linster, & R.-D. Stieglitz (Eds.), *Qualitätsmanagement in der Psychotherapie: Grundlagen, Methoden und Anwendung* (pp. 119–132). Göttingen: Hogrefe.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Jensen-Doss, A. (2011). Practice involves more than treatment: How can evidence-based assessment catch up to evidence-based treatment? *Clinical Psychology: Science and Practice*, *18*, 173–177.

Jensen-Doss, A., & Hawley, K. M. (2011). Understanding clinicians' diagnostic practices: Attitudes toward the utility of diagnosis and standardized diagnostic tools. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 476–485.

Johnson, L. D., & Shaha, S. (1996). Improving quality in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *33*, 225–236.

Joint Comission on Accreditation of Health Care Organizations. (1996). *A guide to performance improvement in behavioral health care organizations*. Oakbrook Terrace, IL: JCAHO.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Jung, F. (2008). *Allgemeine Selbstwirksamkeit: Zusammenhänge mit den Dimensionen des psychischen Wohlbefindens im Fragebogen zur Evaluation von Psychotherapieverläufen (FEP)*. Trier: Unveröffentlichte Diplomarbeit.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65 –81.

Kamphuis, J. H., & Noordhof, A. (2009). On categorical diagnoses in "DSM-V": Cutting dimensions at useful points? *Psychological Assessment*, *21*, 294–301.

Katsikopoulos, K. V., Pachur, T., Machery, E., & Wallin, A. (2008). From Meehl to fast and frugal heuristics (and back). *Theory and Psychology*, *18*, 443–464.

Kazdin, A. E. (1998). *Research design in clinical psychology* (3rd edition). Boston, MA: Allyn and Bacon.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist*, *63*, 146–159.

Kazdin, A. E. (2009). Understanding how and why psychotherapy leads to change. *Psychotherapy Research*, *19*, 418–428.

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*, 21 –37.

Keller, F., Hautzinger, M., & Kühner, C. (2008). Zur faktoriellen Struktur des deutschsprachigen BDI-II. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *37*, 245–254.

Keller, F., & Kempf, W. (1997). Some latent trait and latent class analyses of the Beck-Depression-Inventory (BDI). In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 314–323). Münster: Waxmann. Retrieved December 13, 2012, from http://kops.ub.uni-konstanz.de/handle/urn:nbn:de:bsz:352-opus-82221

Kempf, W. (2003). *Forschungsmethoden der Psychologie: Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik* (Band I: Theorie und Empirie). Berlin: Regener.

Kempf, W. (2008). *Forschungsmethoden der Psychologie: Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik* (Band II: Quantität und Qualität). Berlin: Regener.

Kempf, W. (2010). Quantifizierung qualitativer Daten. *Diskussionsbeiträge der Projektgruppe Friedensforschung*, *65*. Retrieved from December 13, 2012, http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-121212

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., et al. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the united states: Results from the national comorbidity survey. *Archives of General Psychiatry*, *51*, 8–9.

Klotsche, J., Ferger, D., Pieper, L., Rehm, J., & Wittchen, H.-U. (2009). oA novel nonparametric approach for estimating cut-offs in continuous risk indicators with application to diabetes epidemiology. *BMC Medical Research Methodology*, *9*, 63.

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *The British Journal of Psychiatry*, *195*, 15 – 22.

Knox, S., Adrians, N., Everson, E., Hess, S., Hill, C., & Crook-Lyon, R. (2011). Clients' perspectives on therapy termination. *Psychotherapy Research*, *21*, 154–167.

Köck, K. (2012). *Komorbidität in der ambulanten Psychotherapie: Eine Untersuchung ihres Einflusses auf Status, Verlauf und Ergebnis im Rahmen eines Modellprojekts zum Qualitätsmonitoring*. Trier: OPUS. Retrieved December 13, 2012, from http://ubt.opus.hbz-nrw.de/volltexte/2012/756/

Köhler, C. (2012). *Kann das Raschmodell Abweichungen bei der Vorhersage von Psychotherapieverläufen erklären?* Trier: Unveröffentlichte Diplomarbeit.

Kopta, S. M., Howard, K. I., Lowry, J. L., & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Consulting and Clinical Psychology*, *62*, 1009–1016.

Kordy, H., Hannöver, W., & Richard, M. (2001). Computer-assisted feedback-driven quality management for psychotherapy: The Stuttgart-Heidelberg Model. *Journal of Consulting and Clinical Psychology*, *69*, 173–183.

Kraemer, H. C., & Gibbons, R. D. (2009). Why does the randomized clinical trial methodology so often mislead clinical decision making? Focus on moderators and mediators of treatment. *Psychiatric Annals*, *39*, 736–745.

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 1524–1529.

Krampen, G. (2002). *Stundenbogen für die allgemeine und differentielle Einzel-Psychotherapie (STEP)*. Göttingen: Hogrefe.

Krampen, G. (2010). Experimentelle Konstruktion eines Kurzfragebogens zur direkten Veränderungsmessung psychotherapeutischer Effekte im Befinden. *Diagnostica*, *56*, 212–221.

Krampen, G., & Hank, P. (2008). Prozessdiagnostik und kontrollierte Praxis. In B. Röhrle, F. Caspar, & P. F. Schlottke (Eds.), *Lehrbuch der klinisch-psychologischen Diagnostik* (pp. 300–329). Stuttgart: Kohlhammer.

Krampen, G., Schui, G., & Wiesenhütter, J. (2008). Evidenzbasierte Psychotherapie und Therapie-Ressourcen: Ein erweitertes 4-Phasen-Prüfmodell und seine Anwendung auf die klinisch-psychologische Fachliteratur aus dem deutschsprachigen Bereich. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *37*, 43–51.

Krause, M. S. (2010). Trying to discover sufficient condition causes. *Methodology*, *6*, 59–70.

Krause, M. S. (2011a). What are the fundamental facts of a comparison of two treatments' outcomes? *Psychotherapy*, *48*, 234–236.

Krause, M. S. (2011b). Statistical significance testing and clinical trials. *Psychotherapy*, *48*, 217–222.

Krause, M. S., & Howard, K. I. (1999). "Between-group psychotherapy outcome research and basic science" revisited. *Journal of Clinical Psychology*, *55*, 159–169.

Krause, M. S., & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology*, *59*, 751–766.

Krause, M. S., Howard, K. I., & Lutz, W. (1998). Exploring individual change. *Journal of Consulting and Clinical Psychology*, *66*, 838–845.

