Item-Response-Modelle
in der sozialwissenschaftlichen Forschung

Wilhelm Kempf & Rolf Langeheine (Hrsg.)

Item-Response-Modelle in der sozialwissenschaftlichen Forschung

verlag irena regener berlin

© 2012 by verlag irena regener
ostseestr. 109, d – 10409 berlin
http://www.regener-online.de

Alle Rechte vorbehalten, auch die der fotomechanischen Wiedergabe und der Speicherung in elektronischen Medien.

Printed in Germany

ISBN 978-3-936014-29-7

Jürgen Rost zum 60. Geburtstag gewidmet

Inhalt

Einleitung
Wilhelm Kempf & Rolf Langeheine 9

Schachprogramme, Testaufgaben, Populationsunabhängigkeit und das Rasch-Modell – oder: Es ist Zeit für das Rost-Modell
Matthias von Davier 15

Scorebildung, klassische Testtheorie und Item-Response Modelle in der Einstellungsmessung
Wilhelm Kempf 32

Konsequenzen des Ignorierens von globalen Modelltests in Studien wie TIMSS, PISA und PIRLS
Rolf Langeheine 55

Soziale Herkunft, kulturelles Kapital und Lesefähigkeit – ein Vergleich zwischen Finnland und Deutschland mit Hilfe der PISA 2000-Studie
Werner Georg 68

Latent-Class-Modelle von Spracheinstellungen in der berliner Sprachgemeinschaft nach 1989
Irena Regener 83

Modellierung von längsschnittlichen Daten am Beispiel einer quasi-experimentellen Studie zur Erfassung von naturwissenschaftlichem Kompetenzen im Kindergartenalter
Claus H. Carstensen, Eva-Maria Lankes & Mirjam Steffensky 109

Modelle der Multitrait-Multimethod-Analyse für Variablen mit geordneten Antwortkategorien
Michael Eid & Fridtjof W. Nussbeck 127

Ein symmetrisches Rating-Scale-Modell für Fragebogen mit invertierten Items
Thorsten Meiser, David Hoeffler & Fabian Jasper 150

Latent Class- und Mixed Rasch-Modelle zur Identifizierung skalierbarer und unskalierbarer Personengruppen in der Allgemeinen Depressionsskala (ADS)
Ferdinand Keller 171

Zur Skalierung der Lebensqualität: Probabilistische Testmodelle bei der Bestimmung von Nutzwerten für die gesundheitsökonomische Evaluation
Thomas Kohlmann, You-Shan Feng, Bas Janssen & Simon Pickard 189

Was ist aus dem Integrierten Handlungsmodell geworden?
Thomas Martens 210

Literatur 230

Anhang

Über die Autoren 251

Sachregister 255

Einleitung

Im Herbst 2009 verbrachten wir (Irena Regener, Wilhelm Kempf und Rolf Langeheine) einen Tag mit Jürgen Rost in Kiel. Kurz darauf hatte Irena die Idee für dieses Buch und bat uns, als Herausgeber ein Konzept dafür zu entwickeln: für ein Buch über aktuelle Entwicklungen auf einem Forschungsgebiet, zu dem Jürgen Rost entscheidende Beiträge geleistet hat und zu dem er sicher auch weiterhin gern beigetragen hätte, wenn er seine aktive wissenschaftliche Laufbahn nicht aus Gesundheitsgründen bereits in sehr jungen Jahren hätte beenden müssen. Jürgen Rost war über viele Jahre hinweg nicht nur der vielleicht hervorragendste Vertreter der Item-Response-Theorie im deutschen Sprachraum und uns allen ein geschätzter Kollege – für viele von uns ist er auch Freund und/oder Lehrer. Nicht zuletzt deshalb wollen wir ihm das nun vorliegende Buch zu seinem sechzigsten Geburtstag im August 2012 widmen.

Es untersucht die Leistungsfähigkeit von Item-Response-Modellen sowohl in theoretischen als auch in angewandten Beiträgen, deren Gegenstandsbereich von der psychologischen Diagnostik über die Gesundheitspsychologie, die kognitive Psychologie, die Umwelt- und die Sozialpsychologie bis hin zur Soziologie und zur Soziolinguistik reicht. Ganz im Sinne von Jürgen Rost werden nicht nur methodologische Probleme diskutiert, die sich aus den konkreten Fragestellungen der substanzwissenschaftlichen Forschung ergeben; die Rückwirkung der statistischen Modelle auf die inhaltlichen Disziplinen wird ebenfalls demonstriert. Jürgen Rost war und ist es immer sehr wichtig, dass statistische Modelle auch Auswirkungen in den inhaltlichen Disziplinen haben. Seine Arbeiten hatten daher stets die Erweiterung des Anwendungsbereichs der Modelle im Sinn.

"Schachprogramme, Textaufgaben, Populationsunabhängigkeit und das Rasch-Modell – oder: Es ist Zeit für das Rost-Modell". Unter diesem Titel diskutiert *Matthias von Davier* die Frage, warum eine modellbasierte Auswertung von Testdaten einer 'agnostischen' Auswertung mittels einfacher Summenscores vorzuziehen ist. Am Beispiel des Rasch-Modells und dessen Erweiterung zum Mixed-Rasch-Modell werden grundlegende Annahmen der Testauswertung erläutert und die Konsequenzen einer Verletzung dieser Annahmen diskutiert. Mittels des Konzeptes der simultanen probabilistischen Kausalität wird die Annahme der lokalen Unabhängigkeit als Basis vieler Testmodelle etabliert, und anhand eines von Zermelo eingeführten Ansatzes, der eng mit dem Rasch-Modell verwandt ist, zeigt der Autor

schließlich auf, dass die Frage der Parameterseparierbarkeit und des Linkings über Testformen hinweg im Rasch-Modell eine natürliche Lösung hat, die – verglichen mit klassischen Ansätzen – keine künstlichen Zusatzannahmen benötigt.

Summenscores sind ein plausibles, doch leider nicht voraussetzungsloses Maß. 50 Jahre, nachdem Georg Rasch diese Voraussetzungen bewiesen hat, summiert man in der Einstellungsmessung und weiten Teilen des psychologischen Mainstreams immer noch munter darauf los, ohne sich darum zu kümmern. Diesen Missstand nimmt *Wilhelm Kempf* zum Anlass, um ein informationstheoretisches Maß – den Loss-of-Information-Index (LI) – zu konstruieren, das den Informationsverlust beschreibt, der mit der Summenscorebildung einhergeht, wenn ihre Voraussetzungen verletzt sind. Eine Anwendung auf die Messung antisemitischer Einstellungen zeigt, dass die Verwerfung des Rasch-Modells zugunsten von Latent-Class-Modellen zwar eine Abkehr vom Konzept einer quantitativen Einstellungsdimension impliziert, andererseits jedoch genaueren Aufschluss über Struktur und Dynamik der jeweiligen Einstellung liefert. Gerade das, was die klassische Testtheorie als Messfehler vernachlässigt, kann ggf. entscheidende diagnostische Information enthalten.

Während es sich in der Einstellungsmessung meist als zu streng erweist, hat das Rasch-Modell in der Schulleistungsdiagnostik neuerdings seinen Siegeszug angetreten. In vielen großen internationalen Studien wie TIMSS und PISA wird es angewendet. In einem ersten Schritt werden anhand einer Zufallsstichprobe nach Elimination nichtpassender Items Itemparameter geschätzt. Im zweiten Schritt werden mit fixierten Itemparametern aus Schritt Eins Personenscores getrennt für jedes Land geschätzt. Dabei wird allerdings nicht geprüft, ob Iteminvarianz über die Länder vorliegt. *Rolf Langeheine* untersucht dieses Problem exemplarisch an einem kleinen Datensatz mit vier Items und 42 Ländern aus TIMSS95. Die Beurteilung der Anpassungsgüte des Rasch-Modells erfolgt nach dem Konzept des pi*-Indexes mittels eines latenten Klassenmodells mit zwei Komponenten. In der ersten Komponente befindet sich das Rasch-Modell, in der zweiten der damit nicht verträgliche Rest. Die Ergebnisse zeigen eine erhebliche, teils drastische Zunahme der mit dem Rasch-Modell nicht verträglichen Komponente von länderspezifischen zu fixierten Itemparametern.

Die PISA 2000-Studie untersuchte ländervergleichend die Lesefähigkeit von Schülern der neunten Klasse in 32 Ländern. Im Ergebnis schnitt Finnland mit einem durchschnittlichen Testscore von 546 (bei einem internationalen Mittelwert von 500) am besten ab: Deutschland lag, und dieser Umstand führte zu erhitzten öffentlichen Diskussionen, mit 484 Punkten abgeschlagen auf dem 21. Platz. Noch beunruhigender war der Tatbestand, dass die Unterschiede in der Lesefähigkeit zwischen dem untersten und obersten Quartil des International Socio Economic Index (ISEI) mit 111 Skalenpunkten am höchsten von allen Ländern waren. Das Deutsche PISA- Konsortium erklärte diese Unterschiede mit einem familialen Prozessmodell, in dessen Zentrum insbesondere die differenzielle Verfügung über kulturelles und soziales Kapital steht. Dies bildet den Ausgangspunkt des Beitrages von *Werner Georg*, der mittels eines Latent-Class-Modells für Deutschland und Finnland die Effekte des familialen Kulturkapitals in den beiden Ländern vergleicht.

Latent-Class-Analysen (LCA) sind auch konstitutiv für eine Längsschnittstudie, in der diese Methode erstmals in der Soziolinguistik angewendet wurde. *Irena Regener* untersucht am Beispiel der berliner Sprachgemeinschaft, welche Funktion Spracheinstellungen und Sprachverhalten bei den Identitätsfindungs- bzw. -sicherungsprozessen nach 1989 zukam und wie diese zu Annäherung vs. Abgrenzung zwischen Ost und West beitrugen. Die Ergebnisse der LCA zeichnen ein nach wie vor inhomogenes Bild der Stadt am Ende der 1990er-Jahre: Die westberliner Sprachgemeinschaft erscheint zwar zunächst gefestigter und deshalb änderungsresistenter oder auch konservativer, aber es gibt eine Bewegung hin zu positiver Interessiertheit am Ostpendant und dessen Sprache. Die ostberliner Sprachgemeinschaft, unmittelbar nach 1989 in ihrem Selbstverständnis erheblich verunsichert, hat inzwischen einen Teil ihres Selbstbewusstseins und damit auch ihrer sprachlichen Autonomie zurückgewonnen, weil sie ihre eigene Sprache – zunehmend wieder – als einen ihrer zentral identitätsstiftenden Werte wahrnimmt.

Die folgenden Beiträge diskutieren Modellvarianten des Rasch-Modells und deren Anwendung auf verschiedene Problemsituationen. *Claus H. Carstensen, Eva-Maria Lankes & Mirjam Steffensky* thematisieren die Modellierung von Längschnittsdaten in der Bildungsforschung. Als Beispiel dient eine Studie, in der naturwissenschaftliche Kompetenzen bei Kindern im Kindergartenalter in einem Vortest, in einem Nachtest nach einer Interventionsphase und in einem Follow-up-Test erhoben werden. Zur Modellierung der Kompetenzen, für jeden Testzeitpunkt und über die drei Messzeitpunkte hinweg, werden (verallgemeinerte) Rasch-Modelle diskutiert. Die Kompetenzen lassen sich sowohl durch ein Andersen- als auch durch ein Embretson-Modell für drei Zeitpunkte darstellen. Mit dem Andersen-Modell werden Kompetenzzuwächse und Veränderungen in der Varianz der Kompetenzen abgebildet; mit dem Embretson-Modell lassen sich Zusammenhänge zwischen den Kompetenzen zum ersten Zeitpunkt und den Veränderungen über die Zeit darstellen. Ein besonderes Augenmerk des Beitrages liegt auf der Frage, ob sich Kompetenzzuwächse Fünfjähriger als Veränderung entlang einer eindimensionalen, Rasch-homogenen latenten Variable abbilden lassen.

Michael Eid & Fridtjof Nussbeck demonstrieren, wie anhand von Modellen der Item-Response-Theorie Untersuchungen zur konvergenten und diskriminanten Validität durchgeführt werden können. Nach einer Einführung in zentrale Konzepte der Multitrait-Multimethod-Analyse nach Campbell & Fiske werden in der Item-Response-Theorie vorhandene Modelle zur Analyse multimethodal erhobener Daten vergleichend diskutiert. Aufbauend auf der Unterscheidung zwischen strukturell unterschiedlichen und austauschbaren Methoden wird aufgezeigt, wie ein Multitrait-Multimethod-Modell für geordnete Variablen definiert werden kann, das in der Lage ist, beide Arten von Methoden simultan zu berücksichtigen. Das Modell wird als faktorenanalytisches Modell für Variablen mit geordneten Antwortkategorien formuliert, und es wird darauf hingewiesen, dass es äquivalent zu einem multidimensionalen Graded-Response-Modell ist. Die vorgestellten Modellierungsansätze werden mit einem Beispiel aus der differenzialpsychologischen Stimmungsforschung illustriert.

Bei Fragebogenuntersuchungen kommt es nicht selten vor, dass ein Teil der Items invertiert – d.h. negativ formuliert – ist und deshalb für die Analyse rekodiert werden muss. In solchen Fällen können das Rating-Scale-Modell und andere Varianten des Rasch-Modells für ordinale Antwortkategorien, die itemunabhängige Bestandteile von Schwellenabständen postulieren, zu gravierenden Fehlspezifikationen führen. *Thorsten Meiser, David Hoeffler & Fabian Jasper* schlagen daher neue Modelle vor, die der Rekodierung invertierter Items Rechnung tragen und die eine Analyse der zugrundeliegenden Antwortprozesse auf invertierte und nicht-invertierte Items ermöglichen. Die Anwendung und Interpretation der verschiedenen Modelle wird anhand eines empirischen Datensatzes zum Konstrukt des Personal Need for Structure illustriert.

Ferdinand Keller beschäftigt sich mit den psychometrischen Eigenschaften der Allgemeinen Depressionsskala (ADS). Neben einer eindimensionalen Struktur der Skala wird auch eine mehrfaktorielle Struktur und/oder ein Methodenfaktor diskutiert, da der Faktor Wohlbefinden durch vier positiv formulierte Items abgefragt wird, die umgepolt werden, bevor sie in den Summenwert eingehen. In einer früheren Auswertung einer deutschen Jugendlichenstichprobe mit probabilistischen Modellen konnte eine Klasse identifiziert werden, die sich dadurch von allen anderen qualitativ unterscheidet, dass sie bei den negativ formulierten Items auffällig niedrige ADS-Werte ausweist, gleichzeitig aber in den vier positiv formulierten Items vergleichsweise hohe Depressionswerte erreicht. Diesen Befund verfolgt der Autor mittels der Analyse der Normstichprobe für Deutschland, einer neu erhobenen Jugendlichenstichprobe sowie einer Älterenstichprobe weiter.

Erhebungsverfahren zur Messung von Indikatoren der subjektiven Gesundheit haben in der Gesundheitsforschung der letzten Jahre eine immer größere Bedeutung gewonnen. Dabei wurden bei der Entwicklung und Validierung der Messinstrumente zunehmend auch psychometrische Modelle aus der Item-Response-Theorie verwendet. In dem Beitrag von *Thomas Kohlmann, You-Shan Feng, Bas Janssen & Simon Pickard* werden Spezialfälle des generalisierten Partial-Credit-Modells genutzt, um beim Übergang von einer drei- auf eine fünfstufige Antwortvorgabe eines Gesundheitsfragebogens eine methodisch angemessene Skalierung zu erreichen. Es handelt sich bei diesem Fragebogen (dem EQ-5D) um das international am häufigsten eingesetzte Verfahren zur präferenzbasierten Messung der gesundheitsbezogenen Lebensqualität. Die Ergebnisse der Analysen zeigen, wie moderne psychometrische Verfahren zur methodisch abgesicherten Weiterentwicklung von Messinstrumenten auch in der Gesundheitsforschung beitragen können.

In dem Beitrag von *Thomas Martens* wird das von Jürgen Rost entwickelte Integrierte Handlungsmodell reflektiert, auf Lernprozesse übertragen, theoretisch erweitert und beispielhaft empirisch überprüft. Zunächst werden die theoretischen Potenziale des Modells aufgezeigt, indem es auf Lernprozesse angewendet wird. Danach werden die typischen Prozesse der drei Handlungsphasen dargestellt, insbesondere die funktionslogische Integration von bewussten und unbewussten sowie kognitiven und affektiven Prozessen. Abschließend wird eine empirische Studie vorgestellt, die das Lernen von Statistik im Hochschulbereich untersucht. Mit Hilfe

von Mischverteilungsmodellen werden sieben Lerntypen identifiziert, die den entsprechenden motivationalen, intentionalen und volitionalen Mustern der Lern- und Handlungsgenese entsprechen.

Jürgen Rost dieses Buches zu widmen erscheint uns nicht zuletzt auch deshalb angemessen, weil sich alle Buchbeiträge in der einen oder anderen Weise auf seine Arbeiten beziehen und so erkennen lassen, welch entscheidenden Einfluss Jürgen Rost auf unsere Arbeit und die Item-Response-Theorie insgesamt hatte, hat und wohl auch in Zukunft noch haben wird.

Konstanz & Bordesholm, im Frühjahr 2012

Wilhelm Kempf
Rolf Langeheine

Schachprogramme, Testaufgaben, Populationsunabhängigkeit und das Rasch-Modell – oder: Es ist Zeit für das Rost-Modell

Matthias von Davier

1 Einleitung

Ein Testverfahren auszuwerten ist nicht schwer. Schwer ist es nur, es richtig zu machen. Überspitzt dargestellt können zwei sehr unterschiedliche Traditionen der Auswertung psychologischer und bildungswissenschaftlicher Tests identifiziert werden:

Die klassische Tradition wird von einigen ihrer Vertreter als agnostisch bezeichnet. Sie besteht darin, einen einfachen Summenscore zu bilden und nicht weiter darüber nachzudenken, ob diese Summation über Testaufgaben hinweg eine angemessene Zusammenfassung des Phänomens ist, das wir beschreiben und verstehen wollen. Von Kritikern wird diese Art der Testauswertung auch als naive Testtheorie bezeichnet, weil sie die fast ausschließliche Fokussierung auf nur eine von vielen möglichen Zusammenfassungen in Form einer (wenn auch naheliegenden) Summe als unreflektierte Praxis ansehen.

Wir müssen jedoch einschränkend anerkennen, dass diese naive Praxis natürlich das ist, was einem als Erstes in den Sinn kommt, wenn eine Reihe von richtigen und falschen Antworten auf eine große Zahl ähnlicher Aufgaben zusammengefasst werden soll: Wenn wir diese Antworten mit X_i bezeichnen, wobei X_i =1 für eine richtige und X_i = 0 für eine falsche Antwort steht, und annehmen, dass es insgesamt m Aufgaben gibt, dann ist der einfache, ungewichtete Summenscore definiert als

$$S = \sum_{i=1}^{m} X_i = X_1 + X_2 + \ldots + X_m \qquad (1)$$

Dieser Summenscore S ist eine potenziell sehr nützliche Größe, darüber sind sich die unterschiedlichen Traditionen einig. Uneinigkeit jedoch besteht zumeist darüber, ob erst gezeigt werden muss, dass diese Summation sinnvoll ist oder ob einfach davon ausgegangen werden kann, dass der Summenscore S in den meisten Fällen als Zusammenfassung zum Vergleich der Leistungen der Testteilnehmer ausreicht.

"Agnostischer" Glaube (?!) an den Summenscore

Meine Frage, warum ein großes Testprogramm über Jahrzehnte ausschließlich Testrohwerte verwendet, wurde damit beantwortet, dass Praktiker oft Agnostiker sind. Als Agnostizismus kann hier die Tendenz bezeichnet werden, bei fehlendem Wissen nichts auf Glauben basieren zu lassen oder keine Annahmen zu machen. Damit war offensichtlich nicht gemeint, dass Praktiker nicht über (testtheoretisches) Wissen verfügen, sondern, dass sie keine Annahmen machen, wenn der einfache Summenscore als Leistungsindikator verwendet wird.

Dies ist (ebenso offensichtlich) nun schon deswegen falsch, weil die Bildung eines Rohwertes die Annahme macht, dass Indikatoren der Antworten auf Testitems addiert werden können.

Eine etwas weniger kleinliche Kritik an dieser Aussage würde dagegen fragen, ob es nicht notwendig ist, erst einmal zu überprüfen, ob jedes Testitem gleichviel zum Summenscore beitragen soll. Sollte nicht eine Mathematikaufgabe, die mehrere Teilschritte umfasst, größeres Gewicht haben als eine Aufgabe, die zwar viel Lesen, aber wenige mathematische Operationen erfordert? Oder sollte letztere gar ganz aus dem Mathematik-Testwert eliminiert werden? Oder sollten die Mathematikaufgaben mit viel Text und jene ohne – oder fast ohne – getrennt aufsummiert werden?

Summenscores sind jedoch ein so offensichtliches Maß der Leistung, oder etwa nicht?

Jech (1983) und Stob (1984) beschreiben ein interessantes Beispiel aus dem Sport: Die Summe der gewonnenen Spiele in einem Badminton-Turnier kann eine sehr gute Beschreibung der Leistung sein. Dies gilt jedoch nur dann, wenn im Laufe des Turniers alle Spieler gegeneinander antreten. Dagegen ist die Summe der gewonnenen Spiele in einem Turnier, das in einer kompletten Vorrunde mit vier Spielern und dann im Knock-out-Verfahren durchgeführt wird, weniger geeignet, die wahre Leistung eines Spielers abzubilden. Ein mittelmäßiger Spieler kann mit etwas Glück in einer Gruppe sein, in der außer ihm nur Anfänger spielen; ein überdurchschnittlicher Spieler kann durch einen ungünstigen Zufall in der Vorgruppe mit den drei stärksten Spielern konkurrieren. Die Summe der gewonnenen Spiele wird für den mittelmäßigen Spieler wahrscheinlich 3 sein, da dieser nur gegen Novizen spielt, der überdurchschnittliche Spieler hingegen wird wahrscheinlich keines oder maximal ein Spiel gegen die Favoriten gewinnen.

Pech gehabt? Oder gibt es in diesem Fall bessere Maße als den Summenscore? Ernst Zermelo (1929) hat eine Lösung für unvollständige (Schach-)Turniere entwickelt, die erstaunliche Ähnlichkeiten mit dem später (1952) als Bradley-und-Terry-Modell für Paarvergleiche bezeichneten Ansatz besitzt:

$$\text{prob}(\text{Spieler j gewinnt gegen Spieler k}) = \frac{T_j}{T_j + T_k} \qquad (2)$$

Dabei beschreibt T_j die Stärke von Spieler j, und T_k die von Spieler k. Diese Spielerstärken sind durch statistische Schätzverfahren zu ermitteln und nicht mit den Summen (der Anzahl) der gewonnenen oder verlorenen Spiele zwischen zwei Spielern zu verwechseln. Sie gelten über die verschiedenen Spieler-Paarungen hinweg und können daher nicht durch einfache Summen oder Mittelwerte ausgedrückt werden. Zermelo (1929) schlägt eine Lösung für die Schätzung dieser Parameter vor, indem er sie als ein Maximierungsproblem der Wahrscheinlichkeitsrechnung beschreibt und somit angibt, wie Maximum-Likelihood-Schätzer zu bestimmen sind.

Was hat das mit psychologischer oder bildungswissenschaftlicher Leistungsmessung zu tun?

Es scheint nahe liegend, eine gewisse Verwandtschaft zwischen Turnieren und allgemeinen Paarvergleichen mit Leistungstests zu unterstellen. Werden in einem Schachturnier anstelle menschlicher Gegner verschiedenen Schachprogramme als Opponenten der Teilnehmer verwendet, so haben wir etwas, das schon sehr wie ein Schachtest aussieht. Es gibt jedoch weitergehende Symmetrien: Schachprogramme haben Spielstärken, die den Schwierigkeitsgraden von Testaufgaben nicht unähnlich sind, und es zeigt sich empirisch, dass sich diese symmetrisch zur Spielerstärke verhalten; ganz so, wie dies auch in modernen Testmodellen der Fall ist.

$$\text{prob}(\text{Spieler j gewinnt gegen Programm k}) = \frac{T_j}{T_j + D_k} \qquad (3)$$

Dabei sind die T_j wie zuvor definiert, D_k bezeichnet die Schwierigkeit (die Spielstärke) des Schachprogrammes k.

