Inhalt

Zufriedenheitseinschätzungen stationär-psychiatrisch behandelter Kinder im BesT-K: Analysen zu Dimensionalität und Antwortmustern

Ferdinand Keller

1 Einleitung

1.1 Bedeutung der Zufriedenheitsforschung und Kapitelüberblick

Behandlungszufriedenheit ist zu einem wichtigen Bestandteil guter stationärer Behandlung geworden, und dies nicht nur in der sogenannten somatischen Medizin, sondern gerade auch in der Therapie psychischer Erkrankungen. Meist wird das Konzept der Patientenzufriedenheit dabei in vorrangiger Verbindung zur Qualitätssicherung (QS) oder als Aufgabe des Qualitätsmanagements (QM) gesehen. Bezogen auf die Kinder- und Jugendpsychiatrie und Psychotherapie (KJPP) bedeutet dies, dass die Therapieeinrichtungen die Zufriedenheit der behandelten Kinder und Jugendlichen sowie ihrer Eltern/Erziehungsberechtigten mit verschiedenen Aspekten des Behandlungsprozesses (Prozessqualität) und der räumlichen und personellen Ausstattung (Strukturqualität) erheben. Die Einschätzungen der »Kunden« dienen dann als Qualitätsindikator und als Rückmeldung für das »Angebot« der Klinik, das entsprechend für Verbesserungen in den verschiedenen QS-Bereichen genutzt werden kann. Diese eher betriebswirtschaftliche Sichtweise wurde immer wieder kritisiert (z. B. Dür et al. 2000) und ein mindestens ebenso wichtiger Punkt für den Einsatz solcher Erhebungen liegt darin, dass Behandlungszufriedenheit auch auf therapeutische Prozesse einen wichtigen Einfluss haben kann. Brown et al. (2014) verweisen besonders auf die Motivation von Kindern und ihren Familien, in einer Behandlung positiv und engagiert mitzuarbeiten und nicht vorzeitig abzubrechen; die Verbesserung von Behandlungszufriedenheit kann daher helfen, die Abbrecher-Rate zu verringern. Viefhaus et al. (2019) betonen, dass Zufriedenheitserhebungen ein direktes Feedback an die Therapeuten ermöglichen, was die Qualität der Behandlung verbessern hilft.

Während über die Vorteile von Zufriedenheitserhebungen somit weitgehend Einigkeit besteht, ist die theoretische Fundierung des Zufriedenheitskonzeptes nach wie vor nicht geklärt (Bowling et al. 2012; Gill & White 2009) und viele Erhebungsinstrumente sind auf einer ad hoc-Basis entwickelt und unzureichend psychometrisch überprüft worden. Des Weiteren sind sie oft nur auf die eindimensionale Erfassung der globalen Zufriedenheit angelegt, wie das aus acht Items bestehende Client Satisfaction Questionnaire (CSQ; deutsch: ZUF-8), das vor allem in vielen Reha-Kliniken routinemäßig eingesetzt wird (Kriz et al. 2008).

Im Folgenden wird ein kurzer Überblick zur Entwicklung von Messinstrumenten im Rahmen der KJPP gegeben und ein eigener Fragebogen zur Erhebung der Behandlungseinschätzung stationärer Therapie (BesT) vorgestellt. Eine multi-methodale psychometrische Analyse des BesT in der Kinderversion (BesT-K) ist zentraler Gegenstand dieses Kapitels. Sie erfolgt zum einen mit Methoden der

klassischen Testtheorie und mit Modellen aus der Item-Response-Theory (IRT), zum anderen wird als komplementäres Verfahren eine Latent-Class-Analyse (LCA) verwendet, mit der personenorientiert Subgruppen mit spezifischen Antwortmustern gesucht werden. Die methodische Vorgehensweise ist in den Kapiteln 1.3 und 2.3 weiter ausgeführt und die Kapitel 3 und 4 enthalten die jeweiligen Ergebnisse und die Diskussion.

1.2 Zufriedenheitsforschung in der Kinder- und Jugendpsychiatrie

Angelehnt an die angelsächsische consumer-satisfaction-Forschung im (sozial-)psychiatrischen Bereich (Überblick z. B. in Lebow 1983) wurden auch im deutschsprachigen Raum bereits in den 1970er Jahren für Erwachsene Zufriedenheitseinschätzungen erhoben (Gruyters & Priebe 1994; Leimkühler & Müller 1996). In der Kinder- und Jugendpsychiatrie gab es in den 1980er Jahren erste Studien (z. B. Kammerer 1989), in denen Erhebungsinstrumente für Jugendliche und Eltern entwickelt worden sind. Ein psychometrisch gut überprüftes Inventar legten dann Mattejat & Remschmidt (1998) mit ihren Fragebogen zur Bewertung der Behandlung (FBB) vor. Der FBB existiert in drei Versionen für Patient, Eltern und Therapeut, ist jedoch erst für Jugendliche ab ca. 13 Jahren geeignet.

Für Kinder im Alter von ca. 8–12/13 Jahren sind auch international bisher nur wenige Fragebogen verfügbar, z. B. ein kurzes Instrument von Kaplan et al. (2001). In Großbritannien wurde für den vorwiegend ambulanten Bereich des »Mental health care« der »Experience of Service Questionnaire« (ESQ) in Versionen für Kinder, Jugendliche und Eltern entwickelt und im nationalen Gesundheitssystem breit eingesetzt (Brown et al. 2014).

Im Rahmen einer eigenen, zu Beginn der 2000er Jahre begonnen Entwicklung von Fragebögen zur Erfassung von Behandlungszufriedenheit in der KJPP, die in ihren Inhalten möglichst parallel angelegt und Versionen für Kinder, Jugendliche und Eltern umfassen sollte, entstanden die Fragebögen zur Behandlungseinschätzung stationärer Therapie (BesT) (vgl. zusammenfassend in Keller et al. 2018). Die Kinderversion wurde bereits von Keller et al. (2004) publiziert und ist ein inzwischen häufig eingesetzter Fragebogen.

1.3 Methodische Überlegungen und Fragestellungen

Behandlungszufriedenheit ist ein komplexes Konstrukt, das sich aus vielen verschiedenen Aspekten zusammensetzt. Entsprechend muss bei der Entwicklung psychometrisch fundierter Fragebögen die Inhaltsvalidität gewährleistet sein, indem Items aus den verschiedenen Bereichen gesammelt werden (Rost 2004). Als relevant erachtete Bereiche der Zufriedenheit sind dabei neben den bekannten Dimensionen wie therapeutische Beziehung und »Hotelqualität« auch neuere wichtige Aspekte wie Partizipation, Informationsvermittlung und Klima auf der Station. Auf diese Weise fließen (in der Sicht des Forschers) viele Dimensionen der Zufriedenheit in die Konstruktion des Fragebogens ein. Neben der Sammlung der verschiedenen Aspekte des interessierenden Konstrukts Behandlungszufriedenheit gilt es auch, diese in geeignete Fragen und Antwortformate zu überführen, die im vorliegenden Falle auch für Kinder verständlich sind.

In der ersten Publikation des BesT-K (Keller et al. 2004) war die Stichprobe zu klein, um aussagekräftige Faktorenanalysen zur Abschätzung der Dimensionen und zur Gewichtung der einzelnen Items vornehmen zu können. Mit der nun vorliegenden großen Stichprobe soll diesen psychometrischen Fragen weiter nachgegangen werden, insbesondere, ob sich die bei der Konstruktion berücksichtigten unterschiedlichen Aspekte von Zufriedenheit auch in der Einschätzung der Kinder so differenziert wiederfinden lassen, oder ob sie sich auf einige wenige Hauptdimensionen reduzieren.

Die Untersuchungsziele lassen sich in drei Schwerpunkten zusammenfassen:

1. Wie viele und welche Dimensionen lassen sich im BesT-K unterscheiden? Dazu werden Faktorenanalysen angewendet und das Vorgehen folgt insgesamt den Konventionen der klassischen Testtheorie. Es bietet damit auch einen orientierenden Rahmen für die nachfolgende vertiefte Evaluation mittels IRT-Modellen.

2. Wie ist die Itemqualität auf der Ebene der Itemkategorien, d. h. sind die Kategorien ausreichend genutzt und insbesondere: Sind sie aufsteigend geordnet? Wie reliabel misst der Test über verschiedene Ausprägungen der Zufriedenheit hinweg? Dazu werden IRT-Modelle (Rasch-Modell, 2-parametrige Modelle) eingesetzt.

3. Ergänzend zu diesen psychometrischen Analysen, die sich jeweils auf die gesamte Stichprobe beziehen, soll der Frage nachgegangen werden: Lassen sich qualitativ verschiedene Antwortmuster in Form von Subgruppen von Kindern identifizieren? Das Vorhandensein inhaltlich abweichender Antwortmuster würde die Bildung eines einfachen Gesamtwertes in Frage stellen, denn derselbe Summenwert könnte dann z. B. durch das Ankreuzen hoher Zufriedenheitswerte in verschiedenen Itemgruppen zustande kommen. Auch eine Gruppe von Kindern, die die Items/-kategorien nicht oder falsch versteht und deshalb »seltsame« Antwortmuster aufweist (»Unskalierbare«), wäre denkbar (vgl. Keller 2012). Diese Analysen erfolgen mit dem personenorientierten Ansatz der Latent-Class-Analyse (LCA).