Krause, M. S., & Lutz, W. (2009). What should be used for baselines against which to compare treatments' effectiveness? *Psychotherapy Research*, *19*, 358–367.

Krause, M. S., Lutz, W., & Boehnke, J. R. (2011). The role of sampling in clinical trial design. *Psychotherapy Research*, *21*, 243 – 251.

Krause, M. S., Lutz, W., & Saunders, S. M. (2007). Empirically certified treatments or therapists: The issue of separability. *Psychotherapy: Theory, Research, Practice, Training*, *44*, 347–353.

Krueger, R. F., & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment*, *13*, 140–151.

Kubinger, K. D., & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell. *Diagnostica*, *53*, 131–143.

Kuder, G., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lai, J.-S., Crane, P., & Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Quality of Life Research, 15*, 1179–1190.

Laireiter, A.-R., & Vogel, H. (Eds.). (1998). *Qualitätssicherung in der Psychotherapie und psychosozialen Versorgung: Ein Werkstattbuch*. Tübingen: DGVT Verlag.

Lambert, M. J. (2001). Psychotherapy outcome and quality improvement: Introduction to the special section on patient-focused research. *Journal of Consulting and Clinical Psychology, 69*, 147–149.

Lambert, M. J. (2005). Emerging methods for providing clinicians with timely feedback on treatment effectiveness: An introduction. *Journal of Clinical Psychology, 61*, 141–144.

Lambert, M. J. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17*, 1–14.

Lambert, M. J. (2012). Helping clinicians to use and learn from research-based systems: The OQ-analyst. *Psychotherapy, 49*, 109–114.

Lambert, M. J., & Baldwin, S. A. (2009). Some observations on studying therapists instead of treatment packages. *Clinical Psychology: Science and Practice, 16*, 82–85.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training, 38*, 357–361.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy, 3*, 249–258.

Lambert, M. J., & Cattani, K. (2012). Practice-friendly research review: Collaboration in routine care. *Journal of Clinical Psychology: In Session, 68*, 209–220.

Lambert, M. J., Hannöver, W., Nisslmüller, K., Richard, M., & Kordy, H. (2002). Fragebogen zum Ergebnis von Psychotherapie: Zur Reliabilität und Validität der deutschen Übersetzung des Outcome Questionnaire 45.2 (OQ-45.2). *Zeitschrift für Klinische Psychologie und Psychotherapie, 31*, 40–46.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*, 159–172.

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th edition, pp. 139–193). New York: Wiley.

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research, 19*, 493–501.

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy*, *48*, 72–79.

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology & Psychotherapy*, *9*, 149–164.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, *10*, 288–301.

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research*, *11*, 49–68.

Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: a replication. *Clinical Psychology & Psychotherapy*, *9*, 91–103.

Langdridge, D., & Hagger-Johnson, G. (2009). *Introduction to research methods and data analysis in psychology* (2nd edition). Harlow: Pearson.

Lauer, G. (1998). Die Lebensqualitätsdimension in der Qualitätssicherung. In A.-R. Laireiter & H. Vogel (Eds.), *Qualitätssicherung in der Psychotherapie und psychosozialen Versorgung: Ein Werkstattbuch* (pp. 575–591). Tübingen: DGVT Verlag.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston, MA: Houghton Mifflin.

Lee Duckworth, A., Steen, T. A., & Seligman, M. E. P. (2005). Positive psychology in clinical practice. *Annual Review of Clinical Psychology*, *1*, 629–651.

Lehr, D., Hillert, A., Schmitz, E., & Sosnowsky, N. (2008). Screening depressiver Störungen mittels Allgemeiner Depressions-Skala (ADS-K) und State-Trait Depressions Scales (STDS-T). *Diagnostica*, *54*, 61–70.

Levitt, H. M., & Piazza-Bonin, E. (2011). Therapists' and clients' significant experiences underlying psychotherapy discourse. *Psychotherapy Research, 21*, 70–85.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*, 53 –70.

Lilienfeld, S. O., Garb, H. N., & Wood, J. M. (2011). Unsolved questions concerning the effectiveness of psychological assessment as a therapeutic intervention: Comment on Poston and Hanson (2010). *Psychological Assessment, 23*, 1047–1055.

Linacre, J. M. (1994). Sample size and item calibration or person measure stability. *Rasch Measurement Transactions, 7*, 328.

Linacre, J. M. (1999). Category disordering vs. step (threshold) disordering. *Rasch Measurement Transactions, 13*, 675.

Linacre, J. M. (2007a). A user's guide to WINSTEPS/MINISTEPS Rasch-model computer programs. Chicago, IL.

Linacre, J. M. (2007b). How to simulate Rasch data. *Rasch Measurement Transactions, 21*, 1125.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48*, 1181–1209.

Locke, B. D., Buzolitz, J. S., Lei, P.-W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., et al. (2011). Development of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58*, 97–109.

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P.-W., Hayes, J. A., Castonguay, L. G., et al. (2012). Development and initial validation of the Counseling Center Assessment of Psychological Symptoms–34. *Measurement and Evaluation in Counseling and Development, 45*, 151 –169.

Long, J. S. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Los Angeles, CA: Sage.

Longino, H. (1990). *Science as social knowledge*. Princeton, NJ: Princeton University Press.

Lord Darzi. (2008). *High quality care for all: NHS next stage review final report*. Department of Health.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research, 43*, 592–620.

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies: Is it true that "everyone has won and all must have prizes"? *Archives of General Psychiatry, 32*, 995–1008.

Luce, D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology, 46*, 1–26.

Lueger, R. J. (1995). A phase model of psychotherapy outcome. *Psychotherapeut, 40*, 267–278.

Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. *Journal of Consulting and Clinical Psychology*, *69*, 150–158.

Luhmann, M. (2010). *R für Einsteiger: Einführung in die Statistiksoftware für die Sozialwissenschaften*. Weinheim: Beltz.

Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology*, *66*, 400–410.

Lüttinger, P., & Riede, T. (1997). Der Mikrozensus: Amtliche Daten für die Sozialforschung. *ZUMA Nachrichten*, *21*(41), 19–43.

Lutz, W. (1997). *Evaluation eines Qualitätssicherungsprogrammes in der Psychotherapie*. Regensburg: S. Roederer.

Lutz, W. (2002). Patient-focused psychotherapy research and individual treatment progress as scientific groundwork for an empirical based clinical practice. *Psychotherapy Research*, *12*, 251–273.