Für die Begründung dessen, dass ein Ansatz wie der obige gegenüber einer einfachen Summenbildung zu bevorzugen ist, ist es besonders wichtig, dass Summenscores ohne eine Reihe weiterer Annahmen ungeeignet sind, den Ausgang zukünftiger Spiele (oder zukünftige Testergebnisse) vorherzusagen. Dafür reicht die Summe der gewonnenen Spiele nicht aus.

Wenn ein(e) Spieler(in) nur gegen sehr starke Gegner oder Programme spielt, wird er/sie nur wenige Spiele gewinnen. Spielt er/sie jedoch nur gegen sehr leichte Programme, wird er/sie oft gewinnen. Der Zermelo-Bradley-Terry-Ansatz bezieht die gegnerische (Spiel-)Stärke mit ein wie auch die ermittelten Spielerstärken T_j. Deshalb können die Schachprogramm-Schwierigkeiten D_k sehr verschieden von der totalen Anzahl der gewonnenen Spiele sein. Dies ist ein Vorteil, der kaum zu überschätzen ist.

Unter anderem deswegen ist der Ausgang zukünftiger Spiele mit dem Zermelo-Bradley-Terry-Ansatz sehr einfach vorherzusagen ebenso wie der zwischen Gegnern, die noch nie gegeneinander angetreten sind. Die Gleichungen (2) und (3) lassen sich, sofern die T_j und T_k (oder D_k) aus anderen Spielen mit weiteren

(gemeinsamen) Gegnern bekannt sind, ohne Probleme berechnen. Wie das für Sportrankings in der Praxis gemacht wird, beschreiben z. B. Jech (1983) und Stob (1984). Dies ist ebenfalls nur mit erheblichen und oft nicht sehr plausiblen Annahmen zu machen, wenn wir nur die Gesamtzahl der Siege heranziehen und nicht die Spielerstärken T_j. Doch dazu mehr gegen Ende dieses Kapitels.

Die "naive" Form der Testauswertung kennen wir aus Illustrierten ("Der große Psycho-Test: Sind Sie kreativ?"). Bei der Bezeichnung als agnostisch oder naiv geht es weniger darum, ob der Summenscore das Inferenzziel gut oder weniger gut abbildet, sondern vielmehr darum, ob diese Art der Zusammenfassung der Testleistung untersucht wird, oder gar mit alternativen Formen (z. B. mehrere Summenscores die auf Teilmengen der Aufgaben beruhen, oder gewichtete Summenscores) verglichen wird.

Eine modellbasierte Testauswertung versucht die Antwortindikatoren als solche zu verwenden, d. h., diese als binäre, nominale oder ordinale Daten in Modellen für mehrdimensional kategoriale Daten (z. B. Bishop et al. 1975; Agresti 1990) zu analysieren. Ein wesentliches Charakteristikum dieser Modelle ist es, dass sie das Datenniveau der Antwortindikatoren beibehalten und statistische Strukturen verwenden, die diesem Niveau angemessen sind. Die klassische Testauswertung nimmt im Unterschied dazu an, dass sich die (binären oder ordinalen) Antwortindikatoren näherungsweise wie kontinuierliche Daten verhalten. Dies wird besonders dann zu einer sehr starken Annahme, wenn die (tatsächlich diskrete) Verteilung des Summenscores der Itemantworten durch eine kontinuierliche Verteilung approximiert wird, um den Summenscore auf eine andere Skala zu transformieren. Dies ist eine mit dem tatsächlichen Charakter der Variablen inkonsistente Annahme, die nur näherungsweise für sehr lange Tests unproblematisch ist wenn diese auf einer deutlich gröbere Skala (herunter-)transformiert werden.

Ein weiteres Ziel der modellbasierten Testauswertung besteht darin zu ergründen, ob der einfache Summenscore ausreicht, oder ob besser eine gewichtete Summe gebildet werden sollte, oder ob gar eine ganz andere Aggregation der individuellen Antworten angemessen ist, um die Messintention zu unterstützen. Auch kann durch eine modellbasierte Testauswertung ergründet werden, ob eine Personenvariable (ein Summenscore oder eine andere aus den Itemantworten abgeleitete Statistik) für verschiedene Personengruppen in gleicher Weise geeignet ist, die Testleistung abzubilden. Der wesentliche Unterschied zur agnostischen Testauswertung besteht also darin, dass diese keine direkten Kriterien dafür anlegt, ob die Messintention von dem erzeugten Testwert ausreichend gut erfüllt wird, während die modellbasierte Testauswertung eine statistischen Modellfamilie spezifiziert, um die Messintention auszudrücken. Hambleton & Jones (1993) bezeichnen die klassische – agnostische – naive Testauswertung als schwaches, die modellbasierte als starkes Modell. Als statistische Modellfamilien für die Testauswertung stehen üblicherweise mehrere Modellalternativen zur Verfügung. Durch einen Vergleich dieser Alternativen wird geprüft, wie sich die beobachteten Antworten am besten zusammenfassen lassen, um eine möglichst gute Messung zu erzielen. Es gibt natürlich

weitere Unterschiede zwischen einer eher naiven und einer reflektierten Anwendung und Auswertung von Tests. Diese betreffen z. B. die Repräsentativität der Stichprobe, an der die zentralen Parameter der psychometrischen Güte geschätzt werden; die Repräsentation der Messintention (des Konstruktes) in voller Breite oder die Wiederholbarkeit mit konsistent ähnlichen Resultaten (Reliabilität).

Verlassen wir die Psychotests der Illustrierten und sehen uns die Testauswertung in der Wissenschaft und in Testverlagen oder Testorganisationen an. Die dortige Praxis ist oft etwas weniger agnostisch als es die Tests der Illustrierten vermuten lassen, denn der Bildung eines Summenscores geht oft eine Analyse der Summenbestandteile – die Itemanalyse – in Relation zum (provisorisch definierten) Testwert voraus. Die Itemanalyse ermittelt, ob die Gesamtheit der Items in einem Test "zusammenpasst" oder nicht. Zusammenpassen bedeutet in diesem Kontext die Erfüllung gewisser, oft mathematisch formulierter Kriterien, die sich als Annäherung an Gütekriterien für Tests verstehen lassen. Dabei sind wiederum zwei Niveaus der Anlehnung an ein zugrundeliegendes Modell unterscheiden. In der klassischen Form werden Methoden verwendet, die z. B. der Frage nachgehen, wie sich die Items, aus denen sich der Test zusammensetzt, in ihrer linearen Vorhersagbarkeit (Item-Total-Korrelation, Item-Sumscore-Regression) durch den Summenscore unterscheiden. Wenn sich, wie in der Mehrzahl der Fälle, bedeutsame Unterschiede zeigen, dann werden die Items in zwei Gruppen geteilt: 1) jene, die weiterhin Bestandteil des Tests bleiben, und 2) jene, die nicht ausreichend hoch mit dem Summenscore korrelieren und daher aus der Endversion des Tests eliminiert werden. In der modernen modellbasierten Version dieser Itemanalysen wird ein formale Modellfamilie verwendet, in der es Annahmen darüber gibt, wie sich die bedingte Wahrscheinlichkeitsverteilung der binären, nominalen oder ordinalen Antworten verhält, wenn der wahre Testwert der Personen (der nicht notwendigerweise durch eine einfache Aggregation oder Summenbildung zu erhalten ist) bekannt wäre.

In diesem Papier diskutieren wir einige dieser modellbasierten Prozeduren und zeigen, wie sie sich in der modernen Testtheorie, erweitert um Eigenschaften des Rost-Modells, d. h. des Mischverteilungs-Rasch-Modells (Rost 1990, 1991; Rost & von Davier 1995), in kohärenter und formal eleganter Weise überprüfen lassen. Mittels moderner testtheoretischer Verfahren wird die Entwicklung von Tests mit guten psychometrischen Qualitäten in direkter Anlehnung an ein kohärentes formales Modell ermöglicht.

2 Was soll ein Test leisten?

Ein Test ist eine Auswahl von Aufgaben (bzw. Fragen, Meinungen, Einstellungen oder Verhaltenstendenzen), die Probanden zur Beantwortung vorgegeben wird. Das Ziel eines solchen Experimentes besteht üblicherweise darin, etwas über die Testung Hinausgehendes festzustellen: Entweder ist dies eine Fertigkeit oder Fähigkeit, die als Ursache der Testleistung angesehen wird, oder eine Disposition

oder Persönlichkeitseigenschaft, die dazu führt, dass Probanden in dieser oder jener Art antworten. Die Eigenschaft, die wir erfassen wollen, wird im Folgenden mit dem Buchstaben T bezeichnet. T könnte für Trait stehen oder für True-Score, wir wollen uns jedoch hier noch nicht auf eine spezielle Form festlegen, wohl aber anmerken, dass T nicht mit dem Summenscore S identisch sein muss. In vielen Fällen wird S jedoch eine Größe sein, die uns hilft, T (ab-)zuschätzen, oder S wird zumindest als grobe Näherung von T 'durchgehen' können.

T kann eine ein- oder auch mehrdimensionale Variable sein. Ein guter Test lässt sich recht einfach beschreiben, ebenso wie bedingt gute, perfekte und nutzlose Tests: Ein Test ist

- perfekt, wenn sich durch die alleinige Kenntnis von T die X_i fehlerfrei ausrechnen lassen,
- gut, wenn die X_i sich durch die alleinige Kenntnis von T gut vorhersagen lassen,
- bedingt gut, wenn sich die X_i nur dann gut vorhersagen lassen, wenn neben T noch andere Dinge über die Testpersonen, z. B. deren Geschlecht oder das Bildungsniveau der Eltern, bekannt sind,
- nutzlos (für T), wenn T und X_i nichts miteinander zu tun haben, d. h. wenn die Kenntnis von T nichts zur Vorhersage der Antworten beiträgt.

Weshalb diese Beschreibungen von der Eigenschaft auf die Beobachtungen schließend formuliert sind, sollte selbstverständlich sein, denn die interessierende Eigenschaft ist als Ursache, die Beobachtungen (Antworten auf die Testfragen) sind als Wirkung anzusehen, und nicht umgekehrt. So ist die Intelligenz die Ursache der Leistung in einem Intelligenztest, die Extraversion einer Person die ihres extravertierten Verhaltens, und nicht umgekehrt.

Die Kapitel 3 bis 5 stellen grundlegende und notwendige Annahmen über den Zusammenhang der Eigenschaft T mit den beobachteten Antworten X_i vor, die es ermöglichen, aus den Antworten auf die Qualität eines Tests zu schließen. Bei genauerem Hinsehen zeigt sich, dass nur die wenigsten Tests perfekt sind und die meisten Tests, die oberflächlich gut zu funktionieren scheinen, zumindest einige Aufgaben enthalten, die nur bedingt gut sind. Die einzigartige Leistung des Rost-Modells besteht darin, dass es ermöglicht, selbst aus einem bedingt guten Test noch Schlussfolgerungen auf die Eigenschaft T zu ziehen, um die es bei der Testung geht.

3 Bedingte Unabhängigkeit als notwendige Voraussetzung simultaner Kausalität

Die Erklärungen in Kapitel 2 beinhalten eine bedingte Ursachenzuschreibung: Auch wenn ein Proband bei einem Lesetest mit der einen oder anderen Aufgabe Glück oder Pech haben kann, gehen wir davon aus, dass die gemessene Leistung größtenteils auf seiner Lesefähigkeit beruht und nur wenig von anderen Dingen beeinflusst ist. Das bedeutet, dass sich Probanden mit gleicher Lesefähigkeit nur zufällig

voneinander in ihrer Testleistung unterscheiden sollten, wenn der Lesetest primär nur diese Eigenschaft erfasst.

Das primäre formale Kriterium für simultane Kausalität, d. h. die Zuschreibung, dass mehrere simultan beobachtete Ereignisse eine gemeinsame Ursache haben, ist die bedingte Unabhängigkeit dieser Ereignisse bei gegebener konstanter Valenz der Ursache (Suppes & Zanotti 1981, 1995). Diese etwas gestelzt wirkende Aussage ist durchaus plausibel, wenn sie intuitiv formuliert wird: Die gemeinsame Ursache von Ereignissen sollte die beobachteten Korrelationen zwischen ihnen erklären. Auf unser Beispiel bezogen sollen die Zusammenhänge zwischen den Aufgaben verschwinden, wenn wir nur Personen mit gleicher Lesefähigkeit betrachten, da die beobachteten Unterschiede dann nur zufälliger Natur sein sollten.

Ein fast schon stereotypes Beispiel dafür, wie dieses Prinzip NICHT anzuwenden ist, sind die positiven Korrelationen zwischen der Anzahl der Bücher im Elternhaus (manchmal etwas hochtrabend als cultural capital bezeichnet) und der Schulleistung von Kindern, denn die Anzahl der Bücher wird sehr wahrscheinlich auch positiv korreliert sein mit dem Preis der Armbanduhr (oder der Möbel, des Autos, des Hauses) der Eltern. Die Größe der Bibliothek ist sicher nicht die Ursache der Schulleistung oder des Preises der Besitztümer. Auch sind die Ursache(n) der Korrelationen nicht einfach individuell identifizierbare Variablen, sondern ein Konglomerat, welches zumindest den Bildungsstand und/oder den Beruf der Eltern einschließt. Auch wird eine komplette Aufklärung der gemeinsamen Varianz aller beobachteten Größen nicht möglich sein. Es scheint jedoch klar, dass jede der drei zufällig ausgewählten Variablen Schulleistung, Bücher im Haus, Preis der Armbanduhr(en) der Eltern die beiden anderen nicht kausal beeinflussen kann, wohl aber mit diesen positiv korreliert sein wird.

Selbst wenn der Bildungsstand der Eltern die genannten Korrelationen verschwinden ließe, wäre dies kein Beweis für einen kausalen Zusammenhang: Die bedingte Unabhängigkeit ist eine notwendige, jedoch keine hinreichende Bedingung für simultane Kausalität. Daher dürfen diese Ausführungen nicht dahingehend falsch verstanden werden, dass wir die Hypothese aufstellen wollen, dass der Bildungsstand der Eltern einen direkten Einfluss auf die drei Variablen ausübt oder gar deren unmittelbare Ursache ist. Hier geht es nur um ein Beispiel, in dem die positive Korrelation dreier scheinbar sehr unterschiedlicher Variablen plausibel durch dahinterliegende Variablen (z. B. Bildungsstand der Eltern) erklärt werden kann, der direkte Pfad von der Anzahl der Bücher zur Schulleistung jedoch nicht aufrecht zu erhalten ist.

Etwas realistischer ist eine Situation, in der eine Reihe von gut strukturierten Aufgaben ähnlichen Inhaltes betrachtet wird, die von einer Zufallsstichprobe von Testpersonen bearbeitet wird. Wir können annehmen, dass es eine gemeinsame Ursache für die beobachteten Ereignisse (Aufgabe i von Person v gelöst bzw. nicht gelöst) gibt. Dies kann entweder eine Fertigkeit sein, die sich nur auf diesen Aufgabentyp bezieht, oder eine breiter definierte Fähigkeit, die es erlaubt, auch andere Arten von Leistungen zu erklären. Für eine solche Fähigkeit muss nach Sup-

pes & Zanotti gelten, dass diese Variable, wenn sie konstant gehalten wird, zur Unabhängigkeit bei der Aufgabenbearbeitung führt, da die Unterschiede nur zufälliger Natur sein sollten. Ist bedingte Unabhängigkeit gegeben, dann verschwinden die Korrelationen zwischen den Antworten, wenn wir nur Personen mit gleicher Fähigkeit betrachten. Dies ist jedoch nur dann sinnvoll, wenn zuvor, d. h. wenn die Fähigkeit nicht konstant gehalten wird, die simultan beobachteten Ereignisse nicht unabhängig sind. Es muss bereits Kontingenz in der Population vorliegen, um eine gemeinsame ursächliche Variable annehmen zu können. Wohlgemerkt, dies ist nur eine notwendige, aber keine hinreichende Bedingung für eine Ursachenzuschreibung. Wenn diese jedoch nicht gegeben ist, und das ist der zentrale Punkt dieser Betrachtungen, dann kann die untersuchte Variable nicht ursächlich mit der Aufgabenbearbeitung zusammenhängen.

Diese führt zu einem ersten (überprüfbaren) Kriterium, das erfüllt sein muss, wenn wir eine Fähigkeitsvariable wie z. B. den Summenscore auf ihre Güte für die Erklärung von Testleistungen hin untersuchen wollen.

Postulat bedingte Unabhängigkeit: Eine Fähigkeits- oder Fertigkeitsvariable T, die als mögliche Ursache für den Erfolg von Testpersonen bei der Lösung einer Serie von Testaufgaben X_i gelten soll, muss das Kriterium der bedingten Unabhängigkeit der Testaufgaben erfüllen. Formal:

$$\text{prob}\left(X_1, X_2, \ldots, X_m \middle| T\right) = \prod_{i=1}^{m} \text{prob}\left(X_i \middle| T\right) \tag{4}$$

Wie wird überprüft, wenn eine solche Variable angenommen wird oder wenn gar der Summenscore der gelösten Testaufgaben diese ursächliche Rolle annehmen soll?

1. In der klassischen Testanalyse werden die Variablen X_i auf ihre Korrelation mit dem Summenscore S hin untersucht, denn wenn diese Korrelationen hoch sind, kann angenommen werden, dass die Korrelationen zwischen den Einzelvariablen verschwinden, wenn S konstant gehalten wird. Alternativ können auch die Testpersonen nach ihrem Summenscore S geordnet werden. Dann wird geprüft, ob die Korrelationen aller Paare von Antwortvariablen (X_i , X_j) tatsächlich verschwinden, d. h., ob die Bedingung auf die Summenvariable S die Kovarianz zwischen den Items 'klein' wird. Was 'klein' oder 'verschwinden' heißt, muss (z. B. in einer Faustregel) festgelegt werden.

2. In der modernen Testauswertung werden formale Modelle formuliert, die bedingte Unabhängigkeit für eine einfache Variable (z. B. den Summenscore S) annehmen, die als Bedingung konstant gehalten wird. Diese Annahme wird mit Hilfe statistischer Methoden mit allgemeineren Modellen verglichen, z. B. mit Modellen, die eine gewichtete Summe annehmen oder die neben dem Summenscore jeder Testperson noch andere Personeneigenschaften berücksichtigen; oder mit Modellen, die anstelle einer eindimensionalen Fähigkeit eine mehrdimensionale Variable (z. B. mehrere gewichtete Teilsummen) annehmen.

In beiden Fällen wird überprüft, ob bei gegebener (festgehaltener) Fähigkeit die Zusammenhänge zwischen den Antworten (annähernd) verschwinden. Wird die Einflussgröße, die in erster Linie für die Unterschiede verantwortlich ist, gefunden und konstant gehalten, so sollten die verbleibenden Unterschiede nur zufällig sein. Lägen keine unkontrollierbaren Fehlerquellen vor, die die Antworten beeinflussen, dann wären diese sogar perfekt vorhersagbar (determiniert). Dann gäbe es zwischen Probanden mit gleicher Fähigkeit überhaupt keine Unterschiede in ihren Antworten, sofern diese Einflussgröße die einzige Ursache der Antworten ist.

Populationen, Gruppenunterschiede und Störvariablen

Die in Kapitel 2 als bedingt gut bezeichneten Tests sind aus dem folgenden Grunde problematisch: Wenn neben der Eigenschaft T noch eine weitere Variable (z. B. Geschlecht) die Testleistung beeinflusst, wie sollen dann gleiche Leistungen (z. B. gleiche Antworten oder Summenscores) interpretiert werden? Es führte sicher zu berechtigtem Protest, wenn ein Lehrer sagte: "Ja, Franziska, Du hast auch 12 Punkte, genauso wie Fritz, aber für Mädchen ist das Lesen dieser Aufgaben einfacher, daher bekommst Du ein 'Gut' und Fritz ein 'Sehr gut'." Dies würde weder von Schülern noch von Lehrern oder Eltern akzeptiert. Das zeigt, dass ein guter Test gebraucht wird, der es erlaubt, für Mädchen und Jungen gleichermaßen von der Leistung auf die Lesefähigkeit zurückzuschließen.

Postulat Populationsunabhängigkeit bzgl. Z: Die Wahrscheinlichkeit der Antwortvariablen X_i hängt nur von der Eigenschaftsvariablen T ab, nicht aber von Variable Z. Formal lässt sich dies schreiben als

$$\text{prob}\left(X_1,...X_m \middle| T, Z\right) = \text{prob}\left(X_1,...,X_m \middle| T\right). \tag{5}$$

Diese Unabhängigkeitseigenschaft wird oft implizit für alles angenommen, d. h. es wird nicht Fall für Fall geprüft, ob z. B. die Verteilung der Antworten für Frauen und Männer mit gleicher Eigenschaftsausprägung anders ist (Z wäre hier das Geschlecht). Ebenfalls wird üblicherweise nicht getestet, ob z. B. die Leistung in einem Mathematiktest nicht nur von der Mathematikfähigkeit (T), sondern auch von anderen Faktoren wie etwa der Extraversion der Probanden abhängt (Z wäre hier die wie auch immer gemessene Extraversionsneigung). Ähnliches gilt für Schuhgröße, Haarfarbe, Blutgruppe etc., da man davon ausgehen kann, dass diese Variablen nicht mit der Testleistung zusammenhängen.

Es gibt jedoch Variablen, für die nicht unbedingt Populationsunabhängigkeit angenommen werden kann. Als Beispiel kann die Leistung in einem Mathematiktest gelten, der auch Textaufgaben enthält, und die Muttersprache der Schüler. Z wäre in diesem Beispiel 1 bei Schülern, deren Muttersprache mit der Sprache des Tests übereinstimmt, und 0 bei jenen, die eine andere Erstsprache sprechen.

Wir nehmen zur Vereinfachung an, dass die Aufgaben nur gelöst (X = 1), oder nicht gelöst (X = 0) werden, d. h., dass alle beobachteten Variablen binär sind. Wenn es eine zusätzliche Abhängigkeit der Aufgabenlösung von der Muttersprache

gibt, dann vermuten wir, dass für alle Items i, die textbasierte Mathematikaufgaben sind, der folgende Zusammenhang gilt:

$$\text{prob}\left(X_i = 1 \middle| T = t, Z = 1\right) > \text{prob}\left(X_i = 1 \middle| T = t, Z = 0\right). \quad (6)$$

Dies bedeutet, dass bei gleicher Mathematikkompetenz Muttersprachler mit einer Fähigkeit von T= t eine höhere Wahrscheinlichkeit als Nicht-Muttersprachler haben, Textaufgaben zu lösen.

Ist diese Annahme sinnvoll? Wenn nicht, da wir möglichst 'reine' Tests haben wollen, dann ist sie jedoch wenigstens zu überprüfen. Idealerweise sollte gelten, dass beide bedingten Wahrscheinlichkeiten gleich sind, d. h.

$$\text{prob}\left(X_i = 1 \middle| T = t, Z = 1\right) = \text{prob}\left(X_i = 1 \middle| T = t, Z = 0\right). \quad (7)$$

Folglich ist die Lösungswahrscheinlichkeit für alle Items i gleich für Personen mit gleichen Fähigkeiten, ungeachtet sonstiger Unterschiede bei anderen Variablen.