2 Methodik

2.1 Procedere und Stichprobe

Die Stichprobe wurde über einen Zeitraum von drei Jahren hinweg in insgesamt neun Kliniken für Kinder- und Jugendpsychiatrie erhoben. Es wurden Bögen für Kinder, für Jugendliche und für Eltern im Rahmen der QM-Maßnahmen in den Kliniken verteilt (vgl. Keller et al. 2018). Für die hier vorgelegten Analysen wurden nur die Kinderbögen verwendet, die entweder auf den Kinderstationen der Kliniken oder in den Tageskliniken bei den Kindern erhoben wurden, die zwischen acht Jahren und maximal 13 Jahren alt waren. Die Bögen sollten jeweils kurz vor Entlassung, am besten am Entlassungstag, ausgefüllt, in anonymisierter Form in einen vorbereiteten Umschlag gesteckt und verschlossen abgegeben werden. Über die Anzahl ausgegebener Bögen pro Klinik/Station und den jeweiligen Rücklauf waren keine verlässlichen Angaben verfügbar. Eine positive Bewertung der Ethikkommission der Universität Ulm lag vor.

Abbildung 1: *Beispielitem aus dem BesT-Kinderbogen*

Insgesamt standen $n = 1661$ Fragebögen zur Verfügung (31 % weiblich, mittleres Alter = 10.6 Jahre, SD = 2.2). Davon waren $n = 796$ komplett ausgefüllt. Zu beachten ist dabei aber, dass manchmal gar nicht alle Items beantwortbar sind (strukturell fehlende Werte). Ein Kind in tagesklinischer Behandlung kann beispielsweise keine Beurteilung des Zimmers oder betreffend den Wochenendausgang vornehmen. Zentral für die Vollständigkeit der Angaben ist insbesondere Item 2 (Medikamentenaufklärung), die nur vorgenommen wird, wenn das Kind auch ein Medikament erhält. Mit der Beantwortung dieses Items wird zudem eine spezielle Subgruppe definiert, die sich möglicherweise systematisch von den Kindern ohne Medikamentengabe unterscheidet. Nachdem dieses Item nur gering mit den übrigen Items korrelierte (und entsprechend auch eine niedrige Trennschärfe aufwies), wurde es bei den IRT-Analysen weggelassen. Dadurch erhöhte sich die Anzahl vollständig ausgefüllter Bögen für die Items 1 und 3–18 auf $n = 1149$. Vergleichende Analysen wurden aber jeweils vorgenommen und ergaben keine besonderen Unterschiede. Bei den Faktorenanalysen mit M*plus* wurden alle Items und Personen verwendet, da eine missing data-Schätzung vorgenommen wird (s. Kapitel 2.3).

2.2 Erhebungsinstrumente

Der BesT-Kinderbogen besteht aus 18 Items mit fünf Antwortkategorien, die jeweils mit Text erläutert sind. Ein Beispiel ist in Abbildung 1 zu sehen; der Wertebereich liegt zwischen 1 und 5 und hohe Werte bedeuten hohe Zufriedenheit (z. B. Item 1: 5 = »Die Station hat mir sehr gut gefallen«). Die Inhalte aller Items sind in Tabelle 1 aufgeführt. Die Inhalte der erfragten Items wurden gemäß einer bundesweiten Umfrage an KJPP-Abteilungen zu den dort eventuell verwendeten Befragungsinstrumenten und vor allem gemäß den Angaben von Kindern in Fokusgruppen zusammengestellt (vgl. Keller et al. 2004). Die Antworten wurden in anschließenden freien Interviews vertieft und zudem wurden verschiedene nach Meinung der Behandler bedeutsame Aspekte einbezogen. Insgesamt wurden so Items nicht nur zu den bekannten Dimensionen wie therapeutische Beziehung und globale Zufriedenheit, sondern auch zur Beziehung zu Mitarbeiter_innen im Pflege- und Erziehungsdienst (PED), Zugang zu Informationen und Partizipation während des Aufenthalts, Zufriedenheit mit den Regeln, Klima auf der Station und Bewertung der Umgebung (»Hotel-/Serviceleistungen«) aufgenommen.

2.3 Statistischer Ansatz und Analysemethoden

Die deskriptiven Statistiken und die interne Konsistenz wurden mit SAS 9.4 berechnet. Exploratorische Faktorenanalysen (EFA) wurden mit M*plus* in der Version 7.4 (Muthén & Muthén 2012) berechnet unter Verwendung des WLSMV-Schätzers, bei dem die Antwortkategorien als ordinal-kategorial behandelt und fehlende Werte geschätzt werden. Zu den Modellvergleichen werden neben inhaltlichen Kriterien häufig verwendete goodness-of-fit-Maße herangezogen: comparative fit index (CFI), Tucker-Lewis index (TLI), und root mean square error of approximation (RMSEA) (vgl. z. B. Reinecke 2014). Kriterien für akzeptablen Fit sind CFI und TLI > .900 und RMSEA < .080, für guten Fit CFI und TLI > .950 und RMSEA < .060 (Hu & Bentler 1999; Kline 2005).

Für die Analysen zur Fragestellung 2 (Itemqualität und Nutzung der Itemkategorien) mit IRT-Modellen wurde zunächst das Partial Credit Modell (PCM-Masters, 1982) für die Modellschätzung verwendet. Da das PCM mit seinen strengen Annahmen (z. B. gleiche Diskrimination in allen Items) nur selten bei empirischen Daten passt, wurde noch die direkte Erweiterung des PCM, das Generalized Partial Credit Model (GPCM) von Muraki (1992) verwendet, bei dem für jedes Item noch ein Diskriminationsparameter geschätzt wird. Die Modellschätzungen und Ergebnisdarstellungen erfolgten mit dem R-package mirt (Chalmers 2012).

Die Berechnungen der LCAs wurden mit dem Programm LatentGOLD in der Version 4.5 (Vermunt & Magidson 2003) vorgenommen. Nachdem die Anzahl der latenten Klassen selbst kein Modellparameter und damit nicht direkt testbar ist, müssen Lösungen mit unterschiedlicher Klassenzahl miteinander verglichen werden. Zur Bestimmung der optimalen Klassenzahl werden jeweils die Informationskriterien BIC, CAIC und AIC (Rost 2004) herangezogen, jedoch kombiniert mit inhaltlichen Überlegungen zur Herausarbeitung und Interpretation der Klassenprofile. Mögliche Geschlechtsunterschiede wurden durch Einbeziehung der Variablen Geschlecht als aktive (= in das LCA-Modell aufgenommene Variable) und als inaktive Kovariate (= Ausgabe der Geschlechtsanteile pro Klasse gemäß Zuordnungswahrscheinlichkeit) geprüft.

Für einen aussagekräftigen Einsatz der LCA mussten die 18 Items reduziert werden. Dazu wurden die vorangegangenen Analysen zur Itemqualität ebenso herangezogen wie die Diversität der abgefragten Inhalte. Unter Berücksichtigung der Diversität (alle unterschiedenen Bereiche sollten vertreten sein) und Itemqualität (bei Items, die aus demselben Bereich kommen, z. B. therapeutische Beziehung, wurden diejenigen mit der besten Iteminformationsfunktion verwendet) wurden schließlich zehn Items des Fragebogens verwendet, um die LCA handhabbar zu halten. Neben dieser den ganzen Fragebogen abdeckenden LCA sollte auch noch inhaltlich vertieft auf die Items zu »Personen« (Beziehung zu Therapeuten, Betreuer und andere Kinder) und ihre differenzierte Wahrnehmung durch die Kinder eingegangen werden. Bei Betrachtung der Mittelwerte hatte sich z. B. die interessante Unterscheidung in »Therapeut hat Problem verstanden« und »konnte Therapeut alles erzählen, auch wenn es mir unangenehm/peinlich war« angedeutet.

Die Items in den beiden Auswertungsbereichen der LCA umfassen:

1. wichtige Items über den ganzen Fragebogen hinweg: Items 1, 4, 5, 9, 10, 12, 13, 15, 16 und 18 (Iteminhalte siehe Tabelle 1).

Tabelle 1: *Deskriptive Statistiken der BesT-Kinder-Items*

Item-Inhalt	Mittelwert	Standardabweichung	r_{it}	N
1. Zeit auf Station gefallen	3.82	1.06	.57	1661
2. Aufklärung Medikamente	3.94	1.31	.14	1099
3. wohl gefühlt (bei Untersuchung)	3.72	1.17	.48	1629
4. Therapeut_in Problem verstanden	4.19	1.01	.51	1617
5. Therapeut_in alles erzählen	3.75	1.23	.43	1626
6. Zeit gehabt	3.96	0.98	.47	1624
7. Problem jetzt weg	3.74	1.16	.31	1638
8. Personal ansprechbar	4.26	0.93	.52	1652
9. Betreuer_innen Problem verstanden	3.94	1.03	.59	1639
10. Betreuer_innen nett	4.11	0.98	.58	1658
11. Wochenendbeurlaubung	4.06	1.24	.43	1395
12. Regelung Garten, Fernsehen	3.91	1.17	.54	1427
13. Ausgangsregelung	3.99	1.17	.53	1438
14. Zimmer gefallen	3.81	1.26	.54	1490
15. Station gefallen	4.08	1.05	.65	1637
16. Essen geschmeckt	3.38	1.19	.44	1647
17. zusammen sein mit anderen Kindern	4.00	1.02	.38	1650
18. mit anderen Kindern verstanden	3.81	0.94	.32	1649

Die Trennschärfen (r_{it}) basieren auf n = 796 mit vollständigen Daten.

2. Items speziell zu »Personen«: Items 4–6 (Therapeut_in), 8–10 (Betreuungspersonal) und 18 (mit anderen Kindern verstanden).

Zu beiden Auswertungsbereichen wurde jeweils noch das Item 7 (Problem jetzt weg) hinzugefügt, um den Einfluss von Therapieerfolg abschätzen zu können.