Lutz, W. (2010). Was ist Psychotherapie? - Grundlagen und Modelle. In W. Lutz (Ed.), *Lehrbuch Psychotherapie* (pp. 25–47). Bern: Huber.

Lutz, W. (2011). Strengthening psychotherapy outcome: Neue Ansätze in Qualitätssicherung, Evaluation und Versorgungsforschung. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *40*, 221–223.

Lutz, W., & Bittermann, A. (2010). Wie, wann und warum verändern sich Menschen in der Psychotherapie? Forschung zu integrativen und allgemeinen Ansätzen in der Psychotherapie. *Psychotherapie im Dialog*, *11*, 80–84.

Lutz, W., & Böhnke, J. R. (2008). Der "Fragebogen zur Evaluation von Psychotherapieverläufen" (FEP-2): Validierung und Manual. *Trierer Psychologische Berichte*, *35*(3). Retrieved December 13, 2012, from http://www.uni-trier.de/fileadmin/fb1/PSY/tripsyberichte/2008_35_3.pdf

Lutz, W., & Böhnke, J. R. (2010). Psychotherapieforschung: Verläufe, Prozesse, Ergebnisse und Qualitätssicherung. In W. Lutz (Ed.), *Lehrbuch Psychotherapie* (pp. 49–69). Bern: Huber.

Lutz, W., & Böhnke, J. R. (2012). Evaluation klinisch-psychologischer Interventionen. In W. Lutz, U. Stangier, F. Petermann, & A. Maercker (Eds.), *Klinischen Psychologie - Intervention und Beratung* (Band 2, pp. 71–92). Göttingen: Hogrefe.

Lutz, W., Böhnke, J. R., & Köck, K. (2011). Lending an ear to feedback systems: Evaluation of recovery and non-response in psychotherapy in a German outpatient setting. *Community Mental Health Journal*, *47*, 311–317.

Lutz, W., Böhnke, J. R., Köck, K., & Bittermann, A. (2011). Diagnostik und psychometrische Verlaufsrückmeldungen im Rahmen eines Modellprojektes zur Qualitätssicherung in der ambulanten Psychotherapie. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *40*, 283–297.

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C., et al. (in press). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*. online first.

Lutz, W., Ehrlich, T., & Zaunmüller, L. (2010). Richtungen und Verfahren der Psychotherapie im Überblick 2: Neuere Positionen und Entwicklungen im Verständnis von Psychotherapie in Forschung und Praxis. In W. Lutz (Ed.), *Lehrbuch Psychotherapie* (pp. 151–171). Bern: Huber.

Lutz, W., & Grawe, K. (2007). Psychotherapieforschung: Grundlagen, Konzepte und neue Trends. In B. Strauss, F. Caspar, & F. Hohagen (Eds.), *Lehrbuch der Psychotherapie* (pp. 727–768). Göttingen: Hogrefe Verlag.

Lutz, W., Köck, K., & Böhnke, J. R. (2009). Die Wirkung von Rückmeldesystemen aus ambulanten Settings: Das Modellvorhaben zur Psychotherapie der Techniker Krankenkasse und Wege in die stationäre Praxis. *Klinische Verhaltensmedizin und Rehabilitation*, *84*, 118–125.

Lutz, W., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schürch, E., & Stulz, N. (2006). The probability of treatment success, failure and duration — What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology & Psychotherapy*, *13*, 223–232.

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., et al. (2005). Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *Journal of Consulting and Clinical Psychology*, *73*, 904–913.

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology*, *54*, 32–39.

Lutz, W., Lowry, J., Kopta, S. M., Einstein, D. A., & Howard, K. I. (2001). Prediction of dose–response relations based on patient characteristics. *Journal of Clinical Psychology*, *57*, 889–900.

Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology*, *67*, 571–577.

Lutz, W., Martinovich, Z., Howard, K. I., & Leon, S. C. (2002). Outcomes management, expected treatment response, and severity-adjusted provider profiling in outpatient psychotherapy. *Journal of Clinical Psychology*, *58*, 1291–1304.

Lutz, W., Mocanu, S., & Weinmann-Lutz, B. (2010). Differentielle Indikation: Patienten- und Therapeutenmerkmale. In W. Lutz (Ed.), *Lehrbuch Psychotherapie* (pp. 89–104). Bern: Huber.

Lutz, W., Rafaeli, E., Howard, K. I., & Martinovich, Z. (2002). Adaptive modeling of progress in outpatient psychotherapy. *Psychotherapy Research, 12*, 427–443.

Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., et al. (2006). Empirically and clinically useful decision making in psychotherapy: Differential predictions with treatment response models. *Psychological Assessment, 18*, 133–141.

Lutz, W., Schürch, E., Stulz, N., Böhnke, J. R., Schöttke, H., Rogner, J., & Wiedl, K. H. (2009). Entwicklung und psychometrische Kennwerte des Fragebogens zur Evaluation von Psychotherapieverläufen (FEP). *Diagnostica*, *55*, 106–116.

Lutz, W., Stulz, N., & Köck, K. (2009). Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *Journal of affective disorders*, *118*, 60–68.

Lutz, W., Stulz, N., Martinovich, Z., Leon, S. C., & Saunders, S. M. (2009). Methodological background of decision rules and feedback tools for outcomes management in psychotherapy. *Psychotherapy Research*, *19*, 502 – 510.

Lutz, W., Stulz, N., Smart, D. W., & Lambert, M. J. (2007). Die Identifikation früher Veränderungsmuster in der ambulanten Psychotherapie. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *36*, 93–104.

Lutz, W., Tholen, S., Kosfelder, J., Grawe, K., & Schulte, D. (2005). Zur Entwicklung von Entscheidungsregeln in der Psychotherapie: Die Validierung von Vorhersagemodellen mit einer sequenzanalytischen Methode. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *34*, 165–175.

Lutz, W., Tholen, S., Kosfelder, J., Tschitsaz, A., Schürch, E., & Stulz, N. (2005). Evaluation und störungsspezifische Rückmeldung des therapeutischen Fortschritts in der Psychotherapie. *Verhaltenstherapie*, *15*, 168–175.

Lutz, W., Tholen, S., Schürch, E., & Berking, M. (2006). Die Entwicklung, Validierung und Reliabilität von Kurzformen gängiger psychometrischer Instrumente zur Evaluation des therapeutischen Fortschritts in der Psychotherapie und Psychiatrie. *Diagnostica*, *52*, 11–25.