Steigende Erfolgschancen oder monotone Itemfunktionen

Ein Hauptgrund für die Anwendung von Testverfahren besteht darin, dass die Ergebnisse Rückschlüsse auf eine interessierende Personeneigenschaft zulassen sollen. Dafür ist es notwendig, dass die Beziehung zwischen der Testleistung und der Eigenschaft so eindeutig wie nur möglich ist. Wenn die Eigenschaftsausprägung zweier Personen bekannt ist, dann sollte diejenige mit der höheren Ausprägung eine größere Chance haben, die Testaufgaben zu lösen. In der klassischen Testauswertung wird diese Erwartung oft mittels der Berechnung der Korrelationen zwischen Testaufgabe und dem Gesamttestwert überprüft. Leider ist es jedoch nicht ganz einfach festzulegen, was dabei eine 'gute' (ausreichende) oder eine 'weniger gute' (zu schwache) Korrelation ausmacht. Zudem ist die eine Variable eine diskrete, ordinale; die andere binär. Eine einfache Produkt-Moment-Korrelation ist daher nicht ohne Schwierigkeiten zu interpretieren. Dies ist ein weiteres Beispiel dafür, dass die klassische Testauswertung im Wesentlichen die gleichen Ideen wie die modellbasierte verfolgt, die klassische Variante jedoch oft mit Faustregeln arbeiten muss, während die modellbasierte, wie in Kapitel 4 erläutert wird, formale Kriterien entwickeln kann.

Monotonie-Bedingung: Wenn Personen eine Auswahl von Testitems lösen, die einen gemeinsamen Inhaltsbereich abdecken, dann wird angenommen, dass die Antworten von einer (oder einigen wenigen) gemeinsamen Personenvariablen abhängen. Formal bedeutet dies, dass die Antwortvariablen von einer (oder einigen wenigen) nicht direkt beobachtbaren Variablen abhängen. Wir haben diese Variable in den vorherigen beiden Annahmen als T bezeichnet und nicht näher spezifiziert. Oft wird jedoch darüber hinaus angenommen, dass T eine Variable ist, die eine Ordnung besitzt, d. h., dass für zwei Personen mit den Ausprägungen s und t (beide Realisierungen von T) festgestellt werden kann, ob s größer ist als t

(s > t), kleiner (s < t) oder gleich (s = t). Dies ist eine so grundlegende Annahme, dass sie oft nicht hervorgehoben wird, sondern implizit in anderen, spezifischeren enthalten ist. Für eine Vielzahl von Methoden zur Testauswertung nimmt diese Annahme der Wohlgeordnetheit zusammen mit einer Annahme über die Wahrscheinlichkeiten die Form einer positiven Assoziations- oder einer (stärkeren) Monotonie-Bedingung an:

$$\text{prob}\left(X_i = 1 \middle| T = s\right) > \text{prob}\left(X_i = 1 \middle| T = t\right) \quad \text{wenn } s > t \,. \tag{8}$$

In dieser Gleichung wird gefordert, dass für alle Items i die Lösungswahrscheinlichkeit für Personen mit einer höheren Fähigkeitsausprägung (s) höher sein soll als für solche mit einer geringeren Ausprägung (t < s). Dies gilt für alle Testitems i gleichermaßen, d. h. Gleichung (8) soll insbesondere auch unabhängig von der Schwierigkeit der Items sein.

In vielen formalen Testmodellen ist die Annahme über T dahingehend realisiert, dass die Fähigkeiten s, t etc. der Personen als reelle Zahlen angesehen werden können. Zudem wird eine ähnliche Annahme über die Schwierigkeiten der Testitems gemacht, die z. B. als Schwierigkeit b_i von Item i , b_k von Item k bezeichnet werden und von denen angenommen wird, dass sie auch Realzahlen sind. Bei der Schwierigkeit von Testaufgaben wird jedoch von einer monoton fallenden Beziehung ausgegangen, d. h. mit wachsender Schwierigkeit wird eine erfolgreiche Lösung weniger wahrscheinlich.

4 Wie fügen sich diese Annahmen zusammen, oder: Warum all dies?

Die vorgestellten Annahmen sind zentral für das konzeptuelle Verständnis moderner Testauswertung. Sie sind bereits der klassischen Variante implizit, dort jedoch nicht formalisiert. In diesem Kapitel beschreiben wir, inwieweit sie sich in einem weit verbreiteten Testmodell realisieren und überprüfen lassen.

Wir nehmen deshalb an, dass jede von j =1 … n Personen einen Teil von i = 1 … m Aufgaben bearbeitet. Wir erlauben, dass nur einige oder alle Aufgaben bearbeitet werden können, stellen jedoch die Bedingung, dass jede Person mehrere Aufgaben bearbeiten muss und jede Aufgabe von einem Großteil der Personen bearbeitet wird (formeller müssen wir annehmen, dass jede Person mit jeder anderen zumindest indirekt einige Aufgaben gemeinsam hat; wenn zwei Personen völlig verschiedene Aufgaben bearbeiten, so muss es andere geben, die Teile von jeder dieser unterschiedlichen Teilmengen bearbeiten). Wir nehmen an, dass sich die Probanden nur einmal oder aber mehrfach an den Aufgaben messen können, und entwickeln ein Modell, das eine Personenfähigkeit t_j für jede Person j und eine Aufgabenschwierigkeit b_i für jede Aufgabe i beinhaltet und das die oben eingeführten Annahmen realisiert. Wir bezeichnen mit x_{ji} die Anzahl der erfolgreichen, mit y_{ji} die Anzahl der erfolglosen Versuche von Person j, die Aufgabe i zu lösen.

Die Wahrscheinlichkeit, genau x_{ji} Erfolge und y_{ji} Misserfolge zu beobachten, kann als Binomialterm mit genau $n_{ij} = x_{ij} + y_{ij}$ Versuchen geschrieben werden, in dem wir nur Faktoren ausschreiben, die von der Personenfähigkeit t_j und der Aufgabenschwierigkeit b_i abhängen:

$$\text{prob}\left(X_{ij} = x_{ij} \middle| t_j, b_i, n_{ij}\right) = f\left(x_{ij}, n_{ij}\right)\left(\frac{t_j}{t_j + b_i}\right)^{x_{ij}}\left(\frac{b_i}{t_j + b_i}\right)^{n_{ij}-x_{ij}} \tag{9}$$

Beschränken wir uns auf einen Versuch, d. h., $1 = n_{ij} = x_{ij} + y_{ij}$, so dass jede Person nur einmal mit jeder Aufgabe konfrontiert wird, so erhalten wir

$$\text{prob}\left(X_{ij} = x_{ij} \middle| t_j, b_i\right) = \left(\frac{t_j}{t_j + b_i}\right)^{x_{ij}}\left(\frac{b_i}{t_j + b_i}\right)^{1-x_{ij}}. \tag{10}$$

Die Wahrscheinlichkeit, dass Person j Aufgabe i löst, ist folglich

$$\text{prob}\left(X_{ij} = 1 \middle| t_j, b_i\right) = \frac{t_j}{t_j + b_i} = \frac{\frac{t_j}{b_i}}{1 + \frac{t_j}{b_i}}, \tag{11}$$

und die Wahrscheinlichkeit, dass Person j das Item i nicht löst, ist

$$\text{prob}\left(X_{ij} = 0 \middle| t_j, b_i\right) = 1 - \text{prob}\left(X_{ij} = 1 \middle| t_j, b_i\right) = \frac{b_i}{t_j + b_i} = \frac{1}{\frac{1 + t_j}{b_i}}. \tag{12}$$

Diese beiden elementaren Wahrscheinlichkeiten sind so in Zermelo (1929) und Rasch (1960) zu finden. Der Unterschied besteht lediglich darin, dass bei Zermelo Spieler gegen Spieler in (Schach)Turnieren und bei Rasch Testpersonen gegen Aufgaben antreten. Typischerweise wird das Rasch-Modell in logistischer Schreibweise dargestellt, was durch die Transformation $\theta_j = \ln(t_j)$ und $\beta_i = \ln(b_i)$ ermöglicht wird. Es gilt dann

$$\text{prob}\left(X_{ij} = 1 \middle| \theta_j, \beta_i\right) = \frac{\exp\left(\theta_j - \beta_i\right)}{1 + \exp\left(\theta_j - \beta_i\right)} = \frac{\frac{t_j}{b_i}}{1 + \frac{t_j}{b_i}}. \tag{13}$$

Obwohl der Ansatz bereits 1929 entwickelt wurde, ist er noch heute *die* Methode der Wahl bei Paarvergleichen und für das Ranking von Spielern/Sportlern in verschiedenen Sportarten. Auch das Rasch-Modell wird in Testanalysen in Psychologie und Bildungswissenschaft oft verwendet.

Was macht diesen Ansatz so attraktiv? Wenn wir davon ausgehen, dass ein Spieler mehrfach gegen ein Programm oder einen anderen Spieler antreten kann, ohne dass sich die Chancen dramatisch ändern (d. h. dass Spieler- und Gegnerstärken relativ konstant bleiben), dann können wir einen naiven Schätzer der Erfolgswahrscheinlichkeit in der Form

$$\text{prob}(\text{Spieler j gewinnt gegen i}) \approx \frac{x_{ji}}{x_{ji} + y_{ji}} \tag{14}$$

schreiben. Dies ist schon fast die Gleichung (11), nur dass der Bruch in Gleichung (14) doppelt indiziert ist, denn die x_{ji} und y_{ji} sind spezifische Zahlen für die beobachteten Erfolge und Misserfolge, wogegen die Spielerstärken t_j und Programm- oder Gegnerschwierigkeiten b_i unabhängig von dem jeweiligen Opponenten geschrieben werden. Unser Modellansatz ist also

$$\text{prob}(\text{Spieler j gewinnt gegen i}) = \frac{t_j}{t_j + b_i} \approx \frac{x_{ji}}{x_{ji} + y_{ji}} \,. \tag{15}$$

Dabei wird angenommen,

- dass über die Gegner (oder Programme) b_i hinweg die gleiche Fähigkeitsmaßzahl t_j verwendet werden kann, um die Erfolgswahrscheinlichkeiten zu beschreiben;
- dass sich die Fähigkeit eines Gegners (bzw. die Spielstärke des Opponenten oder die Schwierigkeit des Schachprogrammes) nicht ändert, wenn ein anderer Spieler antritt. Dies ist recht einleuchtend, wenn davon ausgegangen werden kann, dass Computerprogramme ihre Spielstärke nicht automatisch dem Spieler anpassen und dass Spieler mit einer gewissen Stärke ihr Spielverhalten durch die Konfrontation mit einem neuen Programm nicht grundlegend ändern. Dann kann näherungsweise angenommen werden, dass die nachfolgenden Spiele nicht wesentlich von den vorhergehenden beeinflusst sind.
  Für Testdaten wird üblicherweise davon ausgegangen, dass dies ebenfalls plausibel ist, solange Aufgaben nicht wiederholt beantwortet werden, denn eine Person kann nicht mehrfach die gleiche Testaufgabe bearbeiten, ohne dass Erinnerungseffekte erwartet werden müssen. In diesen Fällen ist nur eine Wiederholung möglich, d. h., es kann die Gleichung (10) verwendet werden.

Das Rasch-Modell erfüllt die technischen Annahmen

Die einführend dargestellten Annahmen sind im Ansatz von Rasch bzw. Zermelo in der folgenden Form erfüllt:

Eindimensionalität/Populationsunabhängigkeit: Die Fähigkeitsparameter t_j (oder θ_j) sowie die Schwierigkeitsparameter b_i (oder β_i) bestimmen die Lösungswahrscheinlichkeiten vollständig. Dies bedeutet, dass die Antwortwahrscheinlichkeiten ausschließlich von θ_j und β_i abhängen, nicht jedoch von weiteren Variablen.

Für Testdaten wird üblicherweise davon ausgegangen, dass diese Annahme ebenfalls plausibel ist, solange die Testaufgaben logisch unabhängig voneinander sind und nicht wiederholt bearbeitet werden. Gilt die Annahme der Populationsunabhängigkeit nicht, so gibt es weitere Variablen (personen- oder gruppenbezogene Größen), die sich auf die Erfolgswahrscheinlichkeit von Personen oder Personengruppen auswirken, einige oder alle Aufgaben lösen zu können. Nehmen wir an, dass jede Person in der Stichprobe nicht nur durch ihre Fähigkeit θ_j, sondern auch durch eine weitere Variable Z beschrieben wird, so gilt bei einer Verletzung der Populationsunabhängigkeit

$$\text{prob}\left(X_i = 1 \middle| \theta, Z = 1\right) \neq \text{prob}\left(X_i = 1 \middle| \theta, Z = 0\right). \tag{16}$$

Diese Ungleichung bedeutet, dass die Lösungswahrscheinlichkeit nunmehr nicht mehr nur von einer Variablen (θ_j), sondern von zweien (θ_j, Z) abhängt. Dies bedeutet ebenfalls, dass die Eindimensionalität verletzt ist (auch wenn die Variable Z nur eine Dichotomie darstellt, ist sie technisch dennoch als zweite Dimension anzusehen, auf der sich die untersuchten Personen unterscheiden können).

Wenn jedoch die Annahme der Populationsunabhängigkeit gilt, so können wir schreiben:

$$\text{prob}\left(X_i = 1 \middle| \theta, Z = 1\right) = \text{prob}\left(X_i = 1 \middle| \theta, Z = 0\right) = \text{prob}\left(X_i = 1 \middle| \theta\right). \tag{17}$$

D. h. die Wahrscheinlichkeit einer korrekten Antwort ist nur von der Personenfähigkeit (und der Itemschwierigkeit) abhängig, nicht jedoch von weiteren Variablen Z.

Lokale Unabhängigkeit: Bereits der Ansatz von Zermelo nimmt lokale Unabhängigkeit an. Im Rasch-Modell wird die gleiche Annahme gemacht. Daraus folgt

$$\text{prob}\left(x_{j1},\ldots,x_{jm} \middle| \theta_j\right) = \prod_{i=1}^{m} \frac{\exp\left(x_{ji}\left(\theta_j - \beta_i\right)\right)}{1 + \exp\left(\theta_j - \beta_i\right)} \tag{18}$$

für die Wahrscheinlichkeit einer Sequenz von Antworten $x_{j1},\ldots,x_{jm}$ bei gegebener Personenfähigkeit θ_j, was u. a. bedeutet, dass vorhergehende oder nachfolgende Antworten sich nicht gegenseitig beeinflussen.

Strikte Monotonie: Mit wachsender Personenfähigkeit t_j wächst der Bruch in Gleichung (11) gegen Eins, d. h. die Lösung der Aufgabe wird mit steigender Fähigkeit immer wahrscheinlicher. Ebenso wird mit wachsender Aufgabenschwierigkeit b_i die Wahrscheinlichkeit einer korrekten Lösung immer unwahrscheinlicher, der Wert des Bruches geht gegen Null.

5 Das Rost-Modell

Die Kapitel 3 und 4 sollten klar gemacht haben, dass Testmodelle Annahmen machen. Je expliziter diese sind oder je expliziter ein Modell ausgeschrieben wird, des-

to besser lässt sich empirisch überprüfen, ob ein Modell für einen gegebenen Test angemessen ist.

Zwei unserer Beispiele haben verdeutlicht, dass die Annahmen fallspezifisch plausibel oder weniger plausibel sein können:

1. Das Turnier, in dem Spieler gegen Schachprogramme antreten, zeigt, dass die Schwierigkeit einer Aufgabe nicht von der Person abhängt, die diese bearbeitet: Ein Schachprogramm sollte nicht per se für Spieler gleicher Schach-Spielstärke unterschiedlich schwer sein, selbst wenn sie sich in anderen Variablen unterscheiden.
2. Das Beispiel Textaufgaben in einem Mathematiktest macht nachvollziehbar, dass die Erfolgswahrscheinlichkeit von mehr abhängen kann, als es zunächst erscheinen mag: Die Lösung mathematischer Textaufgaben erfordert nicht nur Mathematikfähigkeit, sondern auch das Verständnis des Textes, der die Aufgabe stellt. Zwei Schüler mit gleicher Mathematikfähigkeit, aber verschiedenenen Muttersprachen können sehr unterschiedliche Erfolgschancen bei der Bearbeitung von Textaufgaben haben.

Die zentrale Annahme, die mittels der Beispiele nahegebracht bzw. hinterfragt werden sollte, ist die der Populationsunabhängigkeit der modellbasierten Lösungswahrscheinlichkeiten. Andere hinterfragte Annahmen waren z. B. die der monotonen Itemfunktionen (z. B. Roberts 2001, Rost & Luo 1997, Andrich 1996) oder der lokalen stochastischen Unabhängigkeit (z. B. Haberman 2007).

Die Populationsunabhängigkeit ist jedoch, wie es neudeutsch heißt, besonders tricky, denn die potenzielle Anzahl der Störvariablen, die diese verletzen können, ist sehr groß. Beispiele dafür sind, wie eben noch einmal erwähnt, unterschiedliche Muttersprachen, aber auch Testmotivation oder Testvorbereitung, die gewisse Items oder Itemtypen trainiert; die illegale Veröffentlichung von Testitems auf Webservern sowie unterschiedliche Lösungsstrategien, unabhängig davon, ob trainiert oder nicht.

Da wir nicht alle Störvariablen testen wollen (oder können), sind Lösungen gefragt, die Aussagen darüber treffen, ob potenziell Gruppierungen existieren, die die Annahme der Populationsunabhängigkeit verletzen.

Hier kommt das Rost-Modell (von Jürgen Rost [1990] als Mischverteilungs-Rasch-Modell bezeichnet) ins Spiel: Nehmen wir an, dass es eine Störgröße gibt (geben könnte) und nennen diese c = 1,..,C. Dieses c steht für eine Klassifikation (class membership) bezüglich einer nicht beobachteten Gruppierungsvariablen. Jede Person in der beobachteten Stichprobe ist durch zwei Variablen charakterisiert: die Fähigkeit θ_j, die es zu schätzen gilt, und die Gruppenzugehörigkeit c_j, die die Lösungswahrscheinlichkeiten ggf. modifizieren kann. Für eine Person mit der Fähigkeit θ in Gruppe c wird angenommen, dass die Wahrscheinlichkeit einer Lösung gegeben ist durch

$$\text{prob}\left(X_{ij} = 1 \mid \theta, c, \beta_i\right) = \frac{\exp\left(\theta - \beta_{ic}\right)}{1 + \exp\left(\theta - \beta_{ic}\right)} \,. \tag{19}$$

Damit werden die Itemschwierigkeiten folgendermaßen modifiziert: Anstelle eines Schwierigkeitsparameters β_i gibt es nun für jede Gruppenzugehörigkeit c einen besonderen Parameter β_{ic}. Dies ermöglicht es, dass in jeder der latenten Gruppen andere Itemschwierigkeiten gelten können, womit das Rost-Modell hierarchisch über der Latent-Class-Analyse und dem Rasch-Modell steht und beide umfasst (vgl. Rost 1990, 1991). Wenn das Rost-Modell auf einen Datensatz angewendet wird, werden jeder Person Schätzwerte bezüglich zweier Variablen zugeordnet:

1. eine Schätzung der Wahrscheinlichkeit der Zugehörigkeit zu jeder der (zwei oder mehr) latenten Klassen, in denen das Rasch-Modell mit unterschiedlichen Parametern gilt, sowie
2. eine Schätzung der wahrscheinlichen Lage der Fähigkeit und eine Schätzung der Messgenauigkeit dieser Schätzung bei gegebenen Antworten und für jede der latenten Gruppierungen.

Mit dem Rost-Modell wird es möglich zu ermitteln, welche Gruppierungen anzunehmen sind, um die Daten angemessen beschreiben zu können (von Davier 1997), und wie die wahrscheinlichste Eigenschaftsausprägung einer Person bei gegebenen Antworten ist.

Im Mathematiktest-Beispiel hat jede Textaufgabe für Mutter- und Nicht-Muttersprachler einen anderen Schwierigkeitsgrad. Die besonderen Schwierigkeitsparameter können, müssen aber nicht, für jede Gruppe c unterschiedlich sein. Wichtig ist, sich klar zu machen, dass sowohl die Fähigkeit θ_j, die es zu schätzen gilt, als auch die Gruppenzugehörigkeit c_j, Personenparameter sind. Das Rost-Modell nimmt also an, dass es zwei Größen gibt, die von Person zu Person unterschiedlich sein können, und kann deshalb auch als mehrdimensionales IRT-Modell (Rijmen & De-Boeck 2003) oder als diagnostisches Modell (von Davier 2005, Jiao et al. 2010) geschrieben werden.

Es bietet eine Lösung für Tests, die bedingt gut sind und Punkt 4 meiner Aufstellung in Kapitel 2 "Was soll ein Test leisten?" repräsentieren. Das Rost-Modell erlaubt es, Gruppen zu identifizieren, in denen der Test gut funktioniert, in denen insbesondere das Rasch-Modell gilt. Ohne diese Identifikation können Personen nicht miteinander verglichen werden, da nicht bekannt ist, in welcher Art die Items mit der Personenfähigkeit interagieren, da ohne Gruppenzugehörigkeit die Itemschwierigkeiten unbekannt sind.

6. Ausblick

Das hier als Rost-Modell bezeichnete Mischverteilungs-Rasch-Modell hat in den ca. 20 Jahren seit seiner Veröffentlichung ausgesprochen großes Interesse in der psychometrischen Forschung gefunden und zu Innovationen in der persönlichkeits-

psychologischen, kognitiv psychologischen und klinisch psychologischen Forschung geführt. Zudem gibt es eine Reihe methodischer Entwicklungen (Rost & von Davier 1995, von Davier & Rost 1995, Smit et al. 1999, 2000; Bolt et al. 2002), die sich daraus ableiten. Einen Überblick über diese und neuere Entwicklungen geben die von Hancock & Samuelson (2006) oder von von Davier & Carstensen (2007) herausgegebenen Sammelbände; aber auch der breiter angelegte Band von DeBoeck & Wilson (2004) und ein Übersichtskapitel (von Davier 2009a) greifen Erweiterungen von Mischverteilungsmodellen für Testdaten auf und beschreiben Anwendungsbeispiele.

Meine wissenschaftliche Laufbahn ist deutlich von dem Rost-Modell geprägt worden. Ich hatte das Glück, sehr zeitig mit Jürgen Rost zusammen an der Entwicklung der Software dafür zu arbeiten und gemeinsam mit ihm Anwendungen des Mischverteilungs-Rasch-Modells zu untersuchen (Rost & von Davier 1993). Zudem hat mich diese Arbeit dazu bewogen, danach allgemeinere Item-Response-Modelle zu Mischverteilungsmodellen zu erweitern (von Davier & Yamamoto 2004, von Davier 2007a), und schließlich Multilevel-Erweiterungen der Latent-Class-Analyse (Vermunt 2003) an diese Verallgemeinerungen anzupassen (von Davier 2007b, 2009b, 2010).

Ich bin sicher, dass das Rost-Modell ein sehr aktives Forschungsfeld bleiben wird, das für Anwender und Theoretiker der Testtheorie und Psychometrie von großem Interesse ist.

Scorebildung, klassische Testtheorie und Item-Response-Modelle in der Einstellungsmessung[1]

Wilhelm Kempf

1 Ordinale und strikte Homogenität

Einstellungen sind nicht direkt beobachtbar. Sie manifestieren sich jedoch in der Zustimmung zu oder Ablehnung von bestimmten Aussagen, die als Indikatoren für die Ausprägung einer Einstellung dienen können. Dies macht man sich in der Einstellungsmessung zunutze, indem man den Probanden eine Reihe von Aussagen (Items) vorlegt, zu denen sie auf einer mehrstufigen Likert-Skala Zustimmung oder Ablehnung artikulieren können.

Von diesen Antworten wird auf die zugrunde liegende (latente) Einstellung der Probanden zurückgeschlossen, wobei man in der Regel den aus den Itemscores berechneten Summenscore als Maßzahl für die zu messende Einstellung verwendet.

Auf den ersten Blick erscheint dieses Vorgehen durchaus plausibel. Dass es auch zweckmäßig ist, von den Itemantworten auf eine (quantitative) latente Einstellungsdimension schließen zu wollen, auf der man die Probanden gemäß ihrem Summenscore anordnen kann, liegt jedoch keineswegs auf der Hand. So hat bereits Kracauer (1952) darauf aufmerksam gemacht, dass es in der Inhaltsanalyse nicht die Häufigkeit bestimmter Textmerkmale ist, welche die Bedeutung eines Textes ausmacht, sondern dass es die Muster sind, welche sie bilden, und ähnliches gilt möglicherweise auch für die Einstellungsmessung.