3 Ergebnisse

3.1 Inhalte der Items und allgemeine Ergebnisse zum Fragebogen

Die Mittelwerte der einzelnen Items, die in Tabelle 1 abgebildet sind, zeigen eine insgesamt recht hohe Zufriedenheit der Kinder. Besonders der/die Therapeut_in, aber auch die Betreuer_innen werden gut bewertet (unter Betreuer_innen fallen in der KJPP sowohl das Pflegepersonal wie auch z. B. Heilerziehungspfleger_innen und andere Personen aus dem sozialpädagogischen und Erziehungsdienst).

Eine psychometrische Analyse zeigt eine gute interne Konsistenz der Gesamtskala mit Cronbach α = .86. Lässt man Item 2 weg, erhöht sich das N auf 1149, doch ist die interne Konsistenz mit α = .87 fast identisch und ebenso unterscheiden sich die Trennschärfen kaum. Dies spricht für eine gute Vergleichbarkeit und rechtfertigt das Weglassen von Item 2 für die Berechnung der IRT-Modelle.

Tabelle 2: *Fit-Indizes von exploratorischen Faktorenanalysen mit 1–4 Faktoren (n = 1641)*

Modell	χ^2	df	CFI	TLI	RMSEA (90 %-CI)
1 Faktor	1618.28	135	.894	.880	.082 (.078 – .085)
2 Faktoren	1059.06	118	.933	.913	.070 (.066 – .074)
3 Faktoren	481.25	102	.973	.959	.048 (.043 – .052)
4 Faktoren	217.05	87	.991	.984	.030 (.025 – .035)

3.2 Anzahl der Faktoren und Itemzuordnungen

Eine exploratorische Faktorenanalyse (EFA) des Kinderbogens ergibt die in Tabelle 2 dargestellten Fit-Maße für Lösungen mit 1–4 Faktoren. Die eindimensionale Lösung passt nicht gut und die Lösung mit zwei Faktoren ergibt eine Überlegenheit gegenüber einer eindimensionalen Sichtweise. Die Anpassungsgüte einer 2-Faktorenlösung ist akzeptabel, wird aber deutlich verbessert in einer Lösung mit drei Faktoren. Die 4-Faktorenlösung würde eine weitere Verbesserung im goodness-of-fit (GoF) erzielen, allerdings wird der vierte Faktor dann allein durch die beiden Items 17 und 18 bestimmt. Bereits in der 3-Faktorenlösung zeigt sich, dass der dritte Faktor durch diese beiden Items dominiert ist und die übrigen Items teilweise ähnlich hoch wie auf dem zweiten Faktor laden. Angesichts der deutlichen Steigerung im GoF gegenüber der 2-Faktorenlösung soll aber der dritte Faktor beibehalten werden. Alternativ könnte sonst auch eine 2-Faktorenlösung erwogen werden, die zusätzlich eine Residualkorrelation zwischen den Items 17 und 18 enthält.

Inhaltlich unterscheiden die Kinder gemäß den Ladungen (vgl. Tabelle 3) im Wesentlichen zwei Bereiche, die sich charakterisieren lassen als Therapeutische Beziehung (Therapeut_innen und Betreuungspersonal) und Regeln/Umgebung. Diese Schwerpunkte zeigen sich sowohl in der 2- als auch in der 3-Faktorenlösung. In letzterer zeichnet sich auf den Faktoren 2 und 3 dabei eine Aufteilung in Betreuer/Umgebung und Regeln/Umgebung mit überlappenden (Doppel-)Ladungen ab. Die Korrelation der beiden Faktoren in der 2-Faktorenlösung liegt bei $r = .68$. Bei der 3-Faktorenlösung liegen die Korrelationen der drei Faktoren bei: F1 und F2: $r = .50$; F1 und F3: $r = .53$; F2 und F3: $r = .38$.

Die Korrelationen zwischen den Faktoren in der EFA sind nicht übermäßig hoch, aber doch sehr ausgeprägt, insbesondere in der 2-Faktorenlösung. Sie legen die Betrachtung eines bifactor-Modells nahe, bei dem ein Generalfaktor Zufriedenheit modelliert wird sowie die beiden spezifischen Faktoren Therapeutische Beziehung und Regeln/Umgebung. Dieses Modell, das konfirmatorisch vorgegeben und geschätzt wurde, weist insgesamt eine gute Anpassung auf ($\chi^2 = 586.43$, df = 121, $p < .001$, CFI = .967, TLI = .958, RMSEA = .048) und es zeigt sich, dass alle Items mit Ausnahme der Medikamentenaufklärung moderat bis hoch auf dem Generalfaktor laden (vgl. Tabelle 4). Weiterhin ergibt sich, dass die Items zur Therapeutischen Beziehung zusätzlich noch einen spezifischen Faktor ausbilden. Dagegen laden die Items aus dem Bereich Umgebung (»Zimmer gefallen« u. ä.) hoch auf dem Generalfaktor und darüber hinaus nicht mehr auf dem spezifischen Faktor Regeln/Umgebung. Verwunderlich ist, dass die verbleibenden Items des spezifischen Faktors 2 (Auskommen mit anderen Kindern und die Regelungen) negativ korrelieren, wenn der gemeinsame Varianzanteil in Form des Generalfaktors herausgezogen ist.

Tabelle 3: *Faktorladungen der BesT-Kinder-Items in einer explorativen Faktorenanalyse mit zwei und mit drei Faktoren (n = 1641)*

	2-faktoriell		3-faktoriell		
Itemnummer und -inhalt	F1	F2	F1	F2	F3
1. Zeit auf Station gefallen		.60		.37	.43
2. Aufklärung Medikamente	.34	−.15	.30		
3. wohl gefühlt (bei Untersuchung)	.29	.32	.33	.25	
4. Therapeut Problem verstanden	.70		.70		
5. Therapeut alles erzählen	.61		.68		
6. Therapeut Zeit gehabt	.58		.51		
7. Problem jetzt weg	.21	.19	.29	.19	
8. Personal ansprechbar	.64		.51		.31
9. Betreuer Problem verstanden	.65	.16	.54		.28
10. Betreuer nett	.47	.29	.35		.37
11. Wochenendbeurlaubung		.55	.19	.50	
12. Regelung Garten, Fernsehen		.73		.69	
13. Ausgangsregelung		.63	.20	.54	
14. Zimmer gefallen		.78		.52	.43
15. Station gefallen		.79		.52	.48
16. Essen geschmeckt		.56		.37	.30
17. zusammen sein mit anderen	.24	.29			.67
18. mit anderen verstanden	.29			−.21	.66

nur Ladungen > .15 sind angezeigt; Ladungen > .50 sind unterstrichen;
F1 = Faktor 1 etc.

In der Gesamtbewertung der faktorenanalytischen Befunde lassen sowohl die überlappenden Ladungen auf den Faktoren 2 und 3 bei den EFA als auch die negativen Ladungen auf dem spezifischen Faktor 2 im Rahmen des bifactor-Modells kein eindeutiges Fazit zu. Inhaltlich könnten die gefundenen Zuordnungen die vermischende Denkwelt der Kinder durchaus reflektieren, aber zuvor sollte der Frage nachgegangen werden, ob diese schwierig zu interpretierenden Befunde nicht darauf zurückzuführen sind, dass die Kinder die Items bzw. die Kategorientexte nicht richtig verstehen bzw. die Kategorien nicht in der erwarteten Weise nutzen. Daher werden im nächsten Kapitel IRT-Modelle verwendet, die eine vertiefte Analyse solcher Fragen, z. B. nach der Ordinalität der Itemkategorien, gestatten.

3.3 Analysen zur Itemqualität mit IRT-Modellen

In Modellen gemäß der Item Response Theory (IRT) wird eine probabilistische Beziehung zwischen einer zugrundeliegenden latenten Dimension (»trait«) und dem Ankreuzen von Item(kategorien) angenommen (für eine zusammenfassende Darstellung siehe z. B. Reckase 2009; Rost 2004). Mit IRT-Modellen kann dabei die Informationsfunktion der einzelnen Items und des gesamten Tests/Fragebogens (»Messpräzision«) geschätzt werden. Auf der Ebene der Itemkategorien kann überprüft werden, ob die Kategorien aufsteigend geordnet sind, was sich anhand der Reihenfolge der Schwellenwerte erkennen lässt (siehe unten). Dabei soll aber betont werden, dass die IRT-Ansätze sich nicht

Tabelle 4: *Faktorladungen der BesT-Kinder-Items in einem konfirmatorischen bifactor-Modell (n = 1641)*

Item-Inhalt	General-Faktor	Spezif. Faktor 1	Spezif. Faktor 2
1. Zeit auf Station gefallen	.72		
2. Aufklärung Medikamente	.10	.22	
3. wohl gefühlt (bei Untersuchung)	.53	.21	
4. Therapeut Problem verstanden	.50	.51	
5. Therapeut alles erzählen	.35	.49	
6. Therapeut Zeit gehabt	.40	.39	
7. Problem jetzt weg	.34	.15	
8. Personal ansprechbar	.56	.42	
9. Betreuer Problem verstanden	.63	.43	
10. Betreuer nett	.63	.29	
11. Wochenendbeurlaubung	.54		−.25
12. Regelung Garten, Fernsehen	.67		−.35
13. Ausgangsregelung	.65		−.24
14. Zimmer gefallen	.69		
15. Station gefallen	.83		
16. Essen geschmeckt	.54		
17. zusammen sein mit anderen	.52		.45
18. mit anderen verstanden	.39		.56

Alle Ladungen signifikant mit $p < .0005$, nur bei Item 2: $p = .003$.

fundamental von Item-Faktoranalysen auf Basis von polytomen ordinalen Items unterscheiden, wie sie zuvor mit dem WLSMV-Ansatz der EFA untersucht wurden, sondern inzwischen gezeigt wurde, dass die beiden Ansätze grundsätzlich äquivalent sind (vgl. Wirth & Edwards 2007).
Beim Vergleich des PCM mit dem GPCM, das zusätzlich einen eigenen Diskriminationsparameter pro Item aufweist, zeigte sich in allen von mirt ausgegebenen Informationskriterien eine bessere Anpassung des GPCM, z. B. ein BIC = 67068.6 für das PCM und ein BIC = 66532.2 für das GPCM. Auch der likelihood-ratio-Test war signifikant mit $p < .0001$. Im Folgenden werden daher die Ergebnisse des GPCM berichtet.