Lutz, W., & Tschitsaz, A. (2007). Plötzliche Gewinne und Verluste im Behandlungsverlauf von Angststörungen, depressiven und komorbiden Störungen. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *36*, 298–308.

Lutz, W., Wittmann, W., Böhnke, J., Rubel, J., & Steffanowski, A. (2012). Zu den Ergebnissen des Modellprojektes der Techniker-Krankenkasse zum Qualitätsmonitoring in der ambulanten Psychotherapie aus Sicht des wissenschaftlichen Evaluationsteams. *Psychotherapie Psychosomatik Medizinische Psychologie*, *62*, 413–417.

Maercker, A. (2011). Vom Symptom zur Diagnose: Allgemeine Grundlagen und Beispiele. In F. Petermann, A. Maercker, W. Lutz, & U. Stangier (Eds.), *Klinische Psychologie - Grundlagen* (Band 1, pp. 157–175). Göttingen: Hogrefe.

Mair, P., & Hatzinger, R. (2007a). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9).

Mair, P., & Hatzinger, R. (2007b). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*, 26–43.

Margison, F. R., Barkham, M., Evans, C., McGrath, G., Mellor-Clark, J., Audin, K., & Connell, J. (2000). Measurement and psychotherapy: Evidence-based practice and practice-based evidence. *British Journal of Psychiatry*, *177*, 123–130.

Margraf, J., & Ehlers, A. (2007). *BAI - Beck Angst Inventar: Manual*. Frankfurt am Main: Harcourt Test Services.

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 68*, 438–450.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149 – 174.

Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. In A. T. Panter & S. K. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 159–184). New York: Routledge.

Maydeu-Olivares, A., & McArdle, J. J. (Eds.). (2005). *Contemporary psychometrics*. Mahwah, NJ: Lawrence Erlbaum.

McAllister, M., Dunn, G., Payne, K., Davies, L., & Todd, C. (2012). Patient empowerment: The need to consider it as a measurable patient-reported outcome for chronic conditions. *BMC Health Services Research*, *12*(1), 157.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McFall, R. M. (2005). Theory and utility-key themes in evidence-based assessment: Comment on the special section. *Psychological Assessment, 17*, 312–323.

McGlinchey, J. B., & Zimmerman, M. (2007). Examining a dimensional representation of depression and anxiety disorders' comorbidity in psychiatric outpatients with item response modeling. *Journal of Abnormal Psychology, 116*, 464–474.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meier, S. (1997). Nomothetic item selection rules for tests of psychological interventions. *Psychotherapy Research, 7*, 419–427.

Meijer, R. R., De Vries, R. M., & Van Bruggen, V. (2011). An evaluation of the Brief Symptom Inventory–18 using item response theory: Which items are most strongly related to psychological distress? *Psychological Assessment, 23*, 193–202.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187 –194.

Meiser, T., Stern, E., & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution rasch models for the analysis of repeated observations. *Methods of Psychological Research - online, 3*, 75–93.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165.

Miller, C., & Evans, B. B. (2004). Ethical issues in assessment. In M. Hersen (Ed.), *Psychological assessment in clinical practice: A pragmatic guide* (pp. 21–32). New York: Brunner-Routledge.

Miller, S. D., Duncan, B. L., Brown, J., Sorrell, R., & Chalk, M. B. (2006). Using formal client feedback to improve retention and outcome: Making ongoing, real-time assessment feasible. *Journal of Brief Therapy, 5*, 5–22.

Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome management system. *Journal of Clinical Psychology, 61*, 199–208.

Mockenhaupt, A. (2009). *Grawes Inkonsistenz im Zusammenhang mit der Depression und der Angststörung: Eine Untersuchung anhand der motivationalen Inkongruenz.* Trier: Unveröffentlichte Diplomarbeit.

Mohr, D. C. (1995). Negative outcome in psychotherapy: A critical review. *Clinical Psychology: Science and Practice, 2*, 1–27.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, This time forever. *Measurement, 2*, 201 – 218.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*, 112–117.

Möller, H.-J. (2012). How close is evidence to truth in evidence-based treatment of mental disorders? *European Archives of Psychiatry and Clinical Neuroscience*, *262*, 277-289.

Moosbrugger, H., & Kelava, A. (Eds.). (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.

Muenchen, R. A. (n.d.). r4stats.com: R info for SAS, SPSS, and Stata Users. Retrieved December 13, 2012, from http://r4stats.com/popularity

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Murawski, M. M., & Miederhoff, P. A. (1998). On the generalizability of statistical expressions of health related quality of life instrument responsiveness: A data synthesis. *Quality of Life Research*, *7*, 11–22.

Murphy, D. L., Dodd, B. G., & Vaughn, B. K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, *34*, 424 –437.

Muthén, B. O. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–332). Washington, DC: American Psychological Association.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park, CA: Sage.

Muthén, B. O., & Brown, H. C. (2009). Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling. *Statistics in Medicine*, *28*, 3363–3385.

Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus user's guide* (3rd edition). Los Angeles, CA: Muhén & Muthén.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, *4*, 139–157.

Newman, M. G., Castonguay, L. G., Borkovec, T. D., & Molnar, C. (2004). Integrative therapy for generalized anxiety disorder. In R. G. Heimberg, C. L. Turk, & D. S. Mennin (Eds.), *Generalized anxiety disorder: Advances in research and practice* (pp. 320–350). New York: Guilford Press.

Newnham, E. A., & Page, A. C. (2010). Bridging the gap between best evidence and best practice in mental health. *Clinical Psychology Review*, *30*, 127–142.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

Norcross, J. C., & Lambert, M. J. (2011). Psychotherapy relationships that work II. *Psychotherapy*, *48*, 4–8.

Norcross, J. C., & Wampold, B. E. (2011). Evidence-based therapy relationships: Research conclusions and clinical practices. *Psychotherapy*, *48*, 98–102.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*, 582-592.

Nuevo, R., Leighton, C., Dunn, G., Dowrick, C., Lehtinen, V., Dalgard, O. S., et al. (2010). Impact of severity and type of depression on quality of life in cases identified in the community. *Psychological Medicine*, *40*, 2069–2077.

Nunnally, J. C. (1978). *Psychometric theory* (2nd edition). New York: McGraw-Hill.

Nussbeck, F. W., Eid, M., & Geiser, C. (2010). Mischverteilungsmodelle. In H. Holling & B. Schmitz (Eds.), *Handbuch Statistik, Methoden und Evaluation* (pp. 562–568). Göttingen: Hogrefe.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, *14*, 535–569.

Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy*, *10*, 361–373.

Ollendick, T. H., & King, N. J. (2004). Empirically supported treatments for children and adolescents: Advances toward evidence-based practice. In P. M. Barrett & T. H. Ollendick (Eds.), *Handbook of interventions that work with children and adolescents: Prevention and treatment.* (pp. 3–25). Chichester: John Wiley & Sons.

Orlando Edelen, M., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, *12*, 354–359.

Orlinsky, D. E. (2009). The "Generic Model of Psychotherapy" after 25 years: Evolution of a research-based metatheory. *Journal of Psychotherapy Integration*, *19*, 319–339.

Orlinsky, D. E., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy – noch einmal. In A. E. Bergin & S. E. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th edition, pp. 270–376). New York: Wiley.

Orlinsky, D. E., & Howard, K. I. (1986). Process and outcome in psychotherapy. In S. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (3rd edition, pp. 311–381). New York: Wiley.

Orlinsky, D. E., Ronnestad, M. H., & Willutzky, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. In M. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th edition, pp. 307–389). New York: John Wiley & Sons.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd edition, Vol. 161). Los Angeles, CA: Sage.

Overall, J., & Gorham, D. (1962). The brief psychiatric rating scale. *Psychological Reports, 10*, 799–812.

Padberg, T. (2012). Warum lesen Psychotherapeuten keine Forschungsliteratur? *Psychotherapeutenjournal, 11*(1), 10–17.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*, 1–18.

Patel, V., & Riley, A. (2007). Linking data to decision-making: Applying qualitative data analysis methods and software to identify mechanisms for using outcomes data. *The Journal of Behavioral Health Services and Research*, *34*, 459–474.

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*, 109–118.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*, 150–151.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295–312.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (pp. 125–167). Amsterdam: Elsevier.

Percevic, R., Lambert, M. J., & Kordy, H. (2006). What is the predictive value of responses to psychotherapy for its future course? Empirical explorations and consequences for outcome monitoring. *Psychotherapy Research, 16*, 364–373.

Perrez, M. (2005). Wissenschaftstheoretische Grundlagen: Klinisch-psychologische Intervention. In M. Perrez & U. Baumann (Eds.), *Lehrbuch Klinische Psychologie - Psychotherapie* (3. vollst. überarb. Auflage, pp. 68–88). Bern: Huber.

Persons, J. B., & Silberschatz, G. (1998). Are results of randomized controlled trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology, 66*, 126–135.

Petermann, F., & Müller, J. M. (2001). *Clinical psychology and single-case evidence: A practical approach to treatment planning and evaluation*. Chichester: Wiley.

Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 249–263). New York: Springer.

Piechotta, B. (2008). *PsyQM: Qualitätsmanagement für psychotherapeutische Praxen*. Heidelberg: Springer.

Pina, J. A. L., & Montesinos, M. D. H. (2005). Fitting Rasch model using appropriateness measure statistics. *The Spanish Journal of Psychology, 8*, 100–110.

Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment*, *22*, 203–212.

Presaghi, F., & Desimoni, M. (2010). *random.polychor.pa: A Parallel Analysis With Polychoric Correlation Matrices.* Retrieved December 13, 2012, from http://CRAN.R-project.org/package=random.polychor.pa

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes*, *1*(1), 27.

Psychometric Society. (1979). Announcements: Publication policy regarding monte carlo studies. *Psychometrika*, *44*, 133–134.

Puschner, B., & Kordy, H. (2010). Mit Transparenz und Ergebnisorientierung zur Optimierung der psychotherapeutischen Versorgung: Eine Studie zur Evaluation ambulanter Psychotherapie. *Psychotherapie Psychosomatik Medizinische Psychologie*, *60*, 350–357.

Puschner, B., Schöfer, D., Knaup, C., & Becker, T. (2009). Outcome management in in-patient psychiatric care. *Acta Psychiatrica Scandinavica*, *120*, 308–319.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*, 103–121.

Quilty, L. C., Zhang, K. A., & Bagby, R. M. (2010). The latent symptom structure of the Beck Depression Inventory–II in outpatients with major depression. *Psychological Assessment*, *22*, 603–608.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, *60*, 20–43.

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Retrieved December 13, 2012, from http://www.R-project.org

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Retrieved December 13, 2012, from http://www.R-project.org

Raîche, G., Blais, J.-G., & Magis, D. (2007). Adaptive estimators of trait level in adaptive testing: Some proposals. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Presented at the 2009 GMAC (R) Conference on CAT. Retrieved December 13, 2012, from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat07raiche1%20.pdf

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321–333). Presented at the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press. Retrieved December 13, 2012, from http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.bsmsp/1200512895

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd edition). Thousand Oaks, CA: Sage.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, *52*, 127–141.

Reese, R. J., Toland, M. D., & Hopkins, N. B. (2011). Replicating and extending the good-enough level model of change: Considering session frequency. *Psychotherapy Research*, *21*, 608–619.

Reese, R. J., Toland, M. D., Slone, N. C., & Norsworthy, L. A. (2010). Effect of client feedback on couple psychotherapy outcomes. *Psychotherapy: Theory, Research, Practice, Training*, *47*, 616–630.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. In P. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods of practice* (2nd edition, pp. 55–73). Oxford: Oxford Univ. Press.

Reininghaus, U., McCabe, R., Burns, T., Croudace, T., & Priebe, S. (2011). Measuring patients' views: A bifactor model of distinct patient-reported outcomes in psychosis. *Psychological Medicine*, *41*, 277–289.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*, 228–238.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*, 93–103.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559.

Reise, S. P., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19–31.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*, 164–184.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annu. Rev. Clin. Psychol.*, *5*, 27–48.

Revelle, W. (2010). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved December 13, 2012, from http://personality-project.org/r/psych.manual.pdf

Revicki, D., Gnanasakthy, A., & Weinfurt, K. (2007). Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: The PRO Evidence Dossier. *Quality of Life Research*, *16*, 717–723.

Riley, W., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E., & Cella, D. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Quality of Life Research*, *19*, 1311–1321.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Analysis. *Journal of Statistical Software*, *17*(5).

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1–12.

Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of clinical epidemiology*, *61*, 17–33.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 25–42). New York: Springer.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. überarb. und erw. Auflage). Bern: Huber.

Rounsaville, B. J., Carroll, K. M., & Onken, L. S. (2001). A stage model of behavioral therapies research: Getting started and moving from stage I. *Clinical Psychology: Science and Practice*, *8*, 133–142.