Der Summenscore liefert zwar eine (mehr oder minder angemessene) Quantifizierung des Ergebnisses der Befragung. Diese beschreibt aber zunächst nur die manifesten Antworten der Probanden, (noch) nicht jedoch die zugrunde liegende Einstellung, für welche die Antworten als Indikatoren dienen sollen. Ja, es ist noch nicht einmal ausgemacht, dass der Summenscore eine brauchbare Beschreibung des Befragungsergebnisses darstellt. Es wäre durchaus denkbar, dass manche Antworten stärker zu gewichten sind als andere. Oder es wäre auch denkbar, dass es so etwas wie eine generelle Einstellungsdimension, auf der sich die Probanden anordnen lassen, gar nicht gibt. Vielleicht gibt es stattdessen verschiedene Muster von *qualitativ* verschiedenen Einstellungen.

1 Gefördert von der Deutsche Forschungsgemeinschaft (DFG), Kenzeichen KE 300/8-1.

Als Indikatoren für ein und dieselbe quantitative Einstellungsdimension können die Itemantworten erst dann gelten, wenn jedes der Items dieselbe Ordnungsrelation zwischen je zwei Vpn v und w definiert. Da man davon ausgehen muss, dass die Antworten der Probanden einer Zufallsvariation unterliegen, ist es zweckmäßig, diese Ordnungsrelation ($v >^{o} w$) jedoch nicht auf Grundlage der Itemantworten (x_{vi}) selbst, sondern gemäß $v >^{o} w \Leftrightarrow \tau_{vi} > \tau_{wi}$ auf Grundlage ihrer Erwartungswerte $\tau_{vi} = E(X_{vi})$ zu definieren. Dass die Items ($i = 1,\ldots,k$) dieselbe Einstellungsdimension messen, lässt sich dann an der empirischen Gesetzmäßigkeit ablesen, dass die erwarteten Itemscores unabhängig von der Auswahl der Items stets dieselbe Rangordnung zwischen den Vpn ergeben. Ist dies der Fall, so ist der Fragebogen *ordinal homogen* und die Profillinien in Abbildung 1 sind überschneidungsfrei.[2] Überschneiden sie sich dagegen, so ist der Fragebogen *inhomogen* und die Rückführung der Itemantworten auf eine gemeinsame latente Dimension ist nicht möglich.

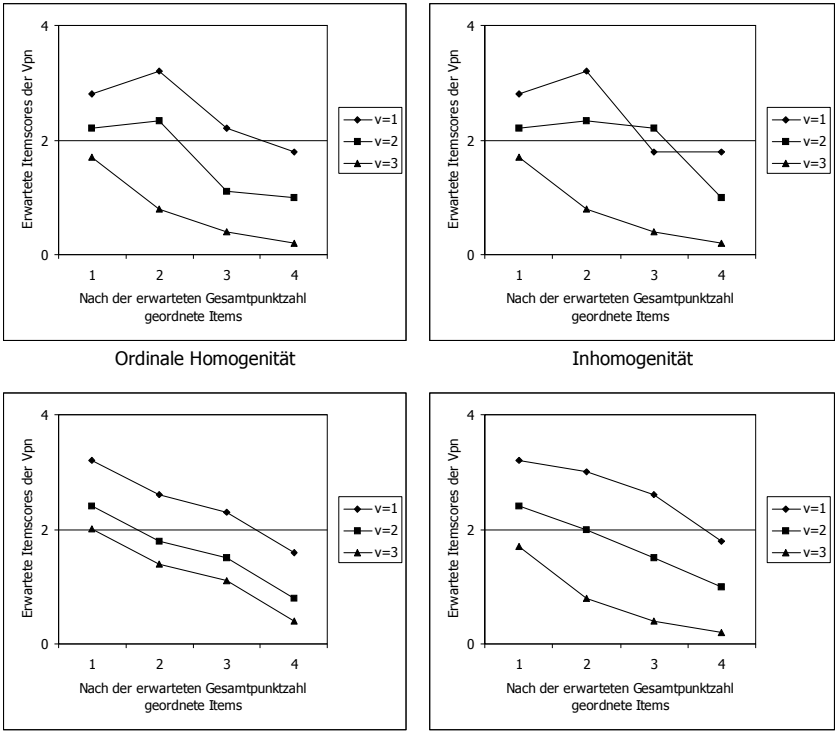


Abbildung 1: Profillinien dreier Vpn

2 Um die Profillinien zu konstruieren, trägt man die erwarteten Itemscores der Vpn (τ_{vi}) auf der Ordinate eines Koordinatensystems ab, auf dessen Abszisse die Items nach der erwarteten Gesamtzahl der bei ihnen erzielten Punkte (τ_{oi}) angeordnet sind, so dass (von links nach rechts) das Item mit der höchsten erwarteten Gesamtpunktzahl zuerst und jenes mit der niedrigsten zuletzt kommt.

Selbst wenn ein Fragebogen ordinal homogen ist, bedeutet dies aber noch nicht, dass die Rangordnung der Probanden durch ihre Summenscores korrekt wiedergegeben wird. Eine notwendige (aber nicht hinreichende) Bedingung dafür besteht darin, dass auch die Items nach ihrer Schwierigkeit geordnet sind und die Regel $i <^{o} j \Leftrightarrow \tau_{vi} > \tau_{vj}$ unabhängig von der Auswahl der Vpn stets dieselbe Schwierigkeitsordnung zweier Items i und j definiert. Graphisch dargestellt bedeutet dies, dass die in Abbildung 1 dargestellten Profillinien nicht nur überschneidungsfrei, sondern auch monoton fallend sein müssen.

Die klassische Testtheorie, auf welche man sich bei der Konstruktion von Einstellungsfragebögen zumeist beruft, wird dieser Voraussetzung gerecht, indem sie die Homogenität von Tests oder Fragebögen durch das Konzept der essenziellen τ-Äquivalenz operationalisiert, wonach sich die erwarteten Scores je zweier Items lediglich durch eine von den Vpn unabhängige Konstante $c_{ij} = \tau_{vi} - \tau_{vj}$ unterscheiden, in welcher der Schwierigkeitsunterschied der beiden Items zum Ausdruck kommt (vgl. Lord & Novick 1968, S. 50, 135). Ist dies der Fall, so verlaufen die in Abbildung 1 dargestellten Profillinien parallel zueinander.

Da die Parallelität der Profillinien für jedes beliebige Paar von Vpn gefordert ist, muss sie auch für jedes Paar von – wie auch immer definierten – Klassen von Vpn gelten. Dies eröffnet die Möglichkeit, die essenzielle τ-Äquivalenz der Items mittels einfacher t-Tests und/oder Varianzanalysen statistisch zu überprüfen. In der Praxis der Fragebogenkonstruktion verzichtet man jedoch meist auf solche Modelltests und bedient sich stattdessen einer mehr pragmatischen Vorgehensweise. Da man kaum damit rechnen kann, dass die Items auch tatsächlich essenziell τ-äquivalent sind, versucht man stattdessen zu quantifizieren, wie gut (oder schlecht) ihre essenzielle τ-Äquivalenz – und damit die interne Konsistenz des Fragebogens – gewährleistet ist.

Tatsächlich ist Cronbachs Koeffizient α, dessen man sich dazu bedient, jedoch durch die Reliabilität des Fragebogens nach oben beschränkt. Je mehr die essenzielle τ-Äquivalenz verletzt ist, desto kleiner wird er, und je besser sie erfüllt ist, desto mehr nähert er sich der Reliabilität an. Um als zuverlässiges Maß für die interne Konsistenz eines Fragebogens dienen zu können, müsste er daher zur Reliabilität des Fragebogens in Beziehung gesetzt werden. Eben dies scheitert jedoch daran, dass alle anderen Reliabilitätsmaße noch strengere Voraussetzungen treffen und die Reliabilität bei Verletzung dieser Voraussetzungen sowohl über- als auch unterschätzen können (vgl. Kempf 2008, S. 173f.).[3] Außer dem Koeffizienten α selbst, der eine untere Schranke der Reliabilität darstellt, vermag die klassische Testtheorie daher kein zuverlässiges Reliabilitätsmaß anzugeben.

Aber es ist nicht nur dieses Defizit der klassischen Testtheorie, welches das gebräuchliche Vorgehen bei der Fragebogenkonstruktion methodisch unbefriedigend erscheinen lässt. Wie seit den richtungsweisenden Arbeiten von Rasch (1960,

3 Die einzige Ausnahme hiervon bildet die Retest-Reliabilität, welche die Reliabilität eines Tests stets unterschätzt, wenn die Voraussetzungen ihrer Berechnung verletzt sind.

1961) bekannt ist, kann die Scorebildung mit einem erheblichen Verlust an statistischer Information einhergehen. Der Summenscore schöpft die in den Antwortmustern der Probanden enthaltene Information über die Ausprägung der latenten Personenvariable (= der zugrunde liegenden Einstellung) lediglich dann (und nur dann) vollständig aus, wenn die Wahrscheinlichkeitsdichte der Antwortvariablen durch die logistische Funktion des Rasch-Modells dargestellt werden kann. Nur wenn dies der Fall ist, stellt der Summenscore eine suffiziente Statistik für die Ausprägung der latenten Personenvariable dar. Die Profillinien sind dann zwar ebenfalls überschneidungsfrei und monoton fallend, sie verlaufen jedoch nicht parallel (vgl. Abb. 1), weshalb die interne Konsistenz schon von vornherein kein geeignetes Kriterium ist, welches die Scorebildung zu rechtfertigen vermag.

Namentlich in der Einstellungsmessung ist die Suffizienz des Summenscores jedoch nur äußerst selten erfüllt. Einstellungsfragebögen sind nur selten strikt homogen im Sinne des Rasch-Modells. Dennoch hält man an der Scorebildung routinemäßig fest und ist schon zufrieden, wenn der Fragebogen einen ausreichend hohen Koeffizienten α aufweist.

Auch wenn es pragmatische Gründe geben mag, die für diese Vorgehensweise sprechen, wirft sie doch die Frage auf, wie groß der Verlust an statistischer Information ist, der damit einhergeht. Oder anders ausgedrückt: Wenn man schon an der Scorebildung festhält, obwohl der Fragebogen nicht strikt homogen ist, sollte man wenigstens ein Relevanzmaß angeben, welches das Ausmaß des damit verbundenen Informationsverlustes quantifiziert. Und wenn sich der Informationsverlust als zu schwerwiegend erweist, sollte man von der Scorebildung Abstand nehmen und die latente Personenvariable nicht als quantitative, sondern als qualitative Variable konzipieren, wie dies in der auf Lazarsfeld (1950) zurückgehenden Latent-Class-Analyse der Fall ist, mittels derer man Typen von Antwortmustern identifizieren kann.

2 Ein allgemeines testtheoretisches Modell

Um ein solches Relevanzmaß zu konstruieren, gehen wir von einem allgemeinen testtheoretischen Modell aus (Kempf 2008, S. 199ff.), das sowohl die klassische Testtheorie als auch die Item-Response-Modelle mit einschließt. Obwohl die klassische und die stochastische Testtheorie verschiedenen Forschungstraditionen entstammen, gehen sie von miteinander kompatiblen Postulaten aus, die in Analogie zu Novicks (1966) Grundannahmen der klassischen Testtheorie in Form der folgenden Modellannahmen ausgedrückt werden können:

1. Jeder Vp v und jedem Item i entspricht eine diskrete Zufallsvariable X_{vi} mit dem Wertevorrat x = 0,...,m-1 und der Wahrscheinlichkeitsdichte $f_{vi}(x)$.
2. Die Items sind lokal unabhängig, so dass die Wahrscheinlichkeit des Antwortvektors einer gegebenen Vp durch die Gleichung

$$\text{prob}(x_{v1},\ldots,x_{vk}) = \prod_{i=1}^{k} f_{vi}(x_{vi}) \qquad (1)$$

dargestellt werden kann.

Liegen die Antworten von n Vpn vor, so können diese in einer Antwortmatrix $X = ((x_{vi}))$ dargestellt werden, deren Zeilen die Antwortvektoren $x_v = (x_{v1},\ldots,x_{vk})$ bilden, welche das Antwortmuster einer Vp wiedergeben und somit eine vollständige Beschreibung ihres Testergebnisses leisten. Dieses Antwortmuster stellt die Datengrundlage dar, von welcher auf die Ausprägung der den Antworten zugrundeliegenden, latenten Personenvariable geschlossen wird.

In der klassischen Testtheorie wird die Dichtefunktion $f_{vi}(x)$ nicht weiter spezifiziert. Stattdessen wird lediglich der als True-Score bezeichnete Erwartungswert der Antwortvariablen $\tau_{vi} = E(X_{vi})$ gemäß

$$x_{vi} = \tau_{vi} + f_{vi} \qquad (2)$$

modelliert, worin $f_{vi} = x_{vi} - \tau_{vi}$ den Messfehler beschreibt, mit welchem die Antwort der Vp belastet ist.

In der stochastischen Testtheorie werden dagegen Zusatzannahmen bezüglich der Wahrscheinlichkeitsdichte der Antwortvariablen getroffen und je nachdem, welche, lassen sich verschiedene Item-Response-Modelle unterscheiden, zwischen denen die in Abbildung 2 dargestellte Hierarchie besteht.

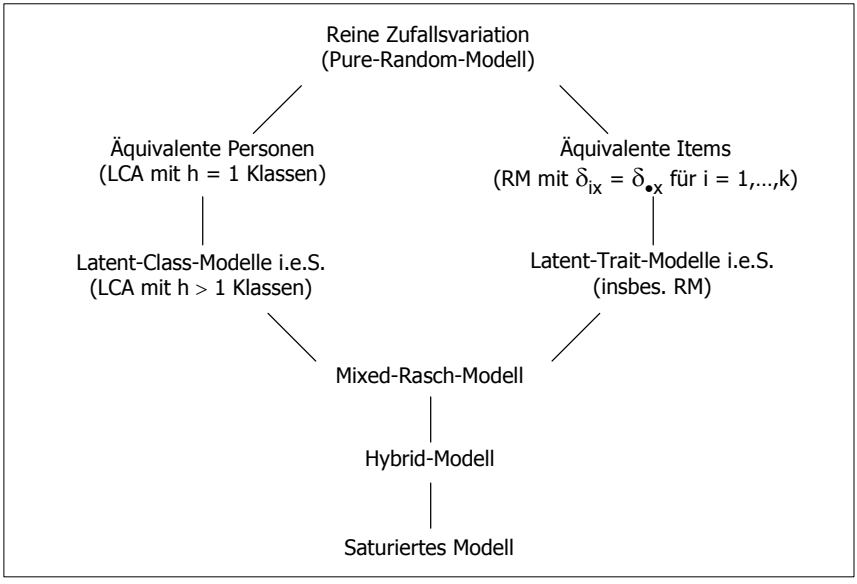


Abbildung 2: Die konzeptionelle Hierarchie der Item-Response-Modelle (nach Kempf 2008, S. 200; modifiziert)

Reine Zufallsvariation: Das Pure-Random-Modell (PR) postuliert, dass die Antwortvariablen X_{vi} sowohl personen- als auch itemunabhängig verteilt sind, so dass alle Vpn und Items dieselben Antwortvariablen mit der Wahrscheinlichkeitsdichte

$$f_{vi}(x) = f_{\bullet\bullet}(x) = p_{\bullet\bullet x} \tag{3}$$

besitzen. Ist diese Annahme erfüllt, liefert der Fragebogen weder Informationen über Einstellungsunterschiede zwischen den Vpn noch über Schwierigkeitsunterschiede zwischen den Items. In diesem Fall ist er als diagnostisches Instrument unbrauchbar. Alle Variation in den Daten ist reine Zufallsvariation und die (marginale) Likelihood der Antwortmatrix hat die Form

$$L_{PR} = \prod_{x=0}^{m-1} p_{\bullet\bullet x}^{n_{oox}} \quad \text{mit} \quad n_{oox} = \sum_{v=1}^{n}\sum_{i=1}^{k} n_{vix}\,, \tag{4}$$

worin

$$n_{vix} = \begin{cases} 1 & \text{wenn } x_{vi} = x \\ O & \text{sonst} \end{cases}. \tag{5}$$

Obwohl das PR-Modell die Antworten der Probanden (noch) nicht auf eine latente Personenvariable zurückführt und auch keine Unterschiede zwischen den Items annimmt, ist es jedoch insofern von theoretischem Interesse, als die Likelihood des PR-Modells die untere Schranke der Likelihood darstellt, welche durch Item-Response-Modelle erzielt werden kann.

Äquivalente Personen: Keinerlei diagnostische Informationen über Einstellungsunterschiede zwischen den Vpn enthält der Fragebogen auch dann, wenn sich zwar die Antwortvariablen verschiedener Items voneinander unterscheiden, diese jedoch personenunabhängig verteilt sind, so dass alle Vpn dieselben Antwortvariablen mit der Wahrscheinlichkeitsdichte

$$f_{vi}(x) = f_{\bullet i}(x) = p_{\bullet ix} \tag{6}$$

besitzen. Formal gesehen entspricht dieses Modell der 1-Klassen-Lösung der Latent-Class-Analyse (LCA).

Latente Personenvariablen: Diagnostische Informationen über die Vpn enthält der Fragebogen erst dann, wenn die Antwortvariablen personenabhängig verteilt sind. Ist dies der Fall, so kann die Wahrscheinlichkeitsdichte der Antwortvariablen als Funktion einer latenten Personenvariable Θ dargestellt werden

$$f_{vi}(x) = f_i(x, \theta_v)\,, \tag{7}$$

die entweder ein- oder mehrdimensional sein kann. In unserem Kontext interessiert zunächst nur der eindimensionale Fall, wobei wir wiederum zwischen diskreten und kontinuierlichen Personenvariablen unterscheiden können.

Latent-Trait-Modelle: Ist die latente Personenvariable kontinuierlich, so wird sie auf einer metrischen Skala mit dem Wertevorrat $-\infty \le \theta_v \le \infty$ gemessen. Die Wahrscheinlichkeit, dass eine zufällig herausgegriffene Vp v das Item i in der Kategorie x beantwortet, kann dann in der Form

$$\text{prob}(X_{vi} = x) = \int_{-\infty}^{\infty} f_i(x,\theta)\, g(\theta)\, d\theta \tag{8}$$

dargestellt werden, worin $g(\theta)$ die Wahrscheinlichkeitsdichte der latenten Variable bezeichnet.

Äquivalente Items: Im einfachsten Fall bestehen keine Unterschiede zwischen den verschiedenen Items, so dass alle dieselben Antwortvariablen mit der Wahrscheinlichkeitsdichte

$$f_i(x,\theta_v) = f_\bullet(x,\theta_v) = p_{v\bullet x} \tag{9}$$

besitzen. Formal gesehen entspricht dies dem Rasch-Modell (RM) mit $\delta_{ix} = \delta_{\bullet x}$ für i = 1,...,k und x = 0,...,m-1.

Latent-Trait-Modelle i.e.S: Grundsätzlich könnte man für $f_i(x,\theta_v)$ jede beliebige stetige Verteilungsfunktion einsetzen. Nur die im RM postulierte logistische Funktion

$$f_{vi}(x,\theta_v) = \frac{\exp(x\theta_v - \delta_{ix})}{\sum_{y=0}^{m-1}\exp(y\theta_v - \delta_{iy})} \qquad \text{mit } \delta_{i0} = 0 \tag{10}$$

gewährleistet jedoch, dass der Summenscore die im Antwortvektor enthaltene statistische Information über die latente Personenvariable vollständig ausschöpft. Darin bezeichnet θ_v die Ausprägung der latenten Personenvariable und δ_{ix} die Schwierigkeit, welche der Beantwortung von Item i in Kategorie x entgegensteht.

Ist das RM erfüllt, so teilt der Summenscore die Personenpopulation in h = k*(m-1)+1 auf einer Ordinalskala gemessene (manifeste) Klassen x_{vo} = 0,...,k*(m-1) ein, so dass die marginale Likelihood der Antwortmatrix

$$L_{RM} = \prod_{v=1}^{n} \text{prob}\left\{(x_{v1},...,x_{vk}) \middle| x_{vo}\right\} \prod_{g=0}^{h} p_g^{n_g} \tag{11}$$

in zwei Faktoren aufgespaltet werden kann, deren erster – die bedingte Likelihood – von der latenten Personenvariable unabhängig ist. Darin bezeichnet n_g die Anzahl der Vpn, deren Score gleich g ist, und p_g die Wahrscheinlichkeit, mit der eine zufällig herausgegriffene Vp diesen Score zeigt.

Latent-Class-Modelle i.e.S.: Ist die latente Personenvariable diskret, so wird sie auf einer Nominal- oder Ordinalskala mit dem Wertevorrat $\theta_g = \theta_1,...,\theta_h$ gemessen, welche die Personenpopulation in h (latente) Klassen g = 1,...,h einteilt, so dass $\theta_v = \theta_g \Leftrightarrow v \in g$ und alle Vpn, die derselben Klasse angehören, dieselben klassenspezifischen Kategorienwahrscheinlichkeiten

$$f_i(x, \theta_g) = p_{gix} \tag{12}$$

besitzen. Die marginale Likelihood der Antwortmatrix nimmt dann die Form

$$L_{LCA} = \prod_{v=1}^{n} \left(\sum_{g=1}^{h} p_g \prod_{i=1}^{k} \prod_{x=0}^{m-1} p_{gix}^{n_{vix}} \right) \tag{13}$$

an, worin die Klassengröße p_g wieder die Wahrscheinlichkeit darstellt, mit der eine zufällig herausgegriffene Vp der Klasse g angehört.

Eine Kombination von quantitativen und qualitativen latenten Personenvariablen ist das auf Rost (1990), Mislevy & Verhelst (1990) und Keldermann & Macready (1990) zurückgehende Mixed-Rasch-Modell (MRM), das die Vp in h latente Klassen einteilt, innerhalb derer das Rasch-Modell gilt.

Ein typisches Anwendungsfeld des MRM ist die Leistungsmessung, und zwar dann, wenn die Testaufgaben mittels unterschiedlicher Lösungsstrategien bearbeitet werden können. In Abhängigkeit davon, mit welcher Strategie die Vpn die Aufgabe zu lösen versuchen, sind unterschiedliche Fähigkeiten gefordert und Aufgaben, die mit der einen Strategie relativ leicht zu lösen sind, können sich bei Anwendung einer anderen als schwierig erweisen und umgekehrt. Das MRM erlaubt dann einen zweistufigen diagnostischen Prozess: Auf der ersten Stufe kann diagnostiziert werden, welcher latenten Klasse die Vpn angehören (d. h. welcher Lösungsstrategie sie sich bedienen), und auf der zweiten Stufe können die Mitglieder derselben Klasse hinsichtlich ihrer Leistungsmenge (d. h. hinsichtlich der Frage, wie gut sie diese Strategie beherrschen) untereinander verglichen werden. In der Einstellungsmessung sind die Ergebnisse des MRM dagegen oft nur schwer oder kaum interpretierbar.

Dasselbe gilt für die Hybrid-Modelle (HM) (z. B. v. Davier & Rost 1997), welche die Vpn ebenfalls in $h_1 \geq 1$ latente Klassen einteilen, innerhalb derer das RM gilt, darüber hinaus aber auch noch $h_2 \geq 1$ Restklassen von Vpn zulassen, deren Antwortmuster nicht damit verträglich sind. Auf unser Beispiel bezogen, könnte es sich dabei um solche Vpn handeln, welche die Testaufgaben nicht mit einer einheitlichen Lösungsstrategie bearbeiten, sondern zwischen verschiedenen Strategien hin und her schwanken.