In Abbildung 2 sind die Antwortwahrscheinlichkeiten (Wk) der Items des BesT Kinderbogens dargestellt (das Item 2 zur Medikamentenaufklärung ist weggelassen für diese Analysen). Beispielhaft sei das Item 1 (BesT_K1) erläutert. Ein Kind mit einer sehr geringen Ausprägung des latenten traits (Zufriedenheit) müsste mit einer hohen Wk die Kategorie 1 ankreuzen. Mit zunehmender Ausprägung des traits sinkt diese ab und die Wk, die Kategorie 2 anzukreuzen, sollte steigen. Der Punkt, an dem die Wk für 1 und für 2 gleich hoch ist, wird als Schwellenwert bezeichnet. Bei dem fünfstufigen Antwortformat ergeben sich also vier Schwellenparameter pro Item, die bei Modellgültigkeit im Sinne der Antwortkategorien aufsteigende Werte besitzen sollten, d. h. Schwelle 1 trennt die Kategorien 1 und 2 und sollte daher einen kleineren Wert haben als Schwelle 2, die die Kategorien 2 und 3 trennt. Ein Schwellenparameter gibt also das Ausmaß der Merkmalsausprägung an, ab dem eine Person wahrscheinlich in die nächsthöhere Antwortkategorie fällt (Rost 2004).

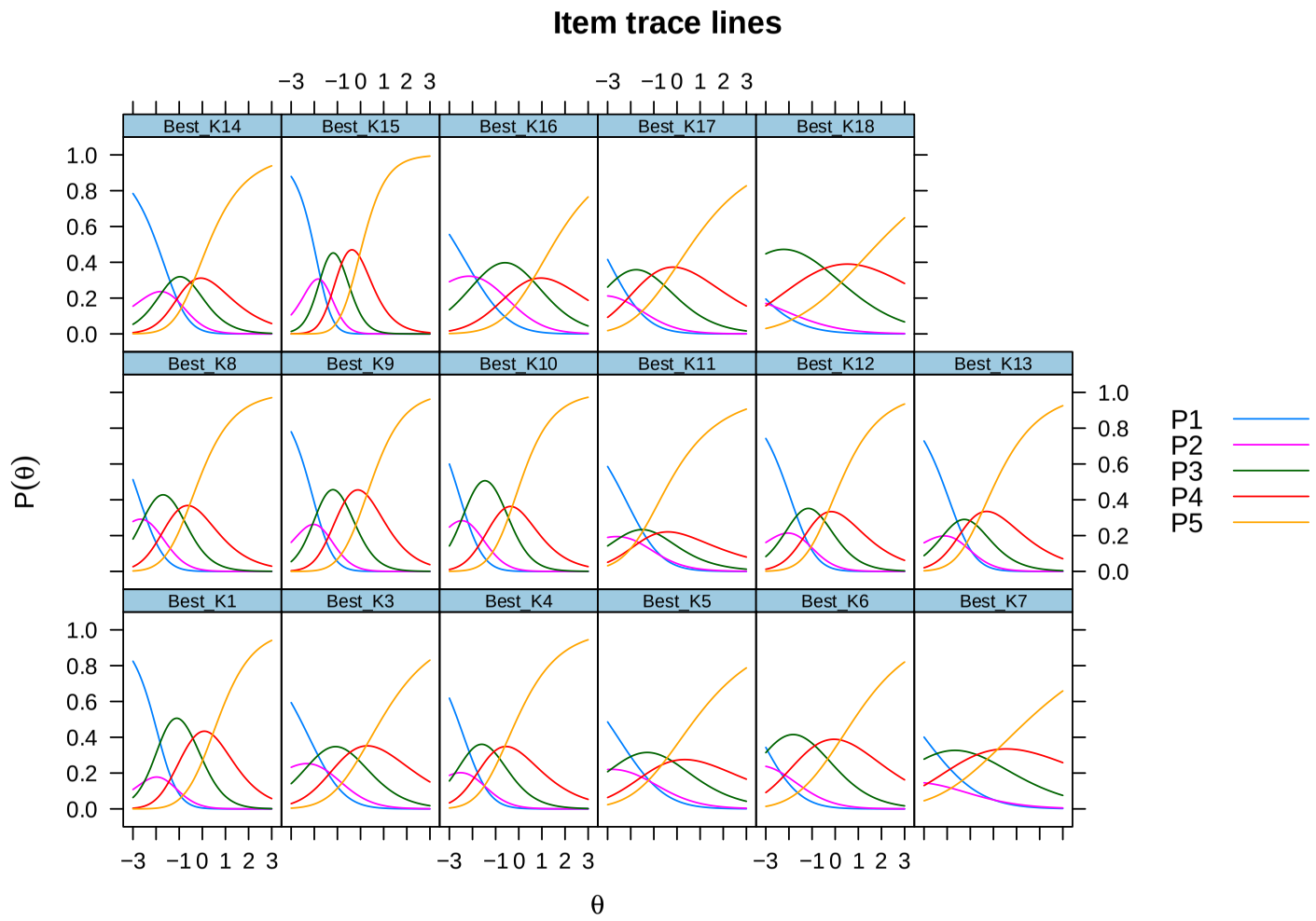

Abbildung 2: *Kategorienspezifische Antwortwahrscheinlichkeiten der einzelnen Items des Zufriedenheitsfragebogens (ohne Item 2), geschätzt mit dem GPCM (P1 beschreibt die Wahrscheinlichkeit, die Kategorie 1 anzukreuzen, P2 die von Kategorie 2 usw.)*

Im GPCM werden die Schwellenparameter für jedes Item gesondert, aber über die Kategorien des Items konstant geschätzt. Im Beispiel sieht man jedoch, dass die Wk für Kategorie 2 an keinem Punkt am höchsten ist (Abbildung 2, erstes Diagramm in Zeile 3); dagegen haben die Kategorien 3 und 4 wieder Bereiche maximaler Wk. Ein Kind mit hoher Zufriedenheit kreuzt mit hoher Wk Kategorie 5 an. Auch in vielen anderen Items weist die Kategorie 2 in keinem Bereich des traits eine höhere Wk als die beiden benachbarten Kategorien auf. Dies spräche für eine Zusammenfassung der Kategorien 1 und 2. Andererseits liegt Kategorie 2 immer im richtigen Bereich des traits und die Reliabilität der Gesamtskala mit fünf Kategorien ist besser als diejenige mit vier Kategorien, weshalb das originale fünfstufige Antwortformat beibehalten wird. Auch die Kategorie 4 hat oft nur einen geringen Anteil maximaler Wk, weist aber nur in zwei Items keinen Bereich maximaler Wk auf (Items 5 und 11) und in zwei weiteren so gut wie keinen (Items 14 und 16). Bei den Items 5 (Therapeut alles erzählen) und Item 16 (Essen geschmeckt) findet sich ein vergleichsweise breiter Bereich für die mittlere Kategorie 3, d. h. diese Items werden vorzugsweise in drei Kategorien beantwortet. Bei Item 11 (Wochenendausgang) hat zusätzlich auch die Kategorie 3 kein ausgeprägtes Maximum, d. h. bei der Einschätzung von Wochenendausgang scheint im Wesentlichen eine dichotome Antworttendenz vorzuliegen. Generell gilt, dass Items, die schon auf dem Generalfaktor der bifactor-Analyse geringe Ladungen hatten, auch eher schlechte Wk-Verläufe aufweisen, z. B. Item 18, während die Items mit hohen Ladungen auch hier gut geordnete Schwellenwerte und Verläufe haben, insbesondere neben Item 1 noch die Items 9 und 15.

Eine Betrachtung der Iteminformationsfunktionen (IIF, hier nicht eigens dargestellt) zeigt ebenfalls und analog zu den Ladungen auf dem Generalfaktor, dass einzelne Items des Fragebogens einen großen Informationsbeitrag leisten, während derjenige anderer Items eher gering ist. Es bleibt jedoch die Frage, wie sich dies auf die Testeigenschaften des Gesamtwertes auswirkt und insbesondere auch, wie gut er in verschiedenen Bereichen des latenten Merkmals misst. Dazu kann die sogenannte Testinformationsfunktion (TIF) herangezogen werden. Sie setzt sich aus der Summe der einzelnen IIF zusammen und ist damit nicht standardisiert. Sie kann jedoch über den Kehrwert der Wurzel aus der TIF in den »standard error of measurement« umgerechnet werden (vgl. Samejima 1994; bei ihr als »standard error of estimation« bezeichnet). Dieser drückt aus, wie groß der Standardschätzfehler bei Prädiktion eines Wertes auf dem latenten Kontinuum wäre. Alternativ lässt sich aus der TIF auch eine gleitende Reliabilität berechnen und im package mirt ausgeben, die die Messgenauigkeit bei den unterschiedlichen Ausprägungen des traits angibt.