Ruberg, S. J., Chen, L., & Wang, Y. (2010). The mean does not mean as much anymore: Finding sub-groups for tailored therapeutics. *Clinical Trials*, *7*, 574–583.

Ruholl, S. (2007). *Selbstwirksamkeit als Indikator für psychische Störungen: Status und Verlauf*. Aachen: RWTH Aachen. Retrieved December 13, 2012, from http://darwin.bth.rwth-aachen.de/opus3/volltexte/2008/2243/pdf/Ruholl_Sabine.pdf

Ruscio, J., Brown, T. A., & Meron Ruscio, A. (2009). A taxometric investigation of DSM-IV major depression in a large outpatient sample. *Assessment*, *16*, 127 –144.

Salvi, G., Leese, M., & Slade, M. (2005). Routine use of mental health outcome assesssments: Choosing the measure. *British Journal of Psychiatry*, *186*, 146–152.

Sapyta, J., Riemer, M., & Bickman, L. (2005). Feedback to clinicians: Theory, research, and practice. *Journal of Clinical Psychology*, *61*, 145–153.

Saß, H., Wittchen, H.-U., Zaudig, M., & Houben, I. (2003). *Diagnostische Kriterien DSM IV TR*. Göttingen: Hogrefe.

Scheidt, C., Brockmann, J., Caspar, F., Rudolf, G., Stangier, U., & Vogel, H. (2012). Das Modellprojekt der Techniker-Krankenkasse: Eine Kommentierung der Ergebnisse aus der Sicht des wissenschaftlichen Projektbeirates. *Psychotherapie Psychosomatik Medizinische Psychologie*, *62*, 405–412.

Schiepek, G., & Strunk, G. (2010). The identification of critical fluctuations and phase transitions in short term and coarse-grained time series—a method for the real-time monitoring of human change processes. *Biological Cybernetics*, *102*, 197–207.

Schiepek, G., Zellweger, A., Kronberger, H., Aichhorn, W., & Leeb, W. (2011). Psychotherapie. In G. Schiepek (Ed.), *Neurobiologie der Psychotherapie* (2. Auflage, pp. 568–592). Stuttgart: Schattauer.

Schindler, A. C., Hiller, W., & Witthöft, M. (2011). Benchmarking of cognitive-behavioral therapy for depression in efficacy and effectiveness studies—How do exclusion criteria affect treatment outcome? *Psychotherapy Research*, *21*, 644–657.

Schindler, A., & Hiller, W. (2010). Therapieeffekte und Responseraten bei unipolar depressiven Patienten einer verhaltenstherapeutischen Hochschulambulanz. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *39*, 107–115.

Schmitz, B. (2000). Auf der Suche nach dem verlorenen Individuum: Vier Theoreme zur Aggregation von Prozessen. *Psychologische Rundschau*, *51*, 83–92.

Schöttke, H., Sembill, A., Eversmann, J., Waldorf, M., & Lange, J. (2011). Therapieziele in der ambulanten kognitiv-verhaltenstherapeutischen oder psychodynamischen Psychotherapie – notwendig oder irrelevant? *Zeitschrift für Klinische Psychologie und Psychotherapie, 40*, 257–266.

Schuck, P., & Zwingmann, C. (2003). The "smallest real difference" as a measure of sensitivity to change: A critical analysis. *International Journal of Rehabilitation Research, 26*, 85-91.

Schulte, D. (1993). Wie soll Therapieerfolg gemessen werden? *Zeitschrift für Klinische Psychologie*, *22*, 374–393.

Schulte, D. (2007). New law for psychological psychotherapists in Germany: Its rules and consequences. *Mental Health and Learning Disabilities Research and Practice*, 219–230.

Schulte, D., & Eifert, G. H. (2002). What to do when manuals fail? The dual model of psychotherapy. *Clinical Psychology: Science and Practice*, *9*, 312–328.

Schürch, E., Lutz, W., & Böhnke, J. R. (2009). Identifikation abweichender Antwortmuster im "Fragebogen zur Evaluation von Psychotherapieverläufen" mithilfe der Rasch-Analyse. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *38*, 135–144.

Schwartz, F. W., Siegrist, J., Von Troschke, J., & Schlaud, M. (2003). Gesundheit und Krankheit in der Bevölkerung. In F. W. Schwartz, B. Badura, R. Busse, R. Leidl, H. Raspe, J. Siegrist, & U. Walter (Eds.), *Public Health: Gesundheit Und Gesundheitswesen*. München: Urban & Fischer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Schwarzer, R. (1994). Optimism, vulnerability, and self-beliefs as health-related cognitions: A systematic overview. *Psychology & Health*, *9*, 161–180.

Seidenstücker, G. (1995). Indikation und Entscheidung. In R. S. Jäger & F. Petermann (Eds.), *Psychologische Diagnostik* (3rd edition). Weinheim: PVU.

Seidenstücker, G., & Baumann, U. (1987). Multimodale Diagnostik als Standard in der Klinischen Psychologie. *Diagnostica*, *33*, 243–258.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, *55*, 5–14.

Sexton, T. L., & Kelley, S. D. (2010). Finding the common core: Evidence-based practices, clinically relevant evidence, and core mechanisms of change. *Administration and Policy in Mental Health and Mental Health Services Research*, *37*, 81–88.

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin*, *92*, 581–604.

Shapiro, J. P. (2009). Integrating outcome research and clinical reasoning in psychotherapy planning. *Professional Psychology: Research and Practice*, *40*, 46–53.

Sharp, C., Goodyer, I., & Croudace, T. J. (2006). The Short Mood and Feelings Questionnaire (SMFQ): A unidimensional Item Response Theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *Journal of Abnormal Child Psychology*, *34*, 365–377.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, *78*, 298–311.

Simon, W., Lambert, M. J., Harris, M. W., Busath, G., & Vazquez, A. (2012). Providing patient progress information and clinical support tools to therapists: Effects on patients at risk of treatment failure. *Psychotherapy Research*, *22*, 638-647.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 553–573.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*, 102–111.

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. B. (2009). On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychological Assessment*, *21*, 272–284.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Sonnanburg, K. (1996). Meaningful measurement in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *33*, 160–170.

Sperry, L., Brill, P. L., Howard, K. I., & Grissom, G. R. (1996). *Treatment outcomes in psychotherapy and psychiatric interventions*. New York: Brunner/Mazel.