Saturiertes Modell: Definiert jedes mögliche Antwortmuster eine eigene Klasse von Vpn mit $\theta_g \Leftrightarrow (x_{g1},\ldots,x_{gk})$, so ist

$$f_i(x, \theta_g) = \begin{cases} 1 & \text{für } x = x_{gi} \\ 0 & \text{sonst} \end{cases}. \tag{14}$$

Da die Antwortmuster der Vpn direkt beobachtet werden können, ist die Personenvariable Θ in diesem Fall nicht mehr latent, sondern manifest, und die marginale Likelihood der Antwortmatrix

$$L_{sat} = \prod_{g=1}^{h_{max}} p_g^{n_g} \quad \text{mit} \quad h_{max} = m^k \tag{15}$$

hängt nur noch von den Klassengrößen p_g ab. Da das saturierte Modell keinerlei einschränkende Modellannahmen trifft, liefert es (absolut gesehen) die bestmögliche Beschreibung der Daten und die marginale Likelihood in Gleichung 15 kann durch keines der anderen Modelle übertroffen werden. Angesichts von m^k möglichen Antwortmustern ist das saturierte Modell jedoch nur in Ausnahmefällen auch interpretierbar. Es ist aber von theoretischem Interesse, weil die Likelihood des saturierten Modells die obere Schranke der Likelihood darstellt, welche durch Item-Response-Modelle erzielt werden kann.

3 Informationstheoretische Relevanzmaße

Ein wesentlicher Unterschied zwischen RM und LCA besteht darin, dass die LCA keinerlei Homogenitätsvoraussetzungen trifft und der Entscheidung über die Ausprägung der latenten Personenvariable nicht nur den Summenscore, sondern das komplette Antwortmuster der Probanden zugrunde legt. Auch darüber, ob sich die Klassen quantitativ oder qualitativ voneinander unterscheiden, wird keinerlei Vorentscheidung getroffen. Aus diesen Gründen ist die LCA auch dann zur Beschreibung der Daten geeignet, wenn das RM gilt: Bei idealer Modellgeltung des RM und entsprechender Klassenanzahl der LCA sind die von der LCA identifizierten latenten

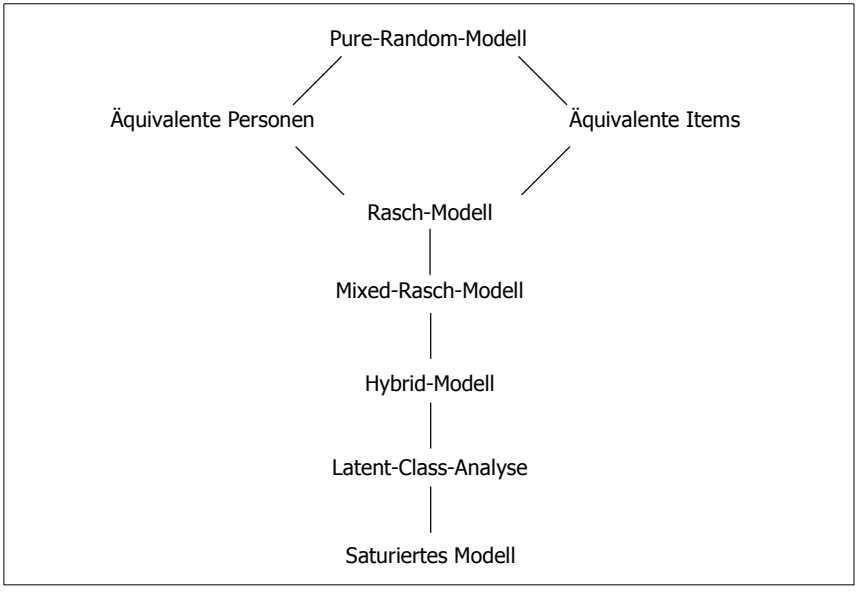


Abbildung 3: Die formale Hierarchie der Item-Response-Modelle

Klassen mit den aufgrund des Summenscores der Vpn gebildeten manifesten Klassen des RM identisch. Ähnliches gilt – wenngleich auch auf komplexerem Niveau – auch für das MRM und das HM.

Die in Abbildung 2 dargestellte Hierarchie der Modelle besteht somit nur konzeptionell. Formal betrachtet, stellen alle anderen Modelle einen Spezialfall der LCA dar, wobei die latenten Klassen teilweise mit manifesten zusammenfallen. Je höher ein Modell in der formalen Hierarchie (vgl. Abb. 3) steht, desto restriktiver sind seine Modellannahmen.

Als Relevanzmaß für den Verlust an statistischer Information, welcher mit der Scorebildung einhergeht, ist es daher zweckmäßig, die Erklärungsmacht (explanatory power) des RM zu jener der LCA in Beziehung zu setzen.

3.1 Der Explanatory-Power-Index

Um ein solches Maß zu konstruieren, gehen wir in mehreren Schritten vor und berechnen zunächst die in bit pro Antwortmuster gemessene (absolute) Redundanz der Daten

$$C_{abs,MX} = H_{max} - H_{MX}\ , \tag{16}$$

die durch ein gegebenes Modell (MX) identifiziert wird. Darin bezeichnet H_{MX} die Entropie (= den mittleren Informationsgehalt) eines Antwortmusters bei Geltung des Modells MX, die gemäß

$$H_{MX} = -\frac{1}{n\ \ln(2)}\ln(L_{MX}) \tag{17}$$

als Funktion der Likelihood (L_{MX}) dargestellt werden kann, welche die Antwortmatrix bei Geltung des Modells MX besitzt (vgl. Anhang 1, S. 51).

H_{max} bezeichnet die maximal mögliche Entropie eines Antwortmusters, die sich gemäß

$$H_{max} = k\ \frac{\ln(m)}{\ln(2)} \tag{18}$$

berechnen lässt (vgl. Anhang 2, S. 52) und von der Annahme ausgeht, dass alle Antwortkategorien dieselbe personen- und itemunabhängige a-priori-Wahrscheinlichkeit $p_{\bullet\bullet x} = m^{-1}$ besitzen, so dass auch alle m^k möglichen Antwortmuster gleich wahrscheinlich sind.

Einsetzen von (17) und (18) in (16) ergibt für die Redundanz der Daten, die durch das Modell identifiziert wird, schließlich die Formel

$$C_{abs,MX} = \frac{n\ k\ \ln(m) + \ln(L_{MX})}{n\ \ln(2)} \tag{19}$$

(vgl. Anhang 3, S. 52). Je größer der Wert ist, den dieser Ausdruck annimmt, desto mehr Information wird von dem Modell ausgeschöpft.

Da schon die Annahme empirischer Kategorienwahrscheinlichkeiten ($p_{\bullet\bullet x}$), die von der a-priori-Wahrscheinlichkeit (m^{-1}) abweichen können, ein gewisses Maß an statistischer Information ausschöpft, ist es sinnvoll, die Differenz

$$VI_{MX} = C_{abs,MX} - C_{abs,PR} \tag{20}$$
$$= \frac{\ln(L_{MX}) - \ln(L_{PR})}{n \ln(2)}$$

zu berechnen, welche die modellspezifische (= über die Annahme empirischer Kategorienwahrscheinlichkeiten hinausgehende) statistische Information misst, die das jeweilige Modell (MX = RM oder LCA) verwertet. Wegen $L_{PR} \leq L_{MX} \leq L_{sat}$ kann der Ausdruck in Gleichung 20 aber höchstens den Wert

$$VI_{max} = C_{abs,sat} - C_{abs,PR} \tag{21}$$
$$= \frac{\ln(L_{sat}) - \ln(L_{PR})}{n \ln(2)}$$

annehmen, welcher die statistische Information angibt, die durch Annahme einer latenten Personenvariable *maximal* ausgeschöpft werden kann.

Indem man die Ausdrücke in Gleichung (20) und (21) zueinander in Beziehung setzt, erhält man ein Maß für die (relative) Erklärungsmacht des Modells MX, welches den Anteil der durch Annahme einer latenten Personenvariable verwertbaren Information angibt, der durch das jeweilige Modell ausgeschöpft wird:

$$EP_{MX} = \frac{VI_{MX}}{VI_{max}} \tag{22}$$
$$= \frac{\ln(L_{MX}) - \ln(L_{PR})}{\ln(L_{sat}) - \ln(L_{PR})}.$$

3.2 Der Loss-of-Information-Index

Der so definierte EP-Index ist analog zu dem in der soziologischen Statistik bekannten PRE-Maß (proportional reduction in error) konstruiert, das von Goodman (1972) eingeführt wurde und auf eine Arbeit von Goodman & Kruskal (1954) zurückgeht. Reynolds (1977) weist darauf hin, dass das PRE-Maß keine Auskunft darüber gibt, ob das jeweilige Modell die Daten adäquat erklärt, sondern lediglich die Passung zweier Modelle (hier des PR- und des MX-Modells) miteinander vergleicht. In der Tat kann man leicht Fälle zeigen, in denen PRE bzw. EP zwar hoch sind, der Likelihood-Quotienten-Test gegen das saturierte Modell aber dennoch signifikant ist. Gerade dies qualifiziert das Maß jedoch dafür, als Grundlage für die

Quantifizierung des Informationsverlustes zu dienen, der bei mangelnder Modellanpassung des RM mit der Scorebildung einhergeht. Als Maß hierfür sei daher der Loss-of-Information-Index

$$LI_{RM,LCA} = 1 - \frac{EP_{RM}}{EP_{LCA}} \qquad (23)$$

$$= \frac{\ln(L_{LCA}) - \ln(L_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

(vgl. Anhang 4, S. 53) vorgeschlagen, der den Anteil an statistischer Information beschreibt, der bei Unterstellung strikter Homogenität (RM) gegenüber der LCA (keine Homogenitätsannahme) ungenutzt bleibt.

Datenbeispiel 1: Als Beispiel betrachten wir einen Fragebogen zur Messung von manifestem Antisemitismus mit den drei Items

1. Mit Juden sollte man keine Geschäfte machen.
2. Es ist besser, mit Juden nichts zu tun zu haben.
3. Ich gehöre zu denen, die keine Juden mögen.

Diese wurden von n = 411 Vpn auf einer 5-stufigen Likert-Skala beantwortet, die von voller Ablehnung (x = 0) bis zu voller Zustimmung (x = 4) reicht. Da eine der Vpn die Items nur unvollständig beantwortet hatte, liegt den folgenden Berechnungen ein Stichprobenumfang von n = 410 zugrunde.

Die interne Konsistenz der Skala beträgt α = 0,836. Mittels LCA wurden gemäß Akaikes (1987) Informationskriterium (AIC) vier latente Klassen von Vpn identifiziert (vgl. Tab. 1). Der Likelihood-Ratio-Test des 4-Klassen-Modells gegen das saturierte Modell spricht für eine sehr gute Modellanpassung (L-Ratio = 33.98; df = 73; n. s.) und die Erklärungsmacht des Modells ist mehr als zufriedenstellend (EP = 94.31%).

Modell	n(P)	ln(L)	L-Ratio	df	p	AIC	EP
PR	4	-1249.82	597.64	120	< 0.001	2507.64	0.00%
LC1	12	-1221.16	540.32	112	< 0.001	2466.32	9.59%
LC2	25	-1026.76	151.52	99	< 0.001	2103.52	74.65%
LC3	38	-986.76	71.52	86	n.s.	2049.52	88.03%
LC4	51	-967.99	33.98	73	n.s.	2037.98	94.31%
LC5	64	-964.30	26.60	60	n.s.	2056.60	95.55%
LC6	77	-958.99	15.98	47	n.s.	2071.98	97.33%
RM	23	-1028.36	154.72	101	< 0.001	2090.72	74.11%
Saturiert	124	-951.00				2150.00	100.00%

Tabelle 1: Datenbeispiel 1, Goodness-of-Fit-Statistiken

Die Profillinien der latenten Klassen sind überschneidungsfrei und (cum crano salis) monoton fallend (vgl. Abb. 4). Dennoch muss das RM verworfen werden (L-Ratio = 154.72; df = 101; $p < 0.001$). Die Skala ist *nicht* strikt, sondern nur ordinal homogen. Die Erklärungsmacht des RM ist eher unbefriedigend (EP = 74.11%) und die Summenscorebildung geht mit einem erheblichen Verlust an statistischer Information einher (LI = 21.42%).

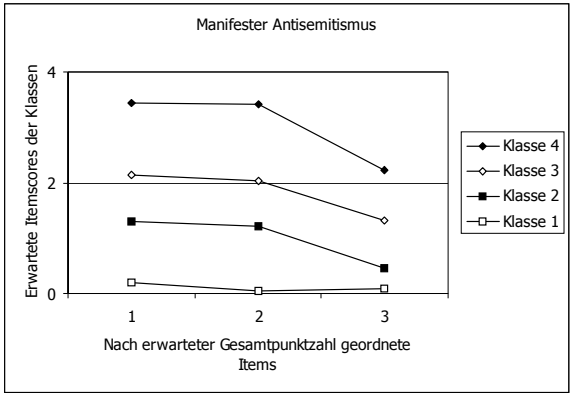


Abbildung 4: Datenbeispiel 1, Profillinien der latenten Klassen

3.3 Der korrigierte Loss-of-Information-Index

Bei Fragebogenerhebungen hat man es häufig mit Missing Data zu tun. Um Summenscores bilden zu können, bleibt einem dann nichts anderes übrig, als entweder – wie in Datenbeispiel 1 – jene Vpn, die den Fragebogen unvollständig beantwortet haben, aus der Auswertung auszuschließen oder die fehlenden Antworten in die neutrale Kategorie ("weder-noch") umzukodieren.

Datenbeispiel 2: Als Beispiel für derart umkodierte Daten betrachten wir einen Fragebogen zur Messung von sekundärem Antisemitismus mit den drei Items

1. Jahrzehnte nach Kriegsende sollten wir nicht mehr so viel über die Judenverfolgung reden, sondern endlich einen Schlussstrich unter die Vergangenheit ziehen.
2. Man sollte endlich mit dem Gerede über unsere Schuld gegenüber den Juden Schluss machen.
3. Das deutsche Volk hat (k)eine besondere Verantwortung gegenüber den Juden.

Diese Items wurden von n = 411 Vpn auf derselben 5-stufigen Likert-Skala beantwortet wie die Items in Datenbeispiel 1.

Die interne Konsistenz der Skala beträgt $\alpha = 0.772$. Um zu kennzeichnen, dass die Berechnungen auf umkodierten Daten beruhen, sind die Modelle in Tabelle 2 mit einem Stern (*) gekennzeichnet. Wie die Tabelle zeigt, identifizierte die LCA gemäß

AIC fünf latente Klassen. Der Likelihood-Ratio-Test des 5-Klassen-Modells gegen das saturierte Modell spricht für eine gute Modellanpassung (L-Ratio = 52.99; df = 60; n. s.) und die Erklärungsmacht des Modells ist zufriedenstellend (EP = 91.77%).

Modell	n(P)	ln(L)	L-Ratio	df	p	AIC	EP
PR*	4	-1933.93	643.52	120	< 0.001	3875.86	0.00%
LC1*	12	-1899.37	574.41	112	< 0.001	3822.74	10.74%
LC2*	25	-1742.37	260.41	99	< 0.001	3534.74	59.54%
LC3*	38	-1693.22	162.12	86	< 0.001	3462.45	74.81%
LC4*	51	-1668.18	112.03	73	< 0.005	3438.36	82.59%
LC5*	64	-1638.66	52.99	60	n.s.	3405.32	91.77%
LC6*	77	-1636.28	48.22	47	n.s.	3426.55	92.51%
RM*	23	-1728.14	231.95	101	< 0.001	3502.28	63.96%
Saturiert*	124	-1612.17				3472.34	100.00%

Tabelle 2: Datenbeispiel 2, Goodness-of-Fit-Statistiken

Das RM musste dagegen verworfen werden (L-Ratio = 231.95; df = 101, p < 0.001), seine Erklärungsmacht ist unbefriedigend (EP = 63.96%) und die Summenscorebildung geht mit einem dramatischen Verlust an diagnostisch relevanter statistischer Information einher (LI = 30.30%):

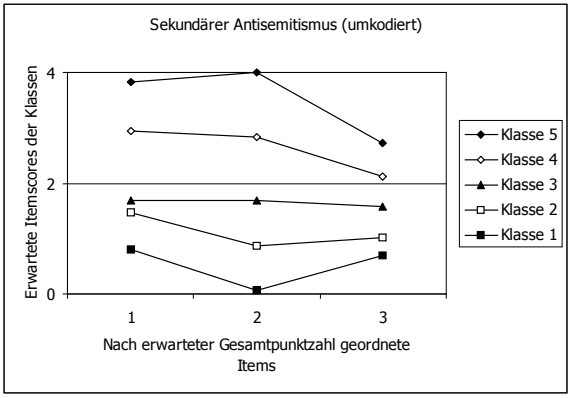


Abbildung 5: Datenbeispiel 2, Profillinien der latenten Klassen

Wie die in Abbildung 5 dargestellten Profillinien der latenten Klassen zeigen, ist die Skala zwar ordinal homogen, so dass jedes der drei Items dieselbe Rangordnung zwischen je zwei Klassen definiert (überschneidungsfreie Profillinien). Die Profillinien verlaufen jedoch nicht monoton, was darauf hinweist, dass sich die Schwierigkeitsrelation zwischen den Items mit zunehmender Ausprägung antise-

mitischer Einstellungen verschiebt. Während Vpn, die den sekundär-antisemitischen Aussagen (eher) ablehnend gegenüberstehen (Klasse 1 und 2), die Forderung, man solle mit "dem Gerede" über unsere Schuld gegenüber den Juden endlich Schluss machen (Item 2), deutlich stärker zurückweisen als die schlichte Schlussstrichforderung, die in Item 1 zum Ausdruck kommt, ist es bei ausgewachsenen Antisemiten (Klasse 5) genau umgekehrt. Sie stimmen gerade der Aussage in Item 2 stärker zu, welche nicht nur einen Schlussstrich unter die Vergangenheit fordert, sondern die Schuldfrage als bloßes "Gerede" abwehrt und sich damit in die Nähe der Holocaust-Leugnung begibt. Diese für die Diagnose antisemitischer Einstellungen höchst relevante Information geht jedoch verloren, wenn man lediglich die Summenscores der Vpn betrachtet.

Sowohl durch den Ausschluss unvollständig ausgefüllter Fragebögen (vgl. Datenbeispiel 1) als auch durch die Umkodierung von Missing Data (vgl. Datenbeispiel 2) geht Information verloren. Vpn, die ein Item unbeantwortet lassen, stimmen der entsprechenden Aussage zwar *weder* zu *noch* lehnen sie diese ab, sie beantworten das Item aber eben *nicht* in der Weder-noch-Kategorie, sondern kreuzen gar keine Antwort an.

Da dieses Antwortverhalten durchaus seine eigene diagnostische Bedeutung haben kann, ist es sinnvoll, die Umkodierung nur für die Rasch-Analyse vorzunehmen und die LCA mit Missing Data als zusätzlicher Antwortkategorie (x = m) zu berechnen.

Da die Anzahl der der Rasch-Analyse zugrunde liegenden Antwortkategorien (= m) in diesem Fall jedoch kleiner ist als jene, die der LCA zugrunde liegt (= m+1), nimmt sowohl die Likelihood des Pure-Random-Modells als auch die des saturierten Modells unter dem reduzierten Datenformat andere Werte $L^*_{PR} \neq L_{PR}$ bzw. $L^*_{sat} \neq L_{sat}$ an als in der LCA. Um zu kennzeichnen, dass auch die Likelihood und die Explanatory Power des RM unter dem reduzierten Antwortformat berechnet werden, kennzeichnen wir sie ebenfalls mit einem Stern und schreiben dafür L^*_{RM} bzw. EP^*_{RM}. Desgleichen bezeichnen wir die durch Annahme einer latenten Personenvariable im reduzierten Datenformat maximal ausschöpfbare Information mit VI^*_{max}.

Um die beiden Modelle (und Datenformate) miteinander vergleichen zu können, kann man sich dann eines korrigierten Loss-of-Information-Index

$$LI^*_{RM,LCA} = 1 - \frac{EP^*_{RM}}{EP_{LCA}} \frac{VI^*_{max}}{VI_{max}} \qquad (24)$$

$$= \frac{\ln(L_{LCA}) - \ln(L^*_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})} - \frac{\ln(L_{PR}) - \ln(L^*_{PR})}{\ln(L_{LCA}) - L(L_{PR})}$$

bedienen (vgl. Anhang 5, S. 53).

Werden LCA und RM auf Grundlage desselben Antwortformats berechnet, ist $\ln(L^*_{PR}) = \ln(L_{PR})$ und das Korrekturglied $\frac{\ln(L_{PR}) - \ln(L^*_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})} = 0$ entfällt, so dass sich $LI^*_{RM,LCA}$ auf $LI_{RM,LCA}$ reduziert.

Datenbeispiel 3: Als Beispiel betrachten wir wieder den Fragebogen zur Messung von sekundärem Antisemitismus und berechnen die LCA mit Antwortverweigerung als einer eigenständigen Antwortkategorie (x = 5), die Rasch-Analyse dagegen wie oben mit umkodierten Daten.

Wie die Goodness-of-Fit-Statistiken in Tabelle 3 zeigen, identifizierte die LCA auch hier wieder fünf latente Klassen (gemäß AIC), deren Profillinien sich ebenfalls kaum von jenen in Datenbeispiel 2 unterscheiden (vgl. Abb. 5). Der Likelihood-Ratio-Test gegen das saturierte Modell demonstriert erneut eine sehr gute Modellanpassung (L-Ratio = 59.44; df = 136; n. s.) und die Erklärungsmacht der LCA ist zwar nur geringfügig, aber doch höher als in Datenbeispiel 2 (EP = 90.91%). Auch der mit Umkodierung und Scorebildung einhergehende Verlust an statistischer Information (LI* = 30.74%) ist nur geringfügig größer als er in Datenbeispiel 2 zu sein schien.

Dies ist aber kein Argument, welches dagegen spricht, Missing Data in der LCA als eigenständige Antwortkategorie zu berücksichtigen und den korrigierten Loss-of-Information-Index anzuwenden. Die Geringfügigkeit der Unterschiede zwischen den Datenbeispielen 2 und 3 ist allein auf die Tatsache zurückzuführen, dass in den vorliegenden Daten nur sehr wenige Missing Data enthalten waren.

Modell	n(P)	ln(L)	L-Ratio	df	p	AIC	EP
PR	5	-1940.41	653.72	210	< 0.001	3890.82	0.00%
LC1	15	-1904.49	581.88	200	< 0.001	3838.98	10.99%
LC2	31	-1746.76	266.42	184	< 0.001	3555.52	59.25%
LC3	47	-1697.21	167.32	168	n.s.	3488.42	74.40%
LC4	63	-1671.45	115.80	152	n.s.	3468.90	82.29%
LC5	79	-1643.27	59.44	136	n.s.	3444.54	90.91%
LC6	95	-1636.77	46.44	120	n.s.	3463.54	92.90%
Saturiert	215	-1613.55				3657.10	100.00%
PR*	4	-1933.93	643.52	120	< 0.001	3875.86	0.00%
RM*	23	-1728.14	231.94	101	< 0.001	3502.28	63.96%
Saturiert*	124	-1612.17				3472.34	100.00%

Tabelle 3: Datenbeispiel 3, Goodness–of-Fit-Statistiken

Um die Profillinien der latenten Klassen darzustellen, kann man in diesem Falle zwei Wege gehen. Der eine besteht darin, die erwarteten Itemscores denn doch

wieder mit umkodierten Daten (keine Antwort = weder-noch) zu berechnen. Dies ist allerdings wenig befriedigend. Besser ist es, die erwarteten Itemscores nur aufgrund der Daten jener Vpn zu berechnen, welche die Antwort auf das jeweilige Item *nicht* verweigert haben. Dies kann man ganz einfach dadurch bewerkstelligen, dass man die klassenspezifischen Kategorienwahrscheinlichkeiten p_{gix} gemäß

$$p^{\#}_{gix} = \frac{p_{gix}}{\sum_{y=0}^{m-1} p_{giy}} \quad \text{für } x = 0,\ldots,m\text{-}1 \tag{25}$$

auf 100% hochrechnet und die erwarteten Itemscores nach

$$\tau_{gi} = \sum_{x=0}^{m-1} x \; p^{\#}_{gix} \tag{26}$$

berechnet. In unserem Datenbeispiel ergeben die beiden Berechnungsarten allerdings kaum einen Unterschied (vgl. Abb. 6), was unter anderem wieder auf die geringe Häufigkeit von Missing Data zurückzuführen ist.