In Abbildung 3 ist die gleitende Reliabilität des Gesamtwertes abgetragen. Die Reliabilität ist in der unteren Hälfte des traits (= Unzufriedenheit) hoch und sinkt im mittleren Bereich dann ab, weist aber immer noch gute Werte auf. Lediglich der obere Bereich (also hohe Zufriedenheitswerte) wird nicht mehr so genau erfasst, d. h. Kinder mit hohen Zufriedenheitswerten lassen sich nicht mehr so gut differenzieren.

Als Fazit aus den IRT-Modellen ergibt sich, dass einzelne Items zwar einen geringen Informationsgehalt aufweisen (wie schon bei den Faktorenanalysen ersichtlich), aber die Reliabilität der Gesamtskala sehr gut bis gut ist, jedenfalls über einen großen Bereich des latenten trait hinweg. Noch wichtiger dürfte sein: Die Itemkategorien sind in der vorgesehenen Weise genutzt, wie sich anhand der ka-

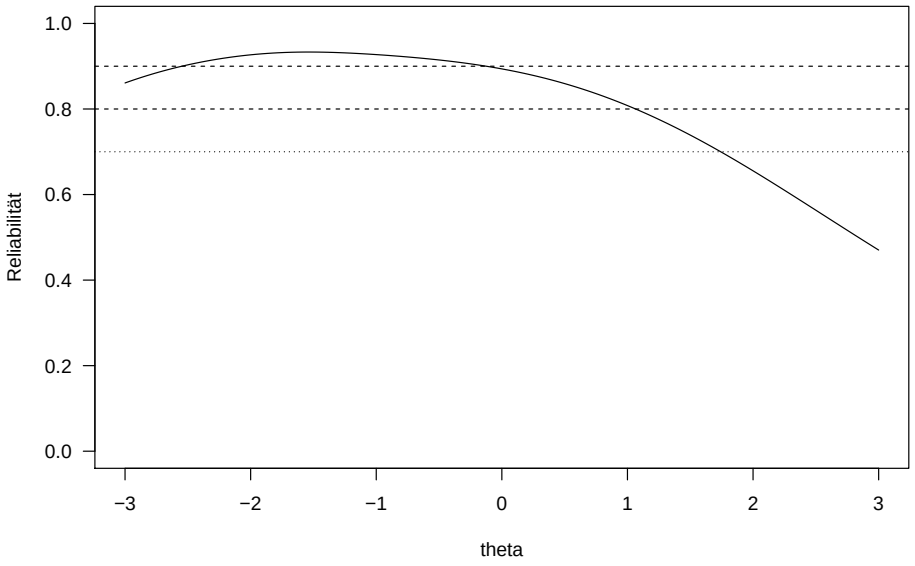

Abbildung 3: *Reliabilität des Gesamtwerts (ohne Item 2) über den latenten trait (Zufriedenheit), geschätzt mit dem GPCM*

tegorienspezifischen Antwort-Wk zeigte, auch wenn in den meisten Items die Kategorie 2 und in wenigen Items die Kategorie 4 nur einen geringen Beitrag leisten. Dieser Befund wurde auch bestätigt in einem nominalen IRT-Modell, bei dem jede Antwortkategorie als nominal behandelt und damit keine Ordinalität der Kategorien vorgegeben wird. Auch in diesem Modell zeigen sich die Kategorien als geordnet in der erwarteten Reihenfolge und die Antwort-Wk sind weitgehend vergleichbar (Ergebnisse nicht dargestellt).

Nachdem sich also keine wesentlichen Hinweise auf Probleme mit den Items bzw. ihren Kategorien im Rahmen eines (unidimensionalen) IRT-Modells zu bestehen scheinen, soll im nächsten Schritt als Alternative zu den Faktorenanalysen der personenorientierte Ansatz in Form der LCA genutzt werden, um eventuell qualitativ abweichende Antwortmuster in Subgruppen zu identifizieren, welche wiederum für die unklaren Befunde zur Faktorstruktur verantwortlich sein könnten.

3.4 Ergebnisse der LCA

3.4.1 LCA für ausgewählte Items des Fragebogens

In Tabelle 5 sind die Informationskriterien für LC-Modelle mit unterschiedlicher Anzahl von Klassen aufgeführt. Gemäß BIC werden fünf latente Klassen als beste Lösung vorgeschlagen und gemäß CAIC würden bereits vier Klassen eine gute Lösung liefern; der AIC hat hingegen noch kein Minimum bei diesen Klassenzahlen, d. h. es könnten noch weitere, über die fünf Klassen hinaus betrachtet werden.

Tabelle 5: *Ausgewählte Items: Vergleich der Modellanpassung für LC-Modelle mit unterschiedlicher Klassenzahl*

Anzahl latenter Klassen	Log-Likelihood	Anzahl Parameter	BIC	CAIC	AIC
3-LC	−17370.9	68	35227.0	35295.0	34877.8
4-LC	−17274.7	80	35120.3	**35200.3**	34709.4
5-LC	−17230.5	92	**35117.4**	35209.4	34645.0
6-LC	−17195.9	104	35134.0	35238.0	34600.0
7-LC	−17160.4	116	35148.6	35264.6	**34552.9**

Nachdem der BIC als Hauptkriterium angesehen werden kann (vgl. Nylund et al. 2007) und die 5-Klassenlösung mit vielen Startwerten stabil repliziert werden konnte, wurde diese Lösung interpretiert. Die Klassenprofile der 5-Klassenlösung sind in Abbildung 4 dargestellt. Die größte Klasse (cl 1) umfasst fast die Hälfte der Kinder und die erwarteten Item-Mittelwerte in dieser Klasse liegen sehr nahe beim Gesamtmittel. Die zweitgrößte Klasse (cl 2) liegt deutlich darunter, folgt aber im Wesentlichen dem Gesamtmuster. Die Klasse 3 (cl 3) ist insgesamt sehr zufrieden mit Therapeuten, Betreuer und den Regelungen und macht lediglich leichte Abstriche beim Essen und dem Auskommen mit den anderen Kindern. Dagegen ist bei der Klasse 4 (cl 4) nur bei Therapeuten und Betreuern eine hohe Einschätzung vorhanden, während sie bei den übrigen Items nahe am Gesamtmittel sind. Die Klasse 5 (cl 5) schließlich ist mit 3.2 % zwar die kleinste Klasse, aber sie ist deutlich unzufrieden mit allen Bereichen. Dies erstreckt sich auch auf eine unterdurchschnittliche Erfolgseinschätzung (»Problem jetzt weg«), während die anderen vier Klassen in diesem Item relativ nahe am Gesamtmittel liegen, d. h. die Zufriedenheit hängt bei ihnen nur gering vom Therapieerfolg ab. Diese unzufriedene Klasse taucht in gleicher Weise auch bei der 4-Klassenlösung auf. Gegenüber der 4-Klassenlösung kommt bei der 5-Klassenlösung die Klasse 4 hinzu, die als einzige die Profile der anderen Klassen überkreuzt und damit ein qualitativ leicht anderes Muster aufweist. Die übrigen vier Klassenprofile sind ordinal geschichtet.

Für die Beurteilung der Qualität einer LCA-Lösung kann auch die mittlere Höhe der Zuordnungswahrscheinlichkeiten herangezogen werden, d. h. wie gut lassen sich die Antwortmuster einer bestimmten Klasse zuordnen. Für die Kinder in der Klasse 5 gelingt dies mit einer hohen Wahrscheinlichkeit (Wk), denn die mittlere Zuordnungs-Wk liegt bei .92. Die mittleren Zuordnungs-Wk der Klassen 1–4 liegen bei: .87, .87, .89, .72. Damit sind auch die anderen Klassen anhand ihrer Profile gut zuordenbar mit Ausnahme der Klasse 4, die nur bedingt abgrenzbar ist von Klasse 1. Ein Zusammenhang der Klassenzugehörigkeit mit dem Geschlecht war nicht vorhanden ($\chi^2 = 3.42$, df = 4, n. s.). Der Einbezug von Geschlecht als zusätzliche Variable in das LCA-Modell änderte ebenfalls nichts Wesentliches an den LCA-Lösungen und -profilen.

3.4.2 LCA für Items aus dem Bereich »Personen«

Die Informationskriterien (vgl. Tabelle 6) liefern kein konsistentes Bild: Gemäß BIC werden sechs latente Klassen als beste Lösung vorgeschlagen, wobei die BIC-Werte nahe beieinanderliegen. Gemäß