Spielmans, G. I., Masters, K. S., & Lambert, M. J. (2006). A comparison of rational versus empirical methods in the prediction of psychotherapy outcome. *Clinical Psychology & Psychotherapy*, *13*, 202–214.

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology*, *86*, 943–953.

StataCorp. (2011). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.

Steck, P. (1997). Psychologische Testverfahren in der Praxis: Ergebnisse einer Umfrage unter Testanwendern. *Diagnostica*, *43*, 267–284.

Steffanowski, A., Kramer, D., Fembacher, A., Glahn, E. M., Bruckmayer, E., Von Heymann, F., et al. (2011). Praxisübergreifende Dokumentation der Ergebnisqualität ambulanter Psychotherapie in Bayern. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *40*, 267–282.

Steinkamp, D., & Schulte, D. (2008). *AMBOS-FG: Auswertung von Testergebnissen mit der Forscher- und Grafikmaske*. Bochum: Ruhr-Universität.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, S. E., Hynan, M. T., & Allen, M. (2000). A meta-analysis of common factor and specific treatment effects across the outcome domains of the phase model of psychotherapy. *Clinical Psychology: Science and Practice*, *7*, 273–290.

Stewart, R. E., & Chambless, D. L. (2007). Does psychotherapy research inform treatment decisions in private practice? *Journal of Clinical Psychology*, *63*, 267–281.

Stewart, R. E., & Chambless, D. L. (2009). Cognitive–behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology*, *77*, 595–606.

Stewart, R. E., & Chambless, D. L. (2010). Interesting practitioners in training in empirically supported treatments: Research reviews versus case studies. *Journal of Clinical Psychology*, *66*, 73–95.

Stewart, R. E., Stirman, S. W., & Chambless, D. L. (2012). A qualitative investigation of practicing psychologists' attitudes toward research-informed practice: Implications for dissemination strategies. *Professional Psychology: Research and Practice*, *43*, 100–109.

Steyer, R., & Eid, M. (2001). *Messen und Testen: Mit Übungen und Lösungen*. Berlin: Springer.

Stieglitz, M. (2003). Psychodiagnostische Verfahren. In M. Härter, H. W. Linster, & R.-D. Stieglitz (Eds.), *Qualitätsmanagement in der Psychotherapie: Grundlagen, Methoden und Anwendung* (pp. 97–117). Göttingen: Hogrefe.

Stiles, W. B., Shapiro, D. A., & Elliott, R. (1986). Are all psychotherapies equivalent. *American Psychologist*, *41*, 165–180.

Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Rothman, A. (2005). Can the randomized controlled trial literature generalize to nonrandomized patients? *Journal of Consulting and Clinical Psychology*, *73*, 127–135.

Strauss, B., & Kaechele, H. (1998). The writing on the wall—Comments on the current discussion about empirically validated treatments in Germany. *Psychotherapy Research, 8,* 158–170.

Stricker, G. (2006). The local clinical scientist, evidence-based practice, and personality assessment. *Journal of Personality Assessment, 86*, 4–9.

Stricker, L. J. (2000). Using just noticeable differences to interpret test scores. *Psychological Methods, 5*, 415–424.

Strupp, H. H. (1963). The outcome problem in psychotherapy revisited. *Psychotherapy: Theory, Research & Practice, 1*, 1–13.

Strupp, H. H., Horowitz, L. M., & Lambert, M. J. (Eds.). (1997). *Measuring patient changes in mood, anxiety, and personality disorders: Towards a core battery*. Washington, DC: American Psychological Association.

Stulz, N., Gallop, R., Lutz, W., Wrenn, G. L., & Crits-Christoph, P. (2010). Examining differential effects of psychosocial treatments for cocaine dependence: An application of latent trajectory analyses. *Drug and Alcohol Dependence, 106*, 164–172.

Stulz, N., & Lutz, W. (2007). Multidimensional patterns of change in outpatient psychotherapy: The phase model revisited. *Journal of Clinical Psychology, 63*, 817–833.

Stulz, N., Lutz, W., Leach, C., Lucock, M., & Barkham, M. (2007). Shapes of early change in psychotherapy under routine outpatient conditions. *Journal of Consulting and Clinical Psychology, 75*, 864–874.

Tang, T. Z., & DeRubeis, R. J. (1999a). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 67*, 894–904.

Tang, T. Z., & DeRubeis, R. J. (1999b). Reconsidering rapid early response in cognitive behavioral therapy for depression. *Clinical Psychology: Science and Practice, 6*, 283–288.

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if Differential Item Functioning makes a difference. *Rasch Measurement Transactions, 20*, 1082–1084.

Thase, M. E., Larsen, K. G., & Kennedy, S. H. (2011). Assessing the "true" effect of active antidepressant therapy v. placebo in major depressive disorder: use of a mixture model. *The British Journal of Psychiatry, 199*, 501 –507.

The Future Vision Coalition. (2009). *A future vision for mental health*. England & Wales. Retrieved December 13, 2012, from http://www.newvisionformentalhealth.org.uk/A_future_vision_for_mental_health.pdf

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the brief symptom inventory. *Psychological Assessment, 24*, 101–113.

Thompson, M., Thompson, L., Gallagher-Thompson, D., & Alto, P. (1995). Linear and nonlinear changes in mood between psychotherapy sessions: Implications for treatment outcome and relapse risk. *Psychotherapy Research*, *5*, 327–336.

Thyer, B. A., & Pignotti, M. (2011). Evidence-based practices do not exist. *Clinical Social Work Journal*, *39*, 328–333.

Tingey, R. C., Lambert, M. J., Burlingame, G. M., & Hansen, N. B. (1996). Assessing clinical significance: proposed extension to method. *Psychotherapy Research*, *6*, 109–123.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology*, *56*, 271–288.

Tran, U. S., Walter, T., & Remmel, A. (2012). Faktoren psychosozialer Beeinträchtigung. *Diagnostica*, *58*, 75–86.

Tschitsaz-Stucki, A., & Lutz, W. (2009). Identifikation und Aufklärung von Veränderungssprüngen im individuellen Psychotherapieverlauf. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *38*, 13–23.

Tura, B., Nicolau, M., & Oliveira, R. (2008). *DiagnosisMed*. Retrieved December 13, 2012, from http://cran.r-project.org/src/contrib/Archive/DiagnosisMed/

Tyron, W. W. (1991). *Activity measurement in psychology and medicine*. New York: Plenum.

U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, & U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. (2006). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: draft guidance. *Health and Quality of Life Outcomes*, *4*(1), 79.