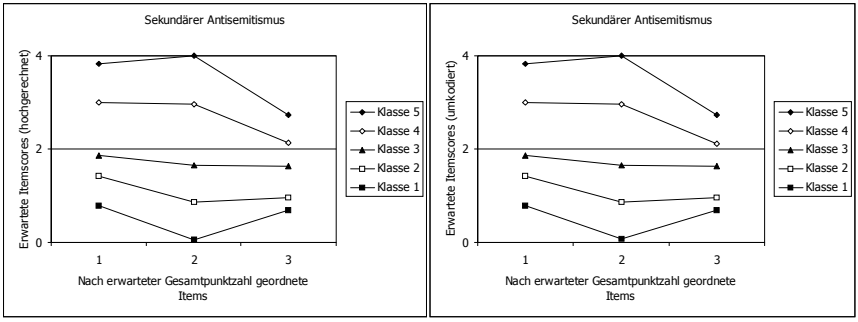


Abbildung 6: Datenbeispiel 3, Profillinien der latenten Klassen

4 Modellauswahl

Bei der Bestimmung der Anzahl latenter Klassen, die zur Beschreibung der Daten angemessenen ist, geht man in der LCA, im MRM und/oder im HM recht pragmatisch vor. Da die Likelihood der Antwortmatrix mit zunehmender Klassenzahl monoton wächst, kann man nicht einfach die Klassenzahl wählen, unter welcher die Antwortmatrix die größte Likelihood besitzt, denn dies würde stets zu einer Entscheidung zugunsten des saturierten Modells führen.

Je mehr latente Klassen man annimmt, desto mehr Modellparameter sind aus den Daten zu schätzen; je mehr Modellparameter zu schätzen sind, desto weniger wird die Verteilung der Antwortmatrix restringiert; und je weniger die Verteilung der Antwortmatrix restringiert wird, desto größer wird ihre Likelihood.

Als Kriterium für die Güte der Beschreibung wählt man daher ein Maß wie den AIC-Index

$$AIC = 2\left|\ln(L_{MX})\right| + 2\ n(P_{MX}), \tag{27}$$

der die Likelihood der Datenmatrix und die Anzahl der zu schätzenden Modellparameter zueinander in Beziehung setzt. Darin bezeichnet L_{MX} die Likelihood und $n(P_{MX})$ die Anzahl der unabhängigen Modellparameter des jeweiligen Modells.

Da die Likelihood eine Zahl zwischen Null und Eins ist, ist ihr natürlicher Logarithmus dem Betrage nach umso kleiner, je größer die Likelihood ist. Der AIC-Index ist daher umso kleiner, je genauer die Daten beschrieben werden (große Likelihood) und je sparsamer die Beschreibung ist (geringe Parameterzahl).

Als bestes Modell wählt man daher jene Klassenanzahl aus, welche den AIC-Index minimiert und somit einen optimalen Kompromiss zwischen Genauigkeit und Sparsamkeit der Beschreibung darstellt.

In großen Stichproben führt der AIC-Index jedoch leicht zu einer Inflation der Klassenzahl, weshalb man sich dann anderer Informationsmaße bedient, die, wie der auf Bozdogan (1987) zurückgehende BIC-Index

$$BIC = 2\left|\ln(L_{MX})\right| + \ln(n)\ n(P_{MX}), \tag{28}$$

die Anzahl der Modellparameter stärker gewichten und somit eine geringere Klassenzahl favorisieren. Als Beispiel hierfür stellen wir in Tabelle 4 die AIC- und BIC-Indizes der oben diskutierten Datenbeispiele einander gegenüber. Wie die Tabelle zeigt, favorisiert der BIC in allen drei Fällen tatsächlich eine geringere Klassenzahl.

Modell	Datenbeispiel 1			Datenbeispiel 2			Datenbeispiel 3		
	AIC	BIC	EP	AIC	BIC	EP	AIC	BIC	EP
LC1	2466.32	2514.51	9.59%	3822.74	3870.96	10.74%	3838.98	3899.26	10.99%
LC2	2103.52	2203.92	74.65%	3534.74	3635.20	59.54%	3555.52	3680.10	59.25%
LC3	2049.52	2202.13	88.03%	3462.45	3615.15	74.81%	3488.42	3677.29	74.40%
LC4	2037.98	2242.80	94.31%	3438.36	3643.31	82.59%	3468.90	3722.07	82.29%
LC5	2056.60	2313.63	95.55%	3405.32	3662.51	91.77%	3444.54	3762.01	90.91%
LC6	2071.98	2381.22	97.33%	3426.55	3735.99	92.51%	3463.54	3845.31	92.90%
$LI_{BIC,AIC}$	6.66%			18.48%			18.15%		

Tabelle 4: Gegenüberstellung der AIC- und BIC-Indizes

Wann eine Stichprobe groß ist, bleibt aber ziemlich vage. Sind n = 410 (Datenbeispiel 1) oder n = 411 (Datenbeispiele 2 und 3) schon große Stichproben? Ist der BIC bereits bei dieser Größe gegenüber dem AIC zu bevorzugen? Führt die Wahl

der vom BIC favorisierten geringeren Klassenzahl (Modell MY) nicht zu einem relevanten Informationsverlust gegenüber jener, die vom AIC favorisiert wird (Modell MX)?

Für die Entscheidung, auf welches der beiden Informationsmaße man sich im konkreten Fall stützen soll, ist es jedenfalls hilfreich, wieder einen Loss-of-Information-Index

$$LI_{MY,MX} = 1 - \frac{EP_{MY}}{EP_{MX}} \qquad (29)$$

$$= \frac{\ln(L_{MX}) - \ln(L_{MY})}{\ln(L_{MX}) - \ln(L_{PR})}$$

zu berechnen, der die Erklärungsmacht der beiden Modelle einander gegenüberstellt.

Für unsere Datenbeispiele (vgl. Tab. 4) zeigt sich, dass die Wahl des BIC in zwei Fällen mit $LI_{AIC,BIC}$ = 18,48% (Datenbeispiel 2) bzw. $LI_{AIC,BIC}$ = 18,15% (Datenbeispiel 3) mit einem recht erheblichen Informationsverlust verbunden wäre. Wollte man sich für die vom BIC vorgeschlagene 3-Klassen-Lösung entscheiden, so würde auch der Informationsvorteil der LCA gegenüber dem RM von $LI_{RM,AIC}$ = 30,30% auf $LI_{RM,BIC}$ = 14,50% (Datenbeispiel 2) bzw. von $LI^*_{RM,AIC} = 30{,}74\%$ auf $LI^*_{RM,BIC} = 15{,}38\%$ (Datenbeispiel 3) absinken. In beiden Datenbeispielen empfiehlt es sich daher, an der vom AIC vorgeschlagenen 5-Klassen-Lösung festzuhalten.

In Datenbeispiel 1 dagegen erscheint der Informationsverlust, der mit einer Entscheidung zugunsten des BIC einhergeht, mit $LI_{AIC,BIC}$ = 6,66% nicht wirklich relevant (LI < 10%) und deutlich geringer als jener, den die Verwendung der Summenscores nach sich zieht ($LI_{RM,LCA}$ = 21,42%). Wenn die vom BIC vorgeschlagene geringere Klassenzahl besser interpretierbar ist als die vom AIC vorgeschlagene, so mag man sich in Fällen wie diesem zugunsten des BIC entscheiden. Für die Quantifizierung des Informationsverlustes, der mit der Summenscorebildung einhergeht, sollte man aber unbedingt von der Klassenzahl ausgehen, welche vom AIC vorgeschlagen wird. Andernfalls würde der Informationsverlust unterschätzt (in unserem Beispiel mit $LI_{RM,BIC}$ = 16,55% statt $LI_{RM,AIC}$ = 21,42%).

5 Zusammenfassung

Die in der Einstellungsmessung gebräuchliche Verwendung von Summenscores ist an Voraussetzungen gebunden, die in der Praxis nur selten überprüft werden. Sind diese Voraussetzungen verletzt, so geht der Summenscore mit einem mehr oder minder gravierenden Verlust an diagnostisch relevanter statistischer Information bezüglich der latenten Personenvariable einher.

Tatsächlich wird diese Information durch den Summenscore nur dann voll ausgeschöpft, wenn der Fragebogen strikt homogen im Sinne des RM ist. Ist diese Voraussetzung nicht erfüllt, sollte man wenigstens ein Relevanzmaß angeben können, welches den Informationsverlust quantifiziert, der mit der Verwendung des Summenscores einhergeht.

Erweist sich der Informationsverlust als zu schwerwiegend, sollte man von der Scorebildung Abstand nehmen und die latente Personenvariable nicht als eine quantitative Einstellungsdimension, sondern als qualitative Variable konzipieren.

Zu diesem Zweck wurde im vorliegenden Aufsatz ein informationstheoretischer Loss-of-Information-Index (Gleichungen 23) entwickelt, der die Erklärungsmacht des RM mit jener der LCA konfrontiert, die im Unterschied zum RM keinerlei Homogenitätsvoraussetzungen trifft.

Während die LCA dazu in der Lage ist, Missing Data als eigenständige Antwortkategorie zu modellieren, ist dies beim RM nicht der Fall. Um Summenscores bilden zu können, muss man dann entweder jene Vpn, die den Fragebogen unvollständig ausgefüllt haben, aus der Datenauswertung ausschließen oder die fehlenden Antworten in die neutrale Kategorie ("weder-noch") umkodieren. Sowohl durch den Ausschluss unvollständig ausgefüllter Fragebögen als auch durch die Umkodierung von Missing Data geht aber ebenfalls Information verloren. Um den Informationsverlust zu quantifizieren, der mit Umkodierung *und* Summenscorebildung einhergeht, wurde daher ein korrigierter Loss-of-Informaton-Index (Gleichung 24) entwickelt, der die Missing Data in der LCA als eigenständige Antwortkategorie berücksichtig und nur das RM auf Grundlage von umkodierten Daten berechnet.

Stützt man sich bei der Interpretation der Ergebnisse der LCA, des MRM oder des HM nicht auf die vom AIC favorisierte Klassenzahl, sondern auf die (geringere) Klassenzahl, welche der BIC vorschlägt, so kann der damit verbundene Informationsverlust ebenfalls wieder mittels eines Loss-of-Information-Index (Gleichung 29) gemessen werden.

Anhang

Anhang 1: Die Entropie eines Antwortmusters bei Geltung des Modells MX

Nach der shannon-wienerschen Formel errechnet sich die Entropie eines Antwortmusters aus

$$H_{MX} = -\sum_{\bar{y}} p_{\bar{y}} \, ld\left(p_{\bar{y}}\right),$$

worin sich das Summenzeichen über alle möglichen Antwortmuster $\left(\bar{y}\right)$ erstreckt und $ld\left(p_{\bar{y}}\right)$ den binären Logarithmus der Wahrscheinlichkeit $\left(p_{\bar{y}}\right)$ des Antwortmusters bezeichnet. In natürlichen Logarithmen ausgedrückt, kann diese Gleichung auch in der Form

$$H_{MX} = -\frac{1}{n}\sum_{\vec{y}} n\, p_{\vec{y}}\ \frac{\ln\left(p_{\vec{y}}\right)}{\ln(2)}$$

$$= -\frac{1}{n\ln(2)}\sum_{\vec{y}} n_{\vec{y}}\ \ln\left(p_{\vec{y}}\right)$$

dargestellt werden, worin $n_{\vec{y}}$ die Häufigkeit des Antwortmusters bezeichnet und der Ausdruck

$$\sum_{\vec{y}} n_{\vec{y}}\ \ln\left(p_{\vec{y}}\right) = \ln\left\{\prod_{\vec{y}} p_{\vec{y}}^{n_{\vec{y}}}\right\}$$

gleich dem natürlichen Logarithmus der Likelihood der Antwortmatrix ist, so dass

$$H_{MX} = -\frac{1}{n\ln(2)}\ln\left(L_{MX}\right).$$

Anhang 2: Die maximale Entropie eines Antwortmusters

Bei k Items mit m Antwortkategorien gibt es m^k mögliche Antwortmuster, deren maximale Entropie

$$H_{max} = \mathrm{ld}\left(m^k\right)$$

wegen $\mathrm{ld}\left(m^k\right) = k\ \mathrm{ld}(m)$ und nach Umrechnung in natürliche Logarithmen auch in der Form

$$H_{max} = k\ \frac{\ln(m)}{\ln(2)}$$

ausgedrückt werden kann.

Anhang 3: Die Redundanz eines Antwortvektors bei Annahme des Modells MX

Einsetzen für H_{max} und H_{MX} in

$$C_{abs,MX} = H_{max} - H_{MX}$$

ergibt

$$C_{abs,MX} = k\,\frac{\ln(m)}{\ln(2)} + \frac{1}{n\ln(2)}\ln\left(L_{MX}\right)$$

$$= \frac{n\,k\,\ln(m) + \ln\left(L_{MX}\right)}{n\ln(2)}.$$

Anhang 4: Der Loss-of-Information Index

Einsetzen für EP_{RM} und EP_{LCA} in

$$LI_{RM,LCA} = 1 - \frac{EP_{RM}}{EP_{LCA}}$$

ergibt

$$LI_{RM,LCA} = 1 - \frac{\frac{\ln(L_{RM}) - \ln(L_{PR})}{\ln(L_{sat}) - \ln(L_{PR})}}{\frac{\ln(L_{LCA}) - \ln(L_{PR})}{\ln(L_{sat}) - \ln(L_{PR})}}$$

$$= 1 - \frac{\ln(L_{RM}) - \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}.$$

Wegen

$$\frac{\ln(L_{LCA}) - L(L_{PR})}{\ln(L_{LCA}) - L(L_{PR})} = 1$$

kann dieser Ausdruck auch in der Form

$$LI_{RM,LCA} = \frac{\ln(L_{LCA}) - L(L_{PR})}{\ln(L_{LCA}) - L(L_{PR})} - \frac{\ln(L_{RM}) - \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

$$= \frac{\ln(L_{LCA}) - \ln(L_{PR}) - \ln(L_{RM}) + \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

$$= \frac{\ln(L_{LCA}) - \ln(L_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

geschrieben werden.

Anhang 5: Der korrigierte Loss-of-Information Index

Einsetzen für EP^*_{RM}, EP_{LCA}, VI^*_{max} und VI_{max} in

$$LI^*_{RM,LCA} = 1 - \frac{EP^*_{RM}}{EP_{LCA}} \frac{VI^*_{max}}{VI_{max}}$$

ergibt

$$LI^*_{RM,LCA} = 1 - \frac{\dfrac{\ln\left(L^*_{RM}\right) - \ln\left(L^*_{PR}\right)}{\ln\left(L^*_{sat}\right) - \ln\left(L^*_{PR}\right)}}{\dfrac{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)}{\ln\left(L_{sat}\right) - \ln\left(L_{PR}\right)}} \frac{\dfrac{\ln\left(L^*_{sat}\right) - \ln\left(L^*_{PR}\right)}{n\,\ln(2)}}{\dfrac{\ln\left(L_{sat}\right) - \ln\left(L_{PR}\right)}{n\,\ln(2)}}$$

$$= 1 - \frac{\ln\left(L^*_{RM}\right) - \ln\left(L^*_{PR}\right)}{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)},$$

welchen Ausdruck wir wegen

$$\frac{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)}{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)} = 1$$

auch in der Form

$$LI^*_{RM,LCA} = \frac{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)}{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)} - \frac{\ln\left(L^*_{RM}\right) - \ln\left(L^*_{PR}\right)}{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)}$$

$$= \frac{\ln\left(L_{LCA}\right) - \ln\left(L^*_{RM}\right)}{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)} - \frac{\ln\left(L_{PR}\right) - \ln\left(L^*_{PR}\right)}{\ln\left(L_{LCA}\right) - \ln\left(L_{PR}\right)}$$

schreiben können.

Konsequenzen des Ignorierens von globalen Modelltests in Studien wie TIMSS, PISA und PIRLS

Rolf Langeheine

1 Internationale Vergleichsstudien und das Problem des Modellfits

Mit den großen internationalen Studien wie TIMSS, PISA oder PIRLS hat das Rasch-Modell (Rasch 1960) weltweit seinen Siegeszug zur Messung von mathematischer, naturwissenschaftlicher und sprachlicher Kompetenz in einer Vielzahl von Ländern angetreten. Das ist verwunderlich, ist doch das Rasch-Modell dafür bekannt, dass es wegen seiner restriktiven Eigenschaften eher selten mit Daten verträglich ist. Sieht man sich diese Studien genauer an, so stellt man jedoch fest, dass formale Tests auf Übereinstimmung von Modell und Daten gar nicht durchgeführt werden. Das Vorgehen bei der Analyse ist wie folgt: Aus dem gesamten Datensatz wird im ersten Schritt eine Zufallsstichprobe von ca. 10% gezogen. Für diese wird das Rasch-Modell angepasst und mittels Itemfitstatistiken werden nicht zufriedenstellende Items eliminiert. Im zweiten Schritt werden getrennt für jedes Land Personenscores mit den aus Schritt Eins gewonnenen, *fixierten* Itemparametern berechnet. Hebt man diese Restriktionen auf und erlaubt für jedes Land in einer weiteren Analyse *freie* Itemparameter, so lässt sich zeigen, dass die Devianz zwischen Daten und Modell in der Regel sehr viel geringer ist.

In diesem Beitrag gehe ich zunächst auf die positiven Aspekte dieser Studien ein. Danach behandle ich verschiedene Möglichkeiten zur Beurteilung der Modellgültigkeit, wobei ich mit pi^* einen nicht mehr ganz neuen, aber relativ unbekannten Modelltest vorstelle. Schließlich berichte ich Ergebnisse für dieses sowie eine Reihe anderer Gütemaße aus Analysen eines kleinen Datensatzes von 42 Ländern und vier Chemieitems aus TIMSS95 und ende mit den sich daraus ergebenden Konsequenzen.

2 TIMSS, PISA, PIRLS/IGLU

In diesen Studien wird bereits seit Anbeginn (z. B. in der First International Mathematics Study, FIMS, und in der First International Science Study, FISS, aus den 60er- bzw. 70er-Jahren, vgl. Medrich & Griffith 1992) sehr viel Energie in die Definition und Ziehung von Stichproben sowie die Kontrolle des Stichprobenfehlers investiert. Gegeben eine repräsentative Stichprobe von einer wohl definierten Po-

pulation und homogene Items, so ist das Rasch-Modell aufgrund seiner Eigenschaften der attraktivste Kandidat zu Messung von Personenfähigkeiten. Leider gibt es keine generelle Repräsentativität (Pötter & Rendtel 1993). Eine Stichprobe kann für einzelne Variablen repräsentativ sein. Zumeist ist jedoch ein interaktives Geflecht von Variablen von Bedeutung, für das keine (amtlichen) Statistiken zur Überprüfung der Repräsentativität vorliegen. Der einzige Ausweg aus diesem Dilemma sind sauber definierte Zufallsstichproben. Auch dieses Kriterium ist in diesen Studien erfüllt. Die Daten werden nach einem Multi-Matrix-Design erhoben, wobei jeder Schüler nach Zufall ein Testheft aus einer Reihe solcher Hefte erhält. Um die Bearbeitungszeit gering zu halten, enthält jedes Testheft nur eine kleine Anzahl von Items des gesamten Itempools. Hinzu kommt eine noch kleinere Zahl von Kernitems, die in allen Testheften identisch sind. Die gemeinsamen Items dienen dazu, die Skala für die gesamte Stichprobe zu fixieren. Dieses Vorgehen hat u. a. den Vorteil, dass die zu messende Kompetenz durch einen langen Test exakter erfasst wird. Es hat den Nachteil, dass es in der Datenmatrix eine große Zahl von Löchern, fehlenden Daten gibt. Diese Löcher werden mittels multipler Imputation unter Rückgriff auf ein Hintergrundmodell gefüllt, genauer: geschätzt. In dieses Hintergrundmodell gehen nicht nur die eigenen Testitems ein, sondern alle weiteren Informationen über den betreffenden Schüler wie Schulnoten, Daten der Eltern etc. etc.

Dies sind gute Voraussetzungen dafür, dass sich das klassische Rasch-Modell als kompatibel mit den Daten erweisen könnte, mit all seinen Vorteilen:

1. Messung von Item- und Personenparametern auf ein und derselben Differenzenskala,
2. Vergleich der Fähigkeiten von Personen,
3. Vergleich der Schwierigkeiten von Items,
4. Vergleich von Personen mit Items.

3 Modellgeltungstests

3.1 Globale Tests

Da ich in dieser Arbeit lediglich vier dichotome Items in Form einer 4-dimensionalen Häufigkeitstabelle mit $k = 2^4$ Zellen/Antwortmustern verwende und Analysen des Rasch-Modells mit loglinearen sowie latenten Klassenmodellen durchführe, gehe ich zunächst auf statistische Tests ein, die "... die Gesamtheit aller Abweichungen ... berücksichtigen" (Rost 1996, S. 336) und dies durch verschiedene Goodness-of-Fit-Statistiken als auch eine Reihe deskriptiver Fitindizes erfassen.

Zur Familie der χ^2-Statistiken gehört die Likelihood-Ratio(LR)-Statistik, die definiert ist als

$$LR = -2\{\ln(L_{MX}) - \ln(L_{sat})\} \quad (1)$$

und mit df = k-1-n(P) Freiheitsgraden asymptotisch nach χ^2 verteilt ist. $\ln(L_{MX})$ steht für die logarithmierte Likelihood des Modells MX, $\ln(L_{sat})$ für jene des saturierten Modells und n(P) für die Anzahl der geschätzten Modellparameter. Da bestimmte Regularitätsbedingungen (vgl. Bishop et al. 1975) für meine Daten erfüllt sind und alternative Statistiken wie das Pearson χ^2 oder die Read-Cressie-Kompromissstatistik für die 42 Länder und das Rasch-Modell zu praktisch gleichen Ergebnissen kommen, gehe ich auf letztere nicht weiter ein.

χ^2-basierte Statistiken haben den Nachteil, dass sie eine direkte Funktion der Größe der Stichprobe sind. Häufig werden daher deskriptive Fitindizes bevorzugt. Zu ihnen gehören informationstheoretische Maße wie AIC (Akaike 1974) und BIC (Schwarz 1978), die sich auf Basis der logarithmierten Likelihood oder auf Basis der Likelihood-Ratio-Statistik definieren lassen. Ich verwende in dieser Arbeit die letztgenannte Version, also

$$AIC^* = LR - 2\,df \tag{2}$$

und

$$BIC^* = LR - \ln(n)\,df \tag{3}$$

wobei n die Stichprobengröße bezeichnet. Je kleiner AIC bzw. BIC sind, desto besser.

Ein weiteres deskriptives Maß ist der Index of Dissimilarity

$$D = \frac{\sum_{s=1}^{k} \left| fo_s - fe_s \right|}{2n} \tag{4}$$

(vgl. Shockey 1988), wobei fo_s und fe_s die beobachteten und erwarteten Häufigkeiten der k Zellen der Tabelle sind. Dies ist ein intuitiv sinnvolles Maß. Bei perfektem Fit nimmt D den Wert Null an. Allerdings ist für D keine Obergrenze definiert.

Während χ^2-basierte Tests die Prüfung erlauben, ob ein Modell mit den Daten verträglich ist, lassen AIC, BIC und D lediglich zu, das "Beste" aus einer Anzahl verschiedener Modelle für ein und denselben Datensatz zu identifizieren. Von Vorteil ist die Möglichkeit, auch nicht ineinander geschachtelte Modelle zu vergleichen.