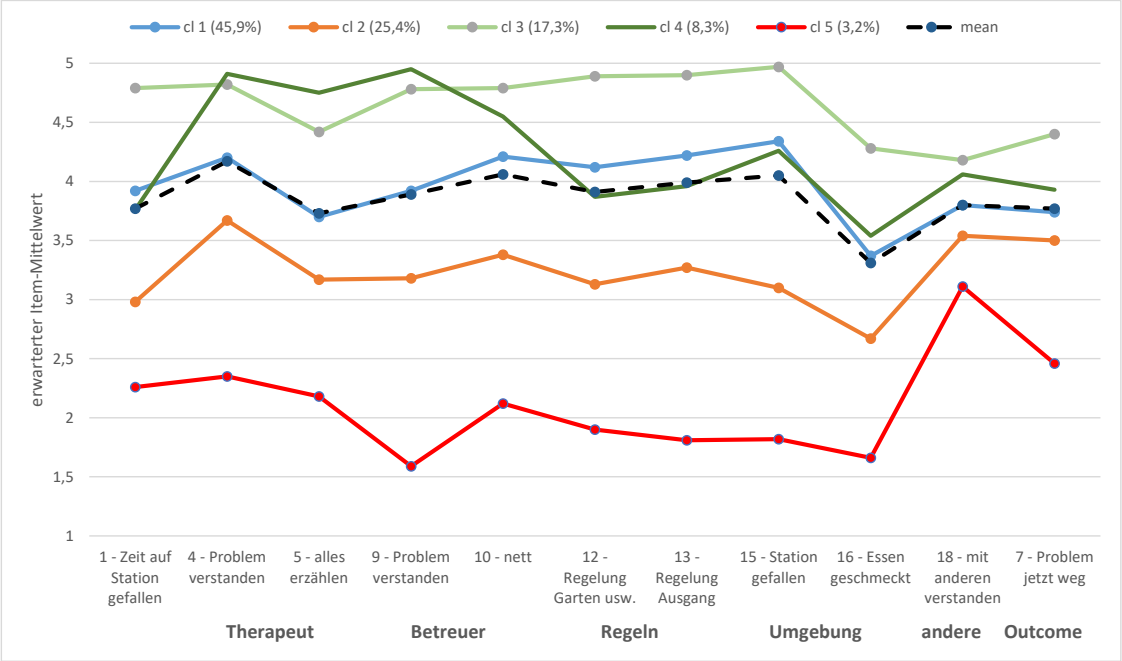

Abbildung 4: *Antwortprofile (erwartete Mittelwerte) für die ausgewählten Items bei einer 5-Klassenlösung (mean = Gesamtmittelwert aller Kinder)*

CAIC würden bereits vier Klassen eine gute Lösung liefern und gemäß AIC könnten noch weitere, über die sechs Klassen hinaus, betrachtet werden.

Nachdem der BIC wieder als Hauptkriterium angesehen wird und die 6-Klassenlösung mit vielen Startwerten stabil replizierbar war, wurde diese Lösung interpretiert. Die Klassenprofile der 6-Klassenlösung sind in Abbildung 5 dargestellt. Anhand des Gesamtmittels ist ersichtlich, dass die Kinder grundsätzlich in den meisten Items eine relativ hohe Behandlungszufriedenheit angeben. Die größte Klasse (cl 1) liegt sehr nahe beim Gesamtmittel, während die zweitgrößte Klasse (cl 2) darunterliegt, aber im Wesentlichen dem Gesamtmuster folgt. Die Klasse 3 (cl 3) ist sehr zufrieden mit Therapeuten und Betreuer, hat aber bezüglich therapeutischem Erfolg (»Problem jetzt weg«) keine besseren Werte. Dagegen ist bei Klasse 4 (cl 4) ihrer eigenen Einschätzung nach das Problem nun weg; auch diese

Tabelle 6: *Ausgewählte Items: Vergleich der Modellanpassung für LC-Modelle mit unterschiedlicher Klassenzahl*

Anzahl latenter Klassen	Log-Likelihood	Anzahl Parameter	BIC	CAIC	AIC
3-LC	−14909.7	50	30186.2	30236.2	29919.3
4-LC	−14850.0	59	30133.1	**30192.1**	29818.1
5-LC	−14817.0	68	30132.9	30200.9	29769.9
6-LC	−14782.8	77	**30130.7**	30207.7	29719.6
7-LC	−14765.1	86	30161.3	30247.3	**29702.2**

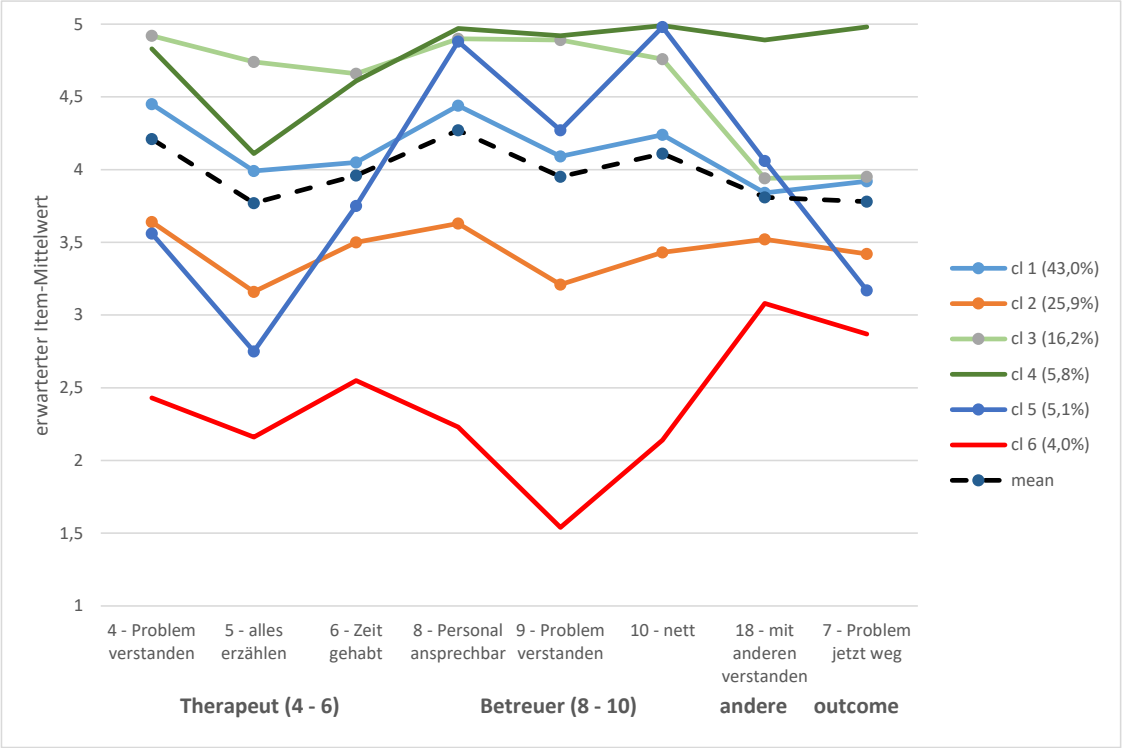

Abbildung 5: *Antwortprofile (erwartete Mittelwerte) bei einer 6-Klassenlösung für den Bereich »Personen« (mean = Gesamtmittelwert aller Kinder)*

Klasse zeichnet eine hohe Zufriedenheit aus, wobei sie trotzdem mit »Therapeut alles erzählen« etwas unzufrieden sind (auf hohem Niveau). Die Klasse 5 (cl 5) ist sehr zufrieden mit den Betreuern, hat aber Probleme mit dem Therapeuten und eine unterdurchschnittliche Erfolgseinschätzung. Die letzte Klasse (cl 6) schließlich ist mit allem unzufrieden und weist in allen Items die niedrigsten Werte auf; sie ist zwar klein, aber klar abgrenzbar und tritt bereits in der 5-Klassenlösung als solche auf.

4 Diskussion

An einer großen Stichprobe von Zufriedenheitserhebungen an Kindern wurden die psychometrischen Eigenschaften des Messinstruments BesT-K untersucht. Dabei wurden zuerst Methoden der klassischen Testtheorie (interne Konsistenz, faktorielle Validität) angewendet. Im Ergebnis zeigten sich insgesamt gute Messeigenschaften und in der faktoriellen Struktur eine empirische Aufgliederung in zwei Faktoren, die als Therapeutische Beziehung und als Regeln/Umgebung bezeichnet werden konnten. Aber auch eine 3-faktorielle Lösung war überlegenswert. Ein bifactor-Modell verwies auf einen starken Generalfaktor, wobei aber die Therapeutische Beziehung noch einen eigenständigen spezifischen

Faktor bildete. Beim zweiten spezifischen Faktor gingen die Items zur Umgebung im Generalfaktor auf und es traten negative Zusammenhänge zwischen den Regel-Items und den Items zu den anderen Kindern auf.

Eine nachfolgende vertiefte Evaluation der Itemqualität hinsichtlich Nutzung und Geordnetheit der Kategorien mittels IRT-Modellen (PCM, GPCM) zeigte, dass der Informationsgehalt der Items unterschiedlich hoch war, aber analog zu den Ladungen in den Faktorenanalysen ausfiel, und dass insbesondere die Itemkategorien in der implizierten ordinalen Weise genutzt wurden, wenngleich nicht in allen Items die fünf Kategorien modellkonform ausgeschöpft zu werden scheinen und insbesondere die Kategorie 2 durch ihre meist geringe Antwortwahrscheinlichkeit vertauschte Schwellenwerte aufwies. Die Reliabilität über das latente Kontinuum hinweg erwies sich im unteren und mittleren Bereich (= niedrige bis mittlere Zufriedenheit) als hoch bis gut, während sie im oberen Bereich absinkt, d. h. Kinder mit hohen Zufriedenheitswerten lassen sich nicht mehr so gut quantitativ voneinander abgrenzen. Solche Differenzierungen zur Messgenauigkeit über den gesamten Bereich des latenten Kontinuums hinweg sind im Übrigen nur in IRT-Modellen möglich, da hier die Testinformationsfunktion und daraus abgeleitete Größen lokal definiert und zudem noch eine Eigenschaft des Tests sind. Im Gegensatz dazu nimmt der (nur aus einem Wert bestehende) Reliabilitätskoeffizient in der klassischen Testtheorie für alle Personen die gleiche Messgenauigkeit an und drückt zudem Eigenschaften des Tests UND der gesamten untersuchten Stichprobe aus (vgl. Samejima 1994).