Uher, R., Muthén, B., Souery, D., Mors, O., Jaracz, J., Placentino, A., et al. (2010). Trajectories of change in depression severity during treatment with antidepressants. *Psychological Medicine, 40*, 1367–1377.

Ullrich, M., Horz, H., Schnotz, W., McElvany, N., Schroeder, S., & Baumert, J. (2012). CML oder MML? Effekte der Kalibrierungsmethode und des Untersuchungsdesigns. Presented at the 48. Kongress der Deutschen Gesellschaft für Psychologie, Bielefeld.

Valderas, J., Kotzeva, A., Espallargues, M., Guyatt, G., Ferrans, C., Halyard, M., et al. (2008a). The impact of measuring patient-reported outcomes in clinical practice: A systematic review of the literature. *Quality of Life Research*, *17*, 179–193.

Valderas, J., Kotzeva, A., Espallargues, M., Guyatt, G., Ferrans, C., Halyard, M., et al. (2008b). A relevant study was missed in our systematic review on the impact of patient-reported outcomes in clinical practice. *Quality of Life Research*, *17*, 965–966.

Van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1988). Consistency of Rasch model parameter estimation: A simulation study. *Applied Psychological Measurement*, *12*, 307–313.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford: Clarendon Press.

Van Rijn, P. W., & Molenaar, P. C. (2005). Logistic models for single subject time series. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 125–145). Mahwah, NJ: Lawrence Erlbaum.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment*, *74*, 242–261.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In *Applied latent class analysis* (pp. 89–106). Cambridge: Cambridge University Press.

Vissers, W. (2010). *The measurement of remoralization: An extension of contemporary psychotherapy outcome research*. Amsterdam: GVO drukkers en vormgevers B.V. | Ponsen & Looijen. Retrieved December 13, 2012, from http://dare.ubn.kun.nl/bitstream/2066/82606/1/82606.pdf

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd edition). Mahwah, NJ: Lawrence Erlbaum Associates: Hillsdale, New Jersey.

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics*, *35*, 5–25.

Waldron, S., Moscovitz, S., Lundin, J., Helm, F. L., Jemerin, J., & Gorman, B. (2011). Evaluating the outcomes of psychotherapies: The Personality Health Index. *Psychoanalytic Psychology*, *28*, 363–388.

Walker, J., Böhnke, J. R., Cerny, T., & Strasser, F. (2010). Development of symptom assessments utilising item response theory and computer-adaptive testing—A practical method based on a systematic review. *Critical reviews in Oncology/Hematology*, *73*, 47–67.

Wampold, B. E. (2001). *The great psychotherapy debate. Models, methods, and findings.* New Jersey: Lawrence Erlbaum Associates.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. -n. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "all must have prizes". *Psychological Bulletin, 122*, 203–215.

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*, 376 – 404.

Warren, J., Nelson, P., & Burlingame, G. (2009). Identifying youth at risk for treatment failure in outpatient community mental health services. *Journal of Child and Family Studies, 18*, 690–701.

Washington, T. (2010). *The effects of using clinical support tools to prevent treatment failure*. Provo, UT: Brigham Young University. Retrieved November 1, 2011, from http://contentdm.lib.byu.edu/ETD/image/etd4178.pdf

Westen, D., Novotny, C. M., & Thompson–Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130*, 631–661.

Westmeyer, H. (1979). Die rationale Rekonstruktion einiger Aspekte psychologischer Praxis. In H. Albert & K. H. Stapf (Eds.), *Theorie und Erfahrung: Beiträge zur Grundlagenproblematik der Sozialwissenschaft* (pp. 139–161). Stuttgart: Klett-Cotta.

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*, 59–68.

Williams, D., & Levitt, H. M. (2008). Clients' experiences of difference with therapists: Sustaining faith in psychotherapy. *Psychotherapy Research, 18*, 256 – 270.

Willke, R. J., Burke, L. B., & Erickson, P. (2004). Measuring treatment impact: A review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Controlled Clinical Trials, 25*, 535–552.

Willse, J. T. (2011). Mixture rasch models with joint maximum likelihood estimation. *Educational and Psychological Measurement, 71*, 5–19.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.

Wirtz, M., & Böcker, M. (2007). Eigenschaften und Nutzen des Rasch-Modells in der klinischen Diagnostik. *Rehabilitation*, *46*, 238-245.

Wittchen, H.-U., & Jacobi, F. (2001). Die Versorgungssituation psychischer Störungen in Deutschland Eine klinisch-epidemiologische Abschätzung anhand des Bundes-Gesundheitssurveys 1998. *Bundes-gesundheitsblatt*, *44*, 993–1000.

Wittmann, W. W., Lutz, W., Steffanowski, A., Kriz, D., Glahn, E. M., Völkle, M. C., et al. (2011). *Qualitätsmonitoring in der ambulanten Psychotherapie: Abschlussbericht*. Hamburg: Techniker Krankenkasse. Retrieved December 13, 2012, from http://www.tk.de/tk/050-publikationen/studien-und-umfragen/qualitaetsmonitoring-in-der-psychotherapie-mai-2011/341996

Wood-Dauphinee, S. (1999). Assessing quality of life in clinical research: From where have we come and where are we going? *Journal of Clinical Epidemiology*, *52*, 355–363.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1–27.

Wulff, E. (1988/2004). Erich Wulf neu gelesen: Sozialpsychiatrischer Krankheitsbegriff? *Sozialpsychiatrische Informationen*, *34*, 4–10.

Yamazaki, S., Fukuhara, S., & Green, J. (2005). Usefulness of five-item and three-item Mental Health Inventories to screen for depressive symptoms in the general population of Japan. *Health and Quality of Life Outcomes*, *3*, 48.

Yen, W. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*, 275–291.

Zayas, L. H., Drake, B., & Jonson-Reid, M. (2011). Overrating or dismissing the value of evidence-based practice: Consequences for clinical practice. *Clinical Social Work Journal*, *39*, 400–405.

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in organizational and social sciences* (pp. 37–59). New York: Routledge.

7. Abbildungsverzeichnis

8. Tabellenverzeichnis

Erklärung

Hiermit erkläre ich, dass die vorliegende Dissertationsschrift von mir selbständig angefertigt wurde und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet wurden. Zudem wurde die Arbeit an keiner anderen Universität zur Erlangung eines akademischen Grades eingereicht.

_____

Trier, 14.12.2012
Jan R. Böhnke