Mit pi^* gibt es einen bisher kaum bekannten, aber sehr attraktiven Index zur Beurteilung der Modellgültigkeit. Die pi^* zugrunde liegende Idee ist denkbar simpel: Zerlege die Variation der Daten in zwei Komponenten, wobei sich die durch das substanzielle Modell erklärbaren Daten in Komponente 1 befinden und der nicht erklärbare Rest in Komponente 2. pi^* ist dann der Anteil der *nicht* erklärbaren Daten und 1 - pi^* gibt den Anteil der Daten wieder, der durch das Modell erklärt werden kann. Es handelt sich also um ein Mischverteilungsproblem, das sich an einem einfachen, hypothetischen Beispiel für eine Häufigkeitstabelle mit zwei dichotomen Items A und B und somit 4 möglichen Antwortmustern veranschaulichen lässt (vgl.

Tab. 1), wobei die Annahme geprüft werden soll, dass zwischen den beiden Items Unabhängigkeit besteht.

Antwortmuster		Häufigkeit	=	Komponente 1	+	Komponente 2
A	B	n(A, B)		n(A, B)		n(A, B)
0	0	20		20		0
0	1	80		80		0
1	0	30		15		15
1	1	60		60		0
Summe		190		175		15

Tabelle 1: Ein hypothetisches Beispiel für zwei 0/1 kodierte Items A und B und die Zerlegung der Daten in zwei Komponenten (Komponente 1 = Unabhängigkeit; Komponente 2 = Rest).
n(A, B) = Häufigkeit des Antwortmusters (A, B)

Wenn das Unabhängigkeitsmodell gelten soll, dann muss das Verhältnis der Häufigkeiten n(B=0) zu n(B=1) für A = 0 und A = 1 dasselbe sein. Das kann man erreichen, indem man 15 der 30 Fälle mit Antwortmuster (1, 0) in die Komponente 2 überträgt. Dann beträgt n(B=0)/n(B=1) in der Komponente 1 jeweils 0.25 und, da 15 der 190 Fälle in Komponente 2 übertragen wurden, ist pi^* = 15/190 = 0.079. Mit anderen Worten: Das substanzielle Modell erklärt 92.1% der Daten.

Erstmalig wurde dieser Index von Rudas et al. (1994) vorgestellt, vgl. auch Clogg et al. (1995) und Rudas (2002). Diese Autoren arbeiten allerdings mit relativ kleinen Häufigkeitstabellen. Zur Prüfung des Rasch-Modells wurde pi^* von Dayton (2003) und Formann (2006) benutzt. Eine Erweiterung auf Mixed-Rasch-Modelle (Rost 1990) findet sich bei Revuelta (2008). In Langeheine (2006) habe ich den pi^*-Ansatz u. a. als Alternative zur Identifikation von Typen in der Konfigurationsfrequenzanalyse (KFA, Krauth & Lienert 1973) verwendet.

Generell lässt sich das Problem wie folgt beschreiben. Die Komponente 1 hat eine durch das substanzielle Modell festgelegte Struktur, die Komponente 2 ist dagegen völlig strukturfrei. Für das Unabhängigkeitsmodell in einer Tabelle mit zwei dichotomen Items werden zwei Parameter geschätzt, so dass df = 4-1-2 = 1. Für Komponente 2 kann somit nur noch ein Parameter geschätzt werden. Generell muss die zweite Komponente df Zellen *ungleich* Null haben (wobei df gleich Anzahl der Freiheitsgrade des *substanziellen Modells* ist). Anders ausgedrückt: Die Komponente 2 muss k - df "Nullzellen" haben. Da alle k Zellen dieser Komponente geschätzt werden, muss für "Null" ein Kriterium spezifiziert werden, z. B. 0.00009. Es gilt also, einen Suchalgorithmus zu entwickeln, der diese "Null"- oder "Nichtnull"zellen findet. Für größere Tabellen ist die Lösung des Problems allerdings alles andere als einfach, da folgende Bedingungen für die Akzeptanz einer Lösung erfüllt sein müssen:

1. Beide Komponenten zusammen müssen die Daten perfekt erklären, d. h. LR und df müssen Null sein.

2. pi^* muss minimiert werden. In obigem Beispiel könnte man anstatt der 15 Fälle mit Antwortmuster (0, 1) auch 40 der 80 Fälle mit Antwortmuster (1, 0) in Komponente 2 übertragen und erhielte so ebenfalls Unabhängigkeit. Dann betrüge pi^* allerdings 0.211.
3. Komponente 2 muss k-df "Nullzellen" haben.
4. Das Gesamtmodell mit beiden Komponenten muss identifiziert sein.

Im Vergleich mit anderen hier aufgeführten Indizes hat pi^* folgende Vorteile:

1. Der Index ist eindeutig definiert im Bereich 0 bis 1 und quantifiziert somit direkt die Diskrepanz zwischen Modell und Daten.
2. Er ist unabhängig von der Stichprobengröße.
3. Es können sowohl Vergleiche verschiedener Modelle für ein und denselben Datensatz als auch Vergleiche ein und desselben Modells für verschiedene Datensätze durchgeführt werden.
4. Sofern die benutzte Software einen Identifikationstest bietet, steht in der Regel auch ein Schätzer für den Standardfehler von pi^* zur Verfügung. Dies erlaubt einen formalen Test der Hypothese, dass das Konfidenzintervall von pi^* den Wert Null einschließt. Ebenso lassen sich Tests auf Gleichheit des Fits zweier Modelle durchführen (Langeheine 2006).
5. Praktiker sind häufig nicht daran interessiert, ob ein Modell nach signifikanzstatistischen Kriterien mit Daten verträglich ist, sondern daran, welchen Teil der Daten ein Modell erklärt. Es kann also a priori ein Kriterium für pi^* spezifiziert werden, das nicht überschritten werden sollte.

3.2 Test auf Homogenität von Items

Wenn sich das Rasch-Modell durch einen globalen Modelltest als kompatibel mit den Daten erweist, dann werden Items als raschhomogen bezeichnet. Bei den internationalen Studien werden allerdings keine formalen globalen Modelltests durchgeführt. In der häufig benutzten Software ConQest (Wu et. al. 1997) wird zwar die deviance = $-2\ln(L_{MX})$ ausgegeben, aber nur zur formalen Testung des relativen Fits jeweils zweier Modelle für ein und denselben Datensatz herangezogen. Die Güte des Rasch-Modells wird hingegen anhand von Itemfitstatistiken für die Übereinstimmung der modellierten mit der empirischen Itemfunktion beurteilt: "The mean square fit statistic is an index of the fit of the data to the assumed scaling model ... Under the null hypothesis that the data and the model are consistent, the expected value of these statistics is one" (Adams et al. 1997, S. 121). Der Umkehrschluss von Itemfitstatistiken nahe Eins auf Konsistenz von Modell und Daten folgt daraus nicht zwingend.

3.3 Test auf Inhomogenität von Personen

Wenn das Rasch-Modell *nicht* mit den Daten verträglich war, dann wurden anfangs Items dafür verantwortlich gemacht und eliminiert. Auch heute ist das weiterhin gängige Praxis. Später wurden raschinhomogene Personen gesucht und ausge-

schieden. Dass dies nicht unbedingt nötig ist, verdanken wir Rost (1990), der mit dem Mixed-Rasch-Modell ein Modell vorgestellt hat, das verschiedene Gruppen von Personen mit einem jeweils eigenen Satz von Itemparametern zulässt, wobei es z. B. auch möglich, ist Gruppen zu identifizieren, die sich lediglich nach einem bestimmten response set verhalten.

4 Vier Chemie-Items aus TIMSS95 – wie passt das Rasch-Modell?

4.1 Daten und Details der Modellierung

In Kapitel 4.2 berichte ich Ergebnisse für vier dichtotome Chemieitems aus TIMSS95, Population 2, 13-14Jährige. Es handelt sich um die 0/1 kodierten Items (0 = falsche, 1 = richtige Antwort)

- F06: Relate rusting iron to the presence of oxygen and moisture.
- A09: Relate fire temperature to oxygen supply.
- G10: Select correct statement regarding the atomic makeup of matter.
- M10: Identify substances which are mixtures.

Dies sind Kernitems. Somit umgehe ich das Problem fehlender Daten aufgrund des Multi-Matrix-Designs. Zusätzlich habe ich Personen mit fehlenden Werten anderer Art ausgeschlossen. Der gesamte Datensatz hat danach n = 36.281 Fälle, summiert über 42 Länder. Datenbasis sind Häufigkeitstabellen der Größe 2^4.

Die Analysen erfolgten in drei Schritten. Im ersten Schritt wurde für jedes der 42 Länder und die Gesamtstichprobe das Rasch-Modell ohne Spezifikation von pi^* angepasst. In diesem wie in allen weiteren Schritten wurde es als loglineares Modell (vgl. Kelderman 1984, Langeheine & van de Pol 1990) spezifiziert. Es handelt sich somit um eine Conditional-Maximum-Likelihiood(CML)-Schätzung, wobei 3 Item- und 4 Scoreparameter geschätzt werden. Das Rasch-Modell hat also 16-1-3-4 = 8 Freiheitsgrade. Von Interesse sind in diesem Schritt die Statistiken LR mit mit assoziiertem p, D, AIC^* und BIC^* (vgl. Tab. 2). Im zweiten Schritt wurde ein latentes Klassenmodell mit zwei Klassen spezifiziert, wobei sich in der ersten Klasse das Rasch-Modell befindet und in der zweiten Klasse mit pi^* der Teil, der durch das Raschmodell *nicht* erklärt wird. Diese Klasse muss acht "Nullzellen" enthalten. Der dritte Schritt ist identisch mit dem zweiten, mit einer Ausnahme: Die Analysen wurden nur für die 42 Länder durchgeführt, wobei die drei Itemparameter auf diejenigen der Gesamtstichprobe aus Schritt Zwei *fixiert* wurden. Daher muss die zweite Klasse nur 8-3 = 5 "Nullzellen" enthalten (und analog die erste Klasse 16-1-4 = 11 Zellen ungleich Null). Dies ist das Vorgehen analog zu TIMSS.

Die Analysen erfolgten mit einer experimentellen Version des Programms LEM von Vermunt (1997). Da es sich bei den Schritten Zwei und Drei um maximal überparametrisierte Modelle handelt (in beiden Fällen müssen so viele Parameter geschätzt werden, dass die gesamte Variation der Daten durch ein saturiertes Modell erklärt wird), ist es unbedingt notwendig, sich nicht auf eine oder wenige Analysen je Datensatz zu verlassen. Ich habe daher generell 20 Replikationen durchgeführt,

wobei bestimmte Steuerparameter von LEM modifiziert wurden. Wenn auch darunter keine Lösung war, die die erforderliche Zahl von "Nullzellen" in der zweiten Komponente auswies, wurden weitere Replikationen durchgeführt Diese Replikationen sind aus einem zweiten Grund notwendig. Selbst bei Tabellen dieser Größe kann man sich nicht absolut sicher sein, dass pi^* minimiert wird. Wenn aber viele aus einem Satz von Replikationen etwa zum gleichen Ergebnis für pi^* kommen, wird man mit hoher Wahrscheinlichkeit nahe beim echten Minimum liegen. Schließlich möchte ich noch anmerken: In Kapitel 3.1 habe ich die Forderung aufgestellt, dass das Modell identifiziert sein muss. Leider bietet die hier verwendete Version von LEM keinen Identifikationstest. Aus meiner Arbeit (Langeheine 2006) mit dem Programm PANMARK (Van de Pol et al. 2000, mit dem sich Rasch-Modelle nicht spezifizieren lassen) ist mir bekannt, dass es Fälle von nicht identifizierten Modellen geben kann, obwohl alle anderen oben genannten Bedingungen erfüllt sind.

4.2 Ergebnisse

In Tabelle 2 finden sich die Ergebnisse verschiedener Modelltestindizes für 42 Länder und die Gesamtstichprobe. Bei 18 dieser Länder zeigt sich nach dem LR-χ^2-Test eine signifikante Abweichung vom Rasch-Modell ($p< 0.05$). Legt man einen Schnittpunkt von 0.5 für p zugrunde, so gehören Kanada, Kolumbien, die Slowakei, Dänemark und Korea zu den Ländern, für die das Rasch-Modell am besten mit den Daten verträglich ist. Nach dem Index of Dissimilarity, der oft als Prozent Nichtübereinstimmung von Modell und Daten interpretiert wird, könnte man vermuten, dass eher eine bessere Übereinstimmung als nach dem LR-Test vorliegt: Der Mittelwert der 42 Länder beträgt 0.040 (vgl. Tab. 4). Das Maximum liegt für Isreal bei 0.100, Kanada zeigt mit nur 0.017 die beste Anpassung. Lediglich acht Länder liegen über 0.05 Nichtübereinstimmung von Modell und Daten. Man beachte: Die Obergrenze von D ist nicht definiert. Nach dem AIC^*-Index (je kleiner, desto besser) weisen die 18 Länder, die im LR-Test signifikant vom Rasch-Modell abweichen, die größten Werte auf. Mit wenigen Ausnahmen trifft dies ebenfalls für den BIC^*-Index zu. Das ist nicht verwunderlich, sind doch beide Indizes LR-basiert.

Nach pi^* beträgt der mit dem Raschmodell *nicht* kompatible Anteil im Mittel 8.9% (vgl. Tab. 4). Mit lediglich 2.9% weist die Slowakei die beste Anpassung auf. Kuwait bildet mit 17.1% das Schlusslicht. Insgesamt überschreiten 17 Länder die 10%-Grenze. Für 11 dieser 17 gilt, dass das mit LR assoziierte $p < 0.05$ ist. Es wird somit deutlich, dass die verschiedenen Indizes durchaus gewisse Gemeinsamkeiten haben.

Die pi^* Werte für das Raschmodell mit fixierten Itemparametern aus der Gesamtstichprobe (Spalte pi^*_{fix} in Tab. 2) zeigen generell eine erheblich schlechtere Gültigkeit des Modells. Der Mittelwert aus allen 42 Ländern steigt auf 33.8% (vgl. Tab. 4) Nichtübereinstimmung von Modell und Daten. Mit Ausnahme von Kanada unterschreitet kein Land die 10%-Grenze. Im Bereich von 10 bis 20% liegen sieben Länder, zehn zwischen 20 und 30%, elf zwischen 30 und 40%, sechs zwischen 40

Land	n	L-Ratio	p	D	AIC*	BIC*	pi*	pi^*_{fix}
Australien	1594	15.4	.052	.028	-0.6	-43.6	.0500	.1985
Österreich	726	12.6	.126	.027	-3.4	-40.1	.0627	.1698
Belgien (fläm.)	700	22.2	.005	.052	6.2	-30.3	.1089	.5521
Belgien (franz.)	601	12.4	.134	.052	-3.6	-38.8	.1290	.2215
Bulgarien	333	7.7	.469	.038	-8.4	-38.8	.0866	.2100
Kanada	2061	6.6	.576	.017	-9.4	-54.4	.0426	.0978
Kolumbien	652	6.8	.564	.030	-9.3	-45.1	.0939	.5798
Zypern	761	32.0	.000	.059	16.0	-21.1	.1208	.3798
Tschechien	820	21.3	.007	.052	5.3	-32.4	.1099	.2975
Slowakei	877	4.3	.826	.020	-11.7	-49.9	.0287	.3539
Dänemark	559	2.8	.948	.025	-13.2	-47.9	.0417	.3927
Frankreich	747	9.3	.322	.037	-6.8	-43.7	.0661	.3126
Deutschland	719	8.5	.384	.035	-7.5	-44.1	.0929	.1505
Griechenland	994	16.5	.035	.037	0.5	-38.7	.1142	.2700
Hongkong	837	14.7	.065	.032	-1.3	-39.1	.0718	.4300
Ungarn	746	8.4	.400	.028	-7.7	-44.6	.0499	.1265
Island	460	7.6	.478	.035	-8.4	-41.5	.0492	.5481
Iran	923	18.6	.017	.031	2.6	-36.0	.0755	.5333
Irland	735	16.1	.041	.046	0.1	-36.9	.1080	.4287
Isreal	173	13.8	.087	.100	-2.2	-27.4	.1223	.1573
Japan	1280	12.1	.146	.026	-3.9	-45.1	.0641	.3324
Korea	724	4.9	.773	.028	-11.1	-47.8	.0591	.2001
Kuwait	184	11.8	.161	.078	-4.2	-29.9	.1711	.5965
Lettland	610	17.9	.022	.050	1.9	-33.5	.1143	.4163
Litauen	641	8.7	.366	.045	-7.3	-43.0	.0895	.4072
Niederlande	508	8.1	.428	.028	-7.9	-41.8	.1068	.7197
Neuseeland	864	16.9	.032	.039	0.9	-37.2	.0877	.3534
Norwegen	722	10.5	.232	.035	-5.5	-42.2	.1137	.5097
Philippinen	1477	20.2	.010	.037	4.2	-38.1	.0825	.3549
Portugal	838	25.7	.001	.062	9.7	-28.1	.1413	.4100
Rumänien	915	21.5	.006	.048	5.5	-33.0	.1143	.3230
Russ. Föd.	1021	8.6	.376	.022	-7.4	-46.8	.0467	.2735
Singapur	1019	23.8	.003	.035	7.8	-31.7	.0683	.2558
Südafrika	1226	21.8	.005	.035	5.8	-35.1	.0746	.3432

Land	n	L-Ratio	p	D	AIC*	BIC*	pi*	pi^*_{fix}
Spanien	946	20.4	.009	.043	4.4	-34.5	.1241	.3562
Schweden	1134	16.4	.037	.037	0.4	-40.0	.0614	.1803
Schweiz	1445	12.2	.141	.026	-3.8	-46.0	.0619	.1917
Thailand	1462	12.0	.153	.029	-4.0	-46.3	.0653	.4812
England	445	20.7	.008	.063	4.7	-28.1	.1620	.2785
Schottland	726	15.1	.058	.040	-0.9	-37.6	.1081	.2079
USA	1356	24.1	.002	.048	8.1	-33.6	.0800	.2572
Slowenien	702	18.1	.021	.048	2.1	-34.4	.1010	.3176
Gesamt	36281	226.5	.000	.022	210.5	142.5	.0645	

Tabelle 2: Modelltestindizes für 42 Länder und die Gesamtstichprobe.
n = Stichprobengröße; L-Ratio = Likelihood-Ratio, p = die damit assoziierte Wahrscheinlichkeit bei df = 8 Freiheitsgraden; D = Index of Dissimilarity; AIC* = Akaike Information Criterion; BIC* = Bayesian Information Criterion; pi* bzw. pi^*_{fix} = nicht raschkonformer Anteil im freien bzw. fixierten Raschmodell

und 50%. Belgien (Flamen), Kolumbien, Island, Iran, Kuwait, die Niederlande und Norwegen überschreiten die 50%-Marke. Generell gilt also: Das Rasch-Modell erklärt unter dieser Bedingung einen viel zu geringen Teil der Daten.

Anwortmuster	beobachtet	RM	Rest	RM_{fix}	$Rest_{fix}$
0 0 0 0	26	26.0	**0**	26.0	**0**
0 0 0 1	13	8.5	4.5	7.6	5.4
0 0 1 0	14	5.8	8.2	8.9	5.1
0 0 1 1	4	2.9	1.1	4.0	**0**
0 1 0 0	38	28.0	10.0	24.2	13.8
0 1 0 1	14	14.0	**0**	11.0	3.0
0 1 1 0	18	9.5	8.5	12.9	5.1
0 1 1 1	27	12.3	14.7	10.9	16.1
1 0 0 0	44	44.0	**0**	44.0	**0**
1 0 0 1	22	22.0	**0**	20.1	1.9
1 0 1 0	26	14.8	11.2	23.5	2.5
1 0 1 1	28	19.2	8.8	19.9	8.1
1 1 0 0	72	72.0	**0**	63.8	8.2
1 1 0 1	93	93.0	**0**	54.0	39.0
1 1 1 0	63	63.0	**0**	63.0	**0**
1 1 1 1	217	217.0	**0**	217.0	**0**
Summe	719	652.0	67.0	610.8	108.2
Prozent	100.0%	90.7%	9.3%	84.9%	15.1%

Tabelle 3: Deutschland: Beobachtete und erwartete Häufigkeiten für das Rasch-Modell bei freien (RM) und fixierten (RM_{fix}) Itemparametern. Der mit dem Rasch-Modell nicht verträgliche Anteil befindet sich jeweils in der Spalte Rest. "Nullzellen" sind fett markiert.

Zur Veranschaulichung des in Kapitel 3.1 beschriebenen Problems der notwendigen "Nullzellen" in der Komponente des 2-Klassen-Modells, die mit pi^* bzw. pi^*_{fix} den Anteil der durch das substanzielle Modell *nicht* erklärten Variation der Daten enthält, sind in Tabelle 3 die Ergebnisse für Deutschland wiedergegeben, in diesem Fall in Form der erwarteten Häufigkeiten. Die "Nullzellen" sind, wie gefordert, acht im Fall freier und fünf im Fall restringierter Itemparameter. Auch die Bedingung, dass die beiden Extremzellen mit den Antwortmustern (0, 0, 0, 0) und (1, 1, 1, 1) für das Rasch-Modell perfekt angepasst sein müssen, ist erfüllt. Deutschland gehört zu den sieben Ländern, die unter der Bedingung fixierter Itemparameter die Grenze von 20% Misfit unterschreiten. Gut ein Drittel des Misfits von 15.1% geht mit 39/719 = 5.4% auf die Zelle mit dem Antwortmuster (1, 1, 0, 1) zurück. Betrachtet man die erwarteten Häufigkeiten des RM wie des RM_{fix}-Parts als "Daten", so lässt sich leicht prüfen, dass sie dem Rasch-Modell perfekt entsprechen.

In Tabelle 4 sind deskriptive Statistiken der Modelltestindizes zusammengestellt. Wie zu erwarten, sind die direkt vergleichbaren Indizes (p, D, AIC^*, BIC^* und pi^*) vergleichsweise hoch korreliert (vgl. Tab. 5). Unterzieht man diese fünf Maße einer Hauptkomponentenanalyse, so erklärt bereits die erste Komponente 74% der Varianz mit Ladungen zwischen $|0.82|$ und $|0.96|$ (vgl. Tab. 5, letzte Zeile). Somit wird bestätigt, dass alle diese Indizes Modellfit messen, wenn auch auf unterschiedliche Art.

	p	D	AIC^*	BIC^*	pi^*	pi^*_{fix}
Mittelwert	.203	.040	-1.507	-38.767	.089	.338
Std. Abw.	.253	.016	6.617	7.003	.033	.144
Minimum	.000	.017	-13.200	-54.400	.029	.098
Maximum	.948	.100	16.000	-21.100	.171	.720

Tabelle 4: Deskriptive Statistiken für die Modelltestindizes p, D, AIC^*, BIC^*, pi^* und pi^*_{fix}

	p	D	AIC^*	BIC^*	pi^*
p	–				
D	-.466	–			
AIC^*	-.845	.464	–		
BIC^*	-.714	.806	.832	–	
pi^*	-.511	.786	.476	.742	–
Ladungen	.83	-.82	-.85	-.96	-.82

Tabelle 5: Interkorrelationen der Modelltestindizes p, D, AIC*, BIC* und pi* sowie Ladungen auf der 1. Komponente der Hauptkomponenten-Analyse

5 Welche Konsequenzen ergeben sich daraus? Was kann man tun?

Wenn bereits nur vier Items bei einer Reihe von Ländern als nicht konform mit dem Rasch-Modell angesehen werden müssen und dies für alle Länder unter der Bedingung fixierter Itemparameter auf der Basis der Gesamtstichprobe gilt, so ist zu erwarten, dass dieses Problem bei höherer Zahl von Items ungeahnte Ausmaße annimmt. Für mich ist dieses Ergebnis nicht verwunderlich. Ende der 90er-Jahre bin ich dieses Problem für diese Daten mit Anpassung von latenten Klassenmodellen sowie dem Rasch-Modell für mehrere Gruppen (hier: Länder) angegangen. Während des 11th International Meeting of the Psychometric Society 1999 in Lüneburg habe ich mit Ray Adams (damals Chair of the Technical Advisory Committee der IEA und somit von TIMSS) meine Ergebnisse diskutiert. Seine Antwort war: "Yes, we know, there is no item invariance across countries". 1999 gab es allerdings keine Möglichkeit, das pi*-Konzept zu nutzen. Wie man aus Tabelle 1 sieht, ist die Modellanpassung nach pi* für eine Reihe von Ländern noch passabel. Dies gilt auch für die Gesamtstichprobe mit pi* = 0.0645. Erst die Fixierung der Itemparameter auf diejenigen der Gesamtstichprobe macht die Heterogenität zwischen den Ländern offensichtlich. Mit anderen Worten: Es herrscht item bias oder differential item functioning (DIF). Van de Vijver & Tanzer (2004) geben einen ausführlichen Überblick über die verschiedenen Typen und Quellen von bias wie über Möglichkeiten, dem Problem zu begegnen. Wichtig sind in diesem Zusammenhang die von ihnen spezifizierten drei Bedingungen, die erfüllt sein müssen, damit Ergebnisse über Länder hinweg vergleichbar sind: 1) In jedem Land wird dasselbe Konstrukt gemessen, 2) die Messeinheit und 3) der Ursprung der Skala sind für alle Länder identisch – erst dann liegt "... completely bias-free measurement" (ebd., S. 122) vor.

In TIMSS95 sind die Bedingungen Zwei und Drei durch die Fixierung der Itemparameter aus Schritt Eins der Analyse bei der Bildung von Skalenwerten in Schritt Zwei erfüllt – allerdings unter Vernachlässigung des Konstruktbias. Offensichtlich hat man bei TIMSS erkannt, dass das Rasch-Modell für diese Art von Messungen wegen seiner restriktiven Eigenschaften von parallelen Itemfunktionen nicht geeignet ist. Seit TIMSS99 (Martin et al. 2000) wird daher das dreiparametrige logistische Modell (3PL) verwendet, das über das Rasch-Modell hinaus itemspezifische Trennschärfen und Ratewahrscheinlichkeiten zulässt. Man beachte allerdings, dass schon mit dem 2PL/Birnbaummodell (Birnbaum 1968) die herausragende Eigenschaft des Rasch-Modells, die spezifisch objektive Messung, verloren geht. Rost (1996, S. 134) gibt ein anschauliches Beispiel von den Folgen. Mit TIMSS99 wurde zugleich TIMSS95 nach dem 3PL reskaliert, um Vergleiche zwischen beiden Zeitpunkten zu ermöglichen. Das 3PL findet ebenso in der internationalen Grundschul-Lese-Untersuchung PIRLS (zu Deutsch IGLU, vgl. Bos et al. 2003) Anwendung.

Die PISA-Studien (2000, 2003 und 2006) verwenden dagegen weiterhin das klassische Rasch-Modell (vgl. Carstensen et al. 2007). Das muss nicht heißen, dass Daten aus PISA und älteren Studien nicht auch mit anderen Modellen untersucht wurden. Ein jüngstes Beispiel sind Oliveri & von Davier (2010) mit zahlreichen Verweisen auf frühere Studien zum Problem der mangelnden Anpassung des Rasch-

Modells. Sie vergleichen den Anteil von Itemmisfit für folgende Modelle: 1) 2PL vs. Rasch-Modell; 2) 2PL 2-Klassen Mischverteilungsmodell vs. 2-Klassen-Rasch-Mischverteilungsmodell; 3) Mehrgruppen 2PL vs. Mehrgruppen-Rasch-Modell; 4) Mehrgruppen 2PL, wobei für einen Teil der Items identische Parameter für alle Gruppen verlangt werden und für einen anderen Teil gruppenspezifische Parameter zugelassen sind. Die Beurteilung der Modellanpassung nach informationstheoretischen Indizes zeigt, dass das 2PL dem Rasch-Modell jeweils überlegen ist. Als am besten angepasst erweist sich das Mehrgruppen-2PL mit für Länder partiell spezifischen Itemparametern durch Reduktion des Itemmisfits um 17% bis 74%. Die interessanteste Botschaft dieser Studie ist allerdings das Ergebnis auf Systemniveau: Die Korrelationen der Ländermittelwerte auf Basis der rein internationalen Itemparameter und der partiell länderspezifischen Itemparameter betragen 0.997 für die Naturwissenschaften, 0.987 für Lesen und 0.997 für Mathematik. Oliveri & von Davier schlagen daher die Verwendung des allgemeineren 2PL mit einer quasi internationalen Skala vor, die sich aus einer großen Zahl von für alle Länder gemeinsamen Itemparametern sowie länderspezifischen Parametern für besonders schlecht passende Items zusammensetzt. Die hinter diesem Vorschlag stehende Idee besteht darin, dass die gemeinsamen Items die Skala definieren und die länderspezifischen Items zur exakteren Messung beitragen.

Welche Konsequenzen lassen sich hieraus ziehen? Das Problem, dass das Rasch-Modell oft mit Daten nicht verträglich ist, ist bekannt. Puristen würden daraus den Schluss ziehen, dass es in solchen Fällen nicht zur Anwendung kommen kann und zuvor kulturfreie bzw. kulturfaire Messinstrumente entwickelt werden müssten. Praktiker dagegen würden für eine von zwei Alternativen plädieren: 1) Anwendung des Raschmodells trotz Misfit oder 2) Wahl eines alternativen, weniger restriktiven Modells (das sich im Prinzip allerdings auch als kompatibel mit den Daten erweisen müsste). Das Ergebnis von Oliveri & von Davier entbehrt für beide Alternativen nicht eines gewissen Charmes. Es wird sich jedoch zeigen müssen, ob es auf Systemebene repliziert werden kann.

Schließlich einige weitere Anmerkungen zum pi*-Konzept. Soweit mir bekannt ist, gibt es immer noch keine generelle Software, insbesondere für größere Probleme, d. h. eine größere Anzahl von Items und größere Modelle. In Kapitel 3.1 habe ich ausgeführt, welche Bedingungen im Prinzip erfüllt sein müssen, um ein Resultat akzeptieren zu können. Die spezielle von mir benutzte Version von LEM versagt z. B., wenn eine Tabelle wenigstens eine "Nullzelle" enthält. Der Grund dafür liegt darin, dass bei der iterativen Schätzprozedur von den beobachteten Häufigkeiten "etwas weggenommen" und in den pi*-Teil übertragen wird. Wo keine Masse vorhanden ist, lässt sich jedoch nichts wegnehmen. Dem kann man begegnen, indem man solche Zellen in beiden Komponenten auf Null fixiert, wie es Revuelta (2008) tut. Alternativ kann man zu allen beobachteten Häufigkeiten eine kleine Konstante addieren, z. B. 0.2. Für komplexe latente Klassenmodelle, und somit umso mehr für jedes hochgradig überparametrisierte pi*-Problem, ist es unbedingt erforderlich, sich nicht auf eine einzige Rechnung zu verlassen. Die beste Garantie zur Vermeidung der Lokalen-Minimum-Falle ist eine 3-stufige Startwerteprozedur wie in

PANMARK. Ist diese nicht verfügbar, so muss man sich mit Replikationen begnügen. In der hier benutzten Version von LEM wird die Iterationsprozedur zusätzlich durch spezielle Steuerparameter beeinflusst.

Nach den in Kapitel 3.1 angeführten Argumenten für das pi^*-Konzept dürfte offensichtlich sein, dass dies ein vielversprechender Ansatz ist. Praktiker, aber auch streng theoretisch orientierte Anwender haben die Möglichkeit, ihre Entscheidung über die Gültigkeit bzw. Brauchbarkeit eines Modells nicht allein nach signifikanzstatistischen Gesichtspunkten zu fällen, sondern können alternativ dazu a priori einen Grenzwert für pi^* definieren. Was bislang aussteht, ist die Entwicklung einer allgemeineren Software dafür. Ich bin relativ sicher, dass wir eines Tages die Publikation eines entsprechenden Algorithmus sehen werden.

Soziale Herkunft, kulturelles Kapital und Lesefähigkeit – ein Vergleich zwischen Finnland und Deutschland mit Hilfe der PISA 2000-Studie

Werner Georg

Die PISA 2000-Studie untersuchte ländervergleichend die Lesefähigkeit von Schülern der neunten Klasse in 32 Ländern. Finnland schnitt mit einem durchschnittlichen Testscore von 546 (bei einem internationalen Mittelwert von 500) am besten ab; Deutschland lag, und dieser Umstand führte zu erhitzten öffentlichen Diskussionen, mit 484 Punkten abgeschlagen auf dem 21. Platz. Noch beunruhigender war der Tatbestand, dass die Unterschiede in der Lesefähigkeit zwischen dem untersten und obersten Quartil des International Socio Economic Index (ISEI, ein metrisches Maß für den Berufsstatus) mit 111 Skalenpunkten am höchsten von allen beteiligten Ländern waren. Das Deutsche PISA-Konsortium erklärte diese Ungleichheit in der Lesefähigkeit mit einem familialen Prozessmodell, bei dem insbesondere die differenzielle Verfügung über kulturelles und soziales Kapital im Zentrum steht. Vor diesem Hintergrund ist es Ziel dieses Beitrags, mit einem Latent-Class-Messmodell die Effekte des familialen Kulturkapitals in Deutschland und Finnland zu vergleichen.

1 Theorie

Ein gut dokumentierter Tatbestand der Bildungssoziologie von den 60er- bis in die 90er-Jahre des letzten Jahrhunderts (Dahrendorf 1965, Peisert 1967, Handl 1985; Ditton 1992; Köhler 1992; Meulemann 1992, Müller 1998; Schimpl-Neimanns 2000; Vester 2004; Blossfeld & Shavit 1993) ist die Persistenz sozialer Ungleichheit im Bildungssystem während der Phase der Bildungsexpansion. Die PISA-Studie 2000 (Baumert & Schümer 2001, S. 323 ff.) zeigte, dass in Deutschland beim Vergleich des unteren und des oberen Verteilungsquartils des ISEI (Ganzeboom et al. 1992) die Lesefähigkeit von allen 32 untersuchten Ländern mit 111 Testpunkten am stärksten auseinanderklaffte. Bei Kontrolle der kognitiven Grundfähigkeiten und der Lesekompetenz war die Wahrscheinlichkeit, ein Gymnasium zu besuchen, für ein Kind aus der oberen Dienstklasse 2.96 mal so hoch wie für ein Kind aus einer Facharbeiterfamilie.

In den letzten 10 Jahren haben einige vergleichende Studien die Persistenzhypothese in Frage gestellt. Unter Verwendung von großen Datensätzen aus 8 europäischen Ländern haben Breen et al. (2009, 2010) gezeigt, dass in den Kohorten zwischen 1908-1924 und 1955-1964 die soziale Ungleichheit in allen untersuchten Ländern deutlich zurückgegangen ist. Andererseits fand Georg (2010) für den relativen Hochschulbesuch eine Fortdauer sozialer Ungleichheit auch während der Bildungsexpansion. So entsteht der Eindruck, dass sich zwar im unteren Bereich der Sekundarstufe deutliche Mobilitätsprozesse ergeben haben (etwa von der Hauptschule zur Realschule), dies für die tertiäre Bildung jedoch nur sehr bedingt der Fall ist.

Fragt man nach soziologischen Erklärungsansätzen dafür, lassen sich zwei Hauptrichtungen unterscheiden, Modelle der rationalen Wahl und konflikttheoretische Ansätze. Im Rational-Choice-Paradigma (Breen & Goldthorpe 1997; Boudon 1974; Becker 2003) wird angenommen, dass klassenspezifisch unterschiedliche Kosten-Nutzen-Kalküle zum Tragen kommen, wobei unter der Randbedingung des Statuserhalts drei Parameter von Bedeutung sind, nämlich die Bildungskosten unter Berücksichtigung der Opportunitätskosten, die erwarteten Bildungsrenditen und die Erfolgswahrscheinlichkeit. Boudon (1974) unterscheidet in diesem Zusammenhang zwei Arten von Herkunftseffekten: primäre und sekundäre. Während die primären Herkunftseffekte die mit den differenziellen Anregungspotenzialen der Schichtungsmilieus verbundenen kognitiven Fähigkeiten zu Beginn der Bildungslaufbahn beschreiben, beziehen sich die sekundären Herkunftseffekte auf die Entscheidungsprozesse an den Übergängen des Bildungssystems. Verschiedene Studien gewichten den Beitrag beider am Zustandekommen der Ungleichheit unterschiedlich. Während Erikson & Rudolphi (2010) für Schweden im Zeitverlauf ein Abnehmen beider feststellen und den Einfluss sekundärer Effekte auf etwa ein Drittel veranschlagen, kommen Maaz & Nagy (2009) für Deutschland in Hinblick auf drei übergangsrelevante Parameter in der vierten Klasse zu einer differenziellen Einschätzung: Für die Benotung wirken primäre Effekte stärker als sekundäre, für die Übergangsempfehlung sind beide etwa gleich wichtig und für den real vollzogenen Übergang sind die sekundären Effekte wichtiger.

Ein Problem der hier beschriebenen Modelle rationaler Wahl liegt darin, dass sie, in den Kategorien von Max Weber (1972), eher einen Idealtypus zweckrationalen Handelns beschreiben: Stellen wir uns vor, ein Mensch würde rational handeln, wie würde er sich dann verhalten? Gedankenexperimentell mag dieses Vorgehen interessant sein, jedoch hat Weber nie einen Zweifel daran gelassen, dass Idealtypen nie die Wirklichkeit sozialen Handelns beschreiben, sondern lediglich eine wissenschaftliche Konstruktion darstellen, die in dieser reinen Form nicht (oder nur selten) anzutreffen ist. Oder, um Pierre Bourdieus (1993a) diesbezüglichen Einwand aufzugreifen: Aufgrund seiner vom Handlungszwang und folglich auch von der Ökonomie der Praxis befreiten Stellung des wissenschaftlichen Subjekts im sozialen Raum konstruiert sich dieses ein Handlungsmodell, in dem das Handeln der Akteure übereinstimmt mit den mathematischen Modellen, die dieses Verhalten abbilden sollen. Diese Modelle verhalten sich aber zum realen Handeln, so Bour-

dieu, wie ein Stadtplan zur realen Stadt. Wohlgemerkt, es soll hier nicht darum gehen zu leugnen, dass sich im Entscheidungsprozess über eine Bildungslaufbahn Analogien und Elemente einer "rationalen" Entscheidung vorfinden lassen. Sicherlich variiert die Einschätzung der Erfolgswahrscheinlichkeit im Bildungssystem schichtspezifisch und ebenso ist eine Risikoaversion bildungsferner Schichten bezüglich der Entscheidung für eine Hochschulausbildung festzustellen. Ob diese Faktoren jedoch aus einer rationalen Wahl oder eher aus einer habituellen Distanz resultieren, lässt sich von außen schwer entscheiden.

Der zweite, konflikttheoretische Ansatz ist insbesondere mit den Schriften Pierre Bourdieus verbunden. Bildungssoziologische Fragestellungen durchziehen dessen gesamtes Werk und werden bereits in seiner gemeinsam mit Passeron veröffentlichten frühen Schrift *Reproduction* (1970) thematisiert. In den *Feinen Unterschieden* (1982) tauchen sie im Zusammenhang mit seiner verallgemeinerten Habitus- und Lebenstiltheorie auf und zielen auf die Frage der sozialen Statusvererbung unter nicht-ständischen Bedingungen ab. In *Reproduction* beginnt Bourdieu mit einer Analyse des relativen Hochschulbesuchs unterschiedlicher sozialer Klassen und der schichtspezifisch differenzierten Studienfachwahl. Um die scharfe soziale Selektivität des Hochschulsystems zu erklären, rekurriert Bourdieu einerseits auf die Ebene der sozialstrukturell differenziellen kulturellen Praxis in den Herkunftsfamilien, andererseits macht- und wissenssoziologisch auf die symbolische Definitionsmacht der oberen Statusgruppen hinsichtlich der legitimen Kultur und deren Passung mit den Anforderungen der studentischen und Fachkultur. Soziale Reproduktion unter meritokratisch legitimierten Bedingungen muss nach Bourdieu ihre Prozesslogik verschleiern, um weiterhin effizient sein zu können. Konkret bedeutet dies, dass die ungleiche Verteilung kulturellen Kapitals in verschiedenen Segmenten der Sozialstruktur in der Weise naturalisiert wird, dass sie als individuell zurechenbarer Begabungs- und Fähigkeitsunterschied überhöht wird. In seiner Kapitaltheorie, die den Kern der sozialen Struktur konzeptualisiert, unterscheidet Bourdieu (1983) drei Formen kulturellen Kapitals: inkorporiertes, institutionalisiertes und objektiviertes Kulturkapital.

Das inkorporierte Kulturkapital wird im familialen Lebensstil der Herkunftsfamilie internalisiert und so zu einem "natürlichen" Teil des Habitus. Die kulturellen Praxen der Familie werden als virtuelles Muster in einem langwierigen Prozess angeeignet und die Aktivierung dieses Musters wirkt auf das Gegenüber als integraler Bestandteil der Persönlichkeit. An verschiedenen Stellen betont Bourdieu, dass es gerade diese scheinbare Natürlichkeit des inkorporierten Kulturkapitals und des mit ihm einhergehenden Habitus ist, die seine Effizienz bei der sozialen Reproduktion der herrschenden Klassenfraktionen ausmacht. Im Vergleich hierzu wirkt der Habitus des Kleinbürgertums, das aufgrund seiner Position im sozialen Raum soziale Unterschiede als Unterschiede in der Bildungskompetenz wahrnimmt (Alladoxia-Effekt), bemüht, zwanghaft und aufgesetzt. Da das Kleinbürgertum sich an trivialisierten und abgestiegenen Elementen der legitimen Kultur orientiert, fehlt ihm einerseits ein "Anlagesinn" für diejenigen kulturellen Praxen und Wissensbestände, die in den Institutionen des Bildungssystems gewinnversprechend sind

und andererseits die Natürlichkeit des Habitus, die einer entsprechenden Selbstrepräsentation Wirksamkeit verleiht. Zwar verfügt der Habitus der Arbeiterklasse, so Bourdieu, nicht über die Künstlichkeit des Habitus des Kleinbürgertums, die demjenigen anhängt, der Fremdes kopiert und imitiert, aber gleichzeitig ist ihr die Bildungsbezogenheit dieser Gruppe fremd und der soziale Aufstieg durch höhere Bildungsinstitutionen wird als Abweichung bestraft.

Das institutionalisierte Kulturkapital wird, im Gegensatz hierzu, in methodisch-didaktischer Weise im Rahmen kanonisierter Wissensbestände in den Bildungsinstitutionen vermittelt und erworben. Diese Institutionen verfügen über die Definitionsmacht hinsichtlich legitimen Wissens und vergeben Patente, die als allgemeine Kompetenzanmutung dienen und den Zugang zu spezifischen Positionen und Laufbahnen ermöglichen. Jedoch wird im Bildungssystem nicht nur Wissen vermittelt und abgefragt, sondern es wird zudem ein spezifischer Habitus erwartet und belohnt, der mit demjenigen der definitionsmächtigen Gruppen korrespondiert. Neben den individuellen Kompetenzen und Fähigkeiten ist nach Bourdieus Auffassung eine habituelle Passung zwischen Herkunftsmilieu und der "Kultur" der Bildungsinstitution oder des Studienfaches förderlich, um einen erfolgreichen Abschluss zu erzielen.

Soziale Ungleichheit im Bildungssystem wird also im bourdieuschen Theorieansatz einerseits durch die differenzielle Verteilung kultureller Praxen in verschiedenen Segmenten der Sozialstruktur, andererseits, auf symbolischer Ebene, durch die Definitionsmacht der herrschenden Kapitalfraktionen hinsichtlich der legitimen Kultur und Wissensbestände in den Bildungsinstitutionen induziert. Mittels dieser Verknüpfung von Macht- und Wissenssoziologie gelingt es Bourdieu, jenseits objektivistischer und reifizierender "Messung" von individuellen Fähigkeiten, die Mechanismen aufzuzeigen, mit denen sich soziale Reproduktion durch Bildung und kulturelles Kapital vollzieht und wie diese mit der Verteilung von symbolischer Macht in Lebensstil und Habitus verknüpft sind.

Wenngleich dieser Versuch der Erklärung sozialer Ungleichheit im Bildungssystem auf einer intuitiven und alltagsweltlichen Ebene einzuleuchten vermag, stößt eine empirische Überprüfung auf Schwierigkeiten, die im Folgenden kurz erläutert werden sollen. Bourdieu selbst stützte sich in seinen statistischen Analysen vor allem auf bivariate Zusammenhangsanalysen zwischen Klassenzugehörigkeit und Bildungsbeteiligung einerseits sowie der kulturellen bzw. lebensstilbezogenen Praxis der verschiedenen Klassenfraktionen andererseits. Ein empirischer Beweis im engeren Sinne (der übrigens nach der reflexiv-soziologischen Methode schnell als kleinbürgerliche Variante innerhalb der soziologischen Wissensbestände desavouiert werden könnte) kann auf dieser Grundlage nicht erbracht werden. Hierzu wäre vielmehr ein theoretisch ausgearbeitetes und begründetes Messkonzept des Konstrukts kulturelles Kapital notwendig und die zugrundeliegende Kausalkette müsste möglichst umfassend modelliert werden. In der bisherigen Forschung ist dies nur teilweise gelungen, wie in Kapitel 2 erläutert werden soll.

2 Bisherige Forschung

Bei einer empirischen Überprüfung der bourdieuschen Reproduktionsthese werden unterschiedliche Fragen aufgeworfen, die im Folgenden systematisch auf der Grundlage des in Kapitel 1 Dargestellten behandelt werden sollen.

Erstens erhebt sich die Frage nach der Operationalisierung und Messung kulturellen Kapitals. Bei Bourdieu (1982, S. 800ff.) wird vor allem kulturelle Praxis (Literatur, Medienkonsum, Musik, Besuch kultureller Veranstaltungen, ästhetische Präferenzen) und das Wissen in kulturellen Bereichen (Komponisten, Regisseure, Musikstücke) erhoben. Insbesondere niederländische Autoren (De Graaf 1986; De Graaf et al. 2000; De Graaf & De Graaf 2002) konnten nachweisen, dass nicht primär Kulturkonsum einen Einfluss auf die schulische Leistung und die Art des Schulabschlusses hat, sondern eher das Leseklima in der Familie und das Leseverhalten des befragten Kindes. Begründbar ist dieser Befund damit, dass im jesuitisch geprägten französischen Schulsystem der 60er-Jahre des letzten Jahrhunderts eine große Nähe zwischen der legitimen Hochkultur einerseits und dem schulischen Kanon andererseits vorhanden war. In einem modernisierten und demokratisierten Schulsystem wie dem niederländischen ist es jedoch weniger die Nähe zwischen der Statuskultur der Oberschicht und den schulischen Anforderungen als vielmehr das methodische Niveau der Lesefähigkeit im Sinne einer abstrakten kognitiven Kompetenz (vgl. hierzu auch das Konzept der reading literacy in der PISA-Studie), das über schulischen Erfolg entscheidet. Zudem kommen verschiedene Studien (Sullivan 2001, De Graaf & De Graaf 2002) zu dem Ergebnis, dass auch anspruchsvoller Fernsehkonsum und der Konsum von populären Lesestoffen einen Einfluss auf die schulische Performanz ausüben. Im Gegensatz hierzu stellten Rössel & Beckert-Zieglschmid (2002) bei einem Vergleich der von Schulze (1992) entwickelten alltagsästhetischen Schemata der Hoch- und der Spannungskultur fest, dass das Hochkulturschema die Schulleistung der Kinder dominiert. Trotz einer gewissen Heterogenität der Befunde spricht inzwischen einiges dafür, dass das bei Bourdieu noch am legitimen Kulturkonsum enggeführte Konzept des kulturellen Kapitals infolge der Zunahme und der Demokratisierung kultureller Stile zu erweitern und die spezifischen Wirkungsrichtungen unterschiedlicher Dimensionen kulturellen Kapitals auf Maße des schulischen und beruflichen Erfolgs zu untersuchen ist.

Zweitens stellt sich die Frage nach der differenziellen Wirkung kulturellen