Die ergänzende Auswertung zur Identifizierung von Antwortmustern mit der LCA, die getrennt für eine Auswahl aus allen Items und speziell noch eine Auswertung der »Personen«-Items vorgenommen wurde, erbrachte keine besonderen Hinweise auf gravierend abweichende Antwortmuster. Einzelne, gering besetzte Klassen mit Überkreuzungen in den Antwortprofilen konnten zwar identifiziert werden, vor allem eine potentiell für die Therapieplanung interessante Subgruppe, die vergleichsweise unzufrieden war mit ihrem/r Therapeut_in, nicht aber mit dem übrigen Klinikpersonal, und auch eine geringere Einschätzung des Therapieerfolgs aufwies, und der in weiteren Studien nachgegangen werden sollte. Insgesamt sind die meisten Profile jedoch weitgehend parallel und bestätigen damit, dass im Wesentlichen Klassen mit verschiedenen Schweregraden an Zufriedenheit vorliegen und insgesamt damit eine quantitative Abstufung sinnvoll ist. Es fanden sich in der LCA auch keine Hinweise auf abweichende Antwortmuster i. S. von »Unskalierbaren«. Eine methodisch vertiefte Analyse könnte über so genannte Hybrid-Modelle erfolgen (vgl. Rost 2004), zum Beispiel im einfachsten Fall über ein 2-Klassenmodell, in dem in einer Klasse das Rasch-Modell (PCM) und damit eine quantitative Abstufung gilt und in einer anderen ein qualitativ abweichendes Antwortmuster besteht (vgl. eine Anwendung in Keller 2012).

Die exploratorische Identifikation der beiden Faktoren »Therapeutische Beziehung« und »Umgebung« deckt sich mit zweien der von Biering (2010) in seiner Literaturübersicht herausgearbeiteten drei »universal components«, die er als »satisfaction with the environment and the organisation of the services«, »satisfaction with the adolescent-caregiver relationship« und »treatment outcome« bezeichnet. Im ambulanten Bereich zeigten sich bei Brown et al. (2014) ebenfalls die Faktoren »satisfaction with care« und »satisfaction with environment«, die freilich hoch korreliert waren. Das Ergebnis der Behandlung (»treatment outcome«) scheint dagegen sowohl in der eigenen Studie als auch generell

(Biering 2010) nur gering mit der Zufriedenheit zusammenzuhängen. Allerdings stand in der eigenen Studie zur Analyse nur das Item 7 (Problem jetzt weg) zur Verfügung, da in den Versorgungskliniken neben dem BesT-K aus Praktikabilitätsgründen Symptommaße nicht systematisch miterhoben wurden. Solche standardisierten Messinstrumente, vorzugsweise auf Basis von Fremdbeurteilungen, wären jedoch notwendig, um dem Zusammenhang zwischen Zufriedenheit und Symptomverlauf weiter nachzugehen. Die Kinder selbst werden üblicherweise von ihren Eltern zur Therapie gebracht, sie sehen häufig bei sich kein besonderes Problem und sie leiden auch nicht besonders unter ihren Symptomen, im Gegensatz zu Eltern und anderen Bezugspersonen (Schwab & Stone 1983).

Geringe Zusammenhänge wies auch das Item »Aufklärung Medikamente« auf, das weder substantiell auf dem Generalfaktor lud noch einem der Faktoren gut zuordenbar war. Die Aufklärung zur Medikation ist normalerweise Aufgabe des Therapeuten und ein Bezug zum Faktor Therapeutische Beziehung wäre naheliegend. Eine Erklärung für diese fehlende Zuordnung liegt eventuell darin, dass die Aufklärung nicht unbedingt vom eigenen Therapeuten, sondern vom zuständigen Arzt/Ärztin, zu dem man ansonsten nur wenig Kontakt hat, vorgenommen wird. Ebenfalls geringe Faktorladungen und eine gemeinsame Residualkorrelation weisen die beiden Items 17 und 18 auf, in denen es um die Beziehung zu den anderen Kindern und generell um das Stationsklima geht. Obwohl dies ein wichtiger Aspekt von Behandlungszufriedenheit ist, könnten sich hier einzelne negative Erlebnisse widerspiegeln, die vom Kind selbst kaum zu beeinflussen sind, und daher eine geringe Korrelation mit den übrigen Items bewirken. Insbesondere könnte es um einzelne Streitereien gehen oder das Gefühl, ungerecht behandelt worden zu sein durch eine bestimmtes anderes Kind; derartige Aussagen finden sich immer wieder in den Freitexten zur Frage »was mich geärgert hat«.

Eine überraschende Gemeinsamkeit ergibt sich für die Items zu den Regeln und den Fragen zur Strukturqualität (»Hotelqualität«), die beide auf einem Faktor laden. Im Erleben der Kinder fließen diese beiden Aspekte anscheinend ineinander und die Regelungen scheinen nicht Teil der therapeutischen Beziehung zu sein. Allerdings gehen die Items zur Strukturqualität im Generalfaktor auf, wenn ein bifactor-Modell angenommen wird.

Insgesamt sind die dimensionalen Differenzierungen in die beiden Hauptbereiche Therapeutische Beziehung und Umgebung/Regeln stabil und zeigten sich in gleicher Weise auch mit anderen Auswertungsmethoden, die hier nicht dargestellt sind, z. B. einem GPCM mit zwei Faktoren. Eine analoge inhaltliche Auftrennung findet sich auch bei Jugendlichen (vgl. Keller et al. 2018). Die noch unklaren Details, v. a. ob der dritte Faktor sinnvoll interpretierbar ist und dass die Ladungen auf dem zweiten spezifischen Faktor im bifactor-Modell teilweise negativ sind, sollten erst noch repliziert werden, zumal auch keine vergleichbaren Ergebnisse aus anderen Zufriedenheitsstudien zur Verfügung stehen. Gerade bei bifactor-Modellen könnten dann weitere Koeffizienten berechnet werden, die die Frage, ob eine Skala essentiell unidimensional ist (Reise 2012), quantifizierbar machen. Neben der »explained common variance« (ECV) des Generalfaktors sind dies unterschiedliche omega-Koeffizienten, die z. B. den verbleibenden Varianzanteil, der auf die spezifischen Faktoren zurückgeht, bestimmen, nachdem der Anteil des Generalfaktors herausgezogen wurde (vgl. Rodriguez et al. 2016).

Ein starker Generalfaktor kann auch als pragmatische Begründung für die Bildung eines Summenwertes über die Items angegeben werden, obwohl das Rasch-Modell (PCM), das als einziges eine strenge

Prüfung und Akzeptanz der Summenbildung als erschöpfender Statistik ermöglicht, nicht so gut passt wie ein Modell, das unterschiedliche Gewichtungen der Items vorsieht (GPCM). Eine separate Berechnung des Gesamtscores für jede Person auf der Basis der Item-Gewichte ist jedoch unrealistisch und in der Praxis unüblich. Außerdem kann auch das Rasch-Modell trotz einzelner Probleme bei Itemparametern eine valide Summenbildung ermöglichen (vgl. Alexandrowicz et al. 2014, die das Beck Depressionsinventar (BDI-II) in verschiedenen Stichproben verglichen haben und weitere Probleme in diesem Zusammenhang erörtern).

Bezüglich des Antwortformats ließe sich noch diskutieren, ob man die Kategorie 2 mit Kategorie 1 zusammenlegt. Wie bereits oben ausgeführt würde die Reliabilität dadurch kleiner werden und die Kategorie 2 ist auch ausreichend besetzt mit mindestens 4 % der Antworten. Bei einigen Items haben zudem noch die Kategorien 3 und 4 eine niedrige Antwortwahrscheinlichkeitskurve. Dennoch scheint es angebracht, bei allen Items durchgängig im fünfstufigen Antwortformat zu bleiben, zumal auch »schlechte« Antwortwahrscheinlichkeitskurven und sogar vertauschte Schwellenparameter nicht unbedingt auf eine Verletzung der Modellannahmen hinweisen. Vertauschungen können auch durch geringe Häufigkeiten in den betroffenen Kategorien bedingt sein und sie sprechen nicht gegen die Ordinalität der Skala und auch nicht dafür, solche Kategorien in jedem Fall zusammenzufassen (Wetzel & Carstensen 2014). In weiteren Analysen sollte aber ein möglicher Alterseffekt untersucht werden, da ein unterschiedlich differenzierteres Ausfüllen der Items bei jüngeren und bei älteren Kindern vermutet werden könnte.

Weitere Analysen zur Nutzung der Antwortkategorien könnten noch zu sogenannten »response sets« durchgeführt werden, d. h. dass beispielsweise manche Personen bevorzugt die extremen Kategorien (hier: 1 und 5) ankreuzen, während andere diese eher vermeiden. Diese Tendenzen sind bekannt aus der Analyse von Persönlichkeits- und Einstellungsskalen (z. B. Rost et al. 1999) und dürften auch in klinischen Skalen vorhanden sein (Keller & Koller 2015), eventuell sogar diagnosespezifisch (Diagnosen standen in der vorliegenden Studie leider in der überwiegenden Mehrzahl der Fälle nicht zur Verfügung). Treten solche Antworttendenzen auf und lassen sie sich z. B. in zwei latente Klassen zuordnen, können die Antworten der Kinder bzw. ihre Personenparameter im IRT-Modell gemäß ihrem »response set« korrigiert werden.

Die erweiterten Auswertungen bezüglich des Einflusses von Alterseffekten und response sets sowie über die hier unterstellte Unidimensionalität bei den IRT-Modellen hinaus sollte dabei nicht in separaten Subgruppenanalysen erfolgen (z. B. jüngere vs. ältere Kinder, zwei Arten von response sets), sondern möglichst unter der Vorgabe von Multidimensionalität bei den IRT-Modellen (z. B. Reckase 2009) und im Rahmen eines Modells, in dem Alter und response sets und eventuell weitere Kovariaten als dimensionale Variablen berücksichtigt werden können. Gut geeignet erscheint hier das multidimensionale Modell von Adams et al. (1997) (bzw. darauf aufbauende Entwicklungen und die entsprechenden Berechnungsmöglichkeiten mit R-packages wie das TAM von Robitzsch, Kiefer und Wu), in dem die genannten Effekte simultan geschätzt werden könnten.

Zusammenfassend lässt sich für die Bestimmung der Behandlungszufriedenheit von Kindern mit ihrer (teil-)stationären Behandlung mit dem BesT-K schlussfolgern, dass für Zwecke der Qualitätssicherung auf die Einzelitems zurückgegriffen werden kann, auch wenn nicht alle Items vollkommene

IRT-Modellkonformität aufweisen und systematische Veränderungsanalysen noch ausstehen. Gerade im Qualitätsmanagment (QM) von Kliniken wird der Fokus häufig auf einzelne Items gelegt, bei denen die eigene Klinik oder eine Station schlechte Werte im benchmarking aufweisen. Konkret ergaben sich in der Stichprobe zum Beispiel klinikspezifische Unterschiede in der Akzeptanz der Ausgangsregelung, was zu QM-Maßnahmen und daraus resultierender Verbesserung in einigen Kliniken führte. Bevorzugt betrachtet wird zudem die Qualität der Medikamentenaufklärung, da die Informiertheit und Teilhabe der Kinder auf diesem Gebiet sehr ernst genommen wird. Viele Kliniken haben deshalb eigens Formulare entwickelt und die Medikamentenaufklärung über die Jahre hinweg verbessert. Für eine zusammenfassende Bewertung der Behandlungszufriedenheit im Rahmen des QM und insbesondere für wissenschaftliche Zwecke bietet sich die Verwendung des BesT-K Gesamtwertes und der beiden Subskalen Therapeutische Beziehung und Umgebung/Regeln an.

Danksagung: Ich danke Rainer Alexandrowicz, Renate Schepker und Daniela Wetzelhütter herzlich für ihre wertvollen Anregungen zur ersten Fassung dieses Beitrags.

Literatur

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). »The multidimensional random coefficient multinomial logit model«. In: *Applied Psychological Measurement* 21, S. 1–23.

Alexandrowicz, R. W., Fritzsche, S. & Keller, F. (2014). »Die Anwendbarkeit des BDI-II in klinischen und nichtklinischen Populationen aus psychometrischer Sicht. Eine vergleichende Analyse mit dem Rasch-Modell«. In: *Neuropsychiatrie* 28, S. 63–73.

Biering, P. (2010). »Child and adolescent experience of and satisfaction with psychiatric care: A critical review of the research literature«. In: *Journal of Psychiatric and Mental Health Nursing* 17, S. 65–72.

Bowling, A., Rowe, G., Lambert, N., Waddington, M., Mahtani, K. R., Kenten, C., Howe, A. & Francis, S. A. (2012). »The measurement of patients' expectations for health care: A review and psychometric testing of a measure of patients' expectations«. In: *Health Technology Assessment* 16.30.

Brown, A., Ford, T., Deighton, J. & Wolpert, M. (2014). »Satisfaction in child and adolescent mental health services: Translating users' feedback into measurement«. In: *Adm Policy Ment Health* 41, S. 434–446.

Chalmers, R. P. (2012). »mirt: A Multidimensional item response theory package for the R environment«. In: *Journal of Statistical Software* 48, S. 1–29.

Dür, W., Grossmann, W. & Schmied, H. (2000). »Patientenzufriedenheit und Patientenerwartung im Krankenhaus: Statistische Analysen als Hilfsmittel im Benchmarking«. In: *Lebensqualitätsforschung aus medizinpsychologischer und -soziologischer Sicht*. Hrsg. von M. Bullinger, J. Siegrist & U. Ravens-Sieberer. Göttingen: Hogrefe, S. 222–243.

Gill, L. & White, L. (2009). »A critical review of patient satisfaction«. In: *Leadership in Health Services* 22, S. 8–19.

Gruyters, T. & Priebe, S. (1994). »Die Bewertung psychiatrischer Behandlung durch den Patienten – Resultate und Probleme der systematischen Forschung«. In: *Psychiatrische Praxis* 21, S. 88–95.

Hu, L. & Bentler, P. (1999). »Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives«. In: *Structural Equation Modelling: A Multidisciplinary Journal* 6, S. 1–55.

Kammerer, E. (1989). »Bewertung stationärer jugendpsychiatrischer Therapie: Eine Gegenüberstellung der Urteile von Jugendlichen und Eltern«. In: *Praxis der Kinderpsychologie und Kinderpsychiatrie* 38, S. 205–209.

Kaplan, S., Busner, J., Chibnall, J. & Kang, G. (2001). »Consumer satisfaction at a child and adolescent state psychiatric hospital«. In: *Psychiatric Services* 52, S. 202–206.

Keller, F. (2012). »Latent-Class- und Mixed-Rasch-Modelle zur Identifizierung skalierbarer und unskalierbarer Personengruppen in der Allgemeinen Depressionsskala«. In: *Item-Response-Modelle in der sozialwissenschaftlichen Forschung*. Hrsg. von W. Kempf & R. Langeheine. Berlin: Verlag irena regener, S. 171–188.

Keller, F., Fegert, J. M. & Naumann, A. (2018). »Fragebögen zur Behandlungseinschätzung stationärer Therapie (BesT) in der Kinder- und Jugendpsychiatrie: Entwicklung und Validierung für Jugendliche und für Eltern«. In: *Zeitschrift für Klinische Psychologie und Psychotherapie* 47, S. 186–197.

Keller, F. & Koller, I. (2015). »Mixed Rasch models for analyzing the stability of response styles across time: An illustration with the Beck Depression Inventory (BDI-II)«. In: *Dependent data in social sciences research: Forms, issues, and methods of analysis*. Hrsg. von M. Stemmler, A. von Eye & W. Wiedermann. Heidelberg: Springer-Verlag, S. 309–324.

Keller, F., Schäfer, S., Konopka, L., Naumann, A. & Fegert, J. M. (2004). »Behandlungszufriedenheit von Kindern in stationär psychiatrischer Behandlung: Entwicklung und psychometrische Eigenschaften eines Fragebogens«. In: *Krankenhauspsychiatrie* 15, S. 3–8.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. 2. Aufl. New York: Guilford Press.

Kriz, D., Nübling, R., Steffanowski, A., Wittmann, W. W. & Schmidt, J. (2008). »Patientenzufriedenheit in der stationären Rehabilitation: Psychometrische Reanalyse des ZUF-8 auf der Basis multizentrischer Stichproben verschiedener Indikation«. In: *Zeitschrift für Medizinische Psychologie* 17, S. 67–79.

Lebow, J. L. (1983). »Research assessing consumer satisfaction with mental health treatment: A review of findings«. In: *Evaluation and Program Planning* 6, S. 211–236.

Leimkühler, A. M. & Müller, U. (1996). »Patientenzufriedenheit – Artefakt oder soziale Tatsache?« In: *Nervenarzt* 67, S. 765–773.

Mattejat, F. & Remschmidt, H. (1998). *Fragebögen zur Beurteilung der Behandlung (FBB)*. Göttingen: Hogrefe.

Muraki, E. (1992). »A generalized partial credit model: Application of an EM algorithm«. In: *Applied Psychological Measurement* 16, S. 159–176.

Muthén, L. & Muthén, B. O. (2012). *Mplus user's guide*. 7. Aufl. Los Angeles, CA: Muthén & Muthén.

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«. In: *Structural Equation Modeling* 14, S. 535–569.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer-Verlag.

Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften*. 2. Aufl. München: Oldenbourg.

Reise, S. P. (2012). »The rediscovery of bifactor measurement models«. In: *Multivariate Behavioral Research* 47, S. 667–696.

Rodriguez, A., Reise, S. P. & Haviland, M. G. (2016). »Evaluating bifactor models: Calculating and interpreting statistical indices«. In: *Psychological Methods* 21.2, S. 137–150.

Rost, J. (2004). *Lehrbuch Testtheorie Testkonstruktion*. 2. Aufl. Bern: Verlag Hans Huber.

Rost, J., Carstensen, C. H. & von Davier, M. (1999). »Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten«. In: *Diagnostica* 45.3, S. 119–127.

Samejima, F. (1994). »Estimation of reliability coefficients using the test information function and its modifications«. In: *Applied Psychological Measurement* 18, S. 229–244.

Schwab, M. E. & Stone, K. (1983). »Conceptual and methodologic issues in the evaluation of children's satisfaction with their mental health care«. In: *Evaluation and Program Planning* 6, S. 283–289.

Vermunt, J. K. & Magidson, J. (2003). *LatentGOLD user's guide*. Belmont, MA: Statistical Innovations Inc.

Viefhaus, P., Döpfner, M., Dachs, L., Goletz, H., Görtz-Dorten, A., Kinnen, C., Perri, D., Rademacher, C., Schürmann, S., Woitecki, K., Wolff Metternich-Kaizman, T. & Walter, D. (2019). »Treatment satisfaction following routine outpatient cognitive-behavioral therapy of adolescents with mental disorders: A triple perspective of patients, parents and therapists«. In: *European Child & Adolescent Psychiatry* 28, S. 543–556.

Wetzel, E. & Carstensen, C. H. (2014). »Reversed thresholds in partial credit models: A reason for collapsing categories?« In: *Assessment* 21.6, S. 765–774.

Wirth, R. J. & Edwards, M. C. (2007). »Item factor analysis: Current approaches and future directions«. In: *Psychological Methods* 12, S. 58–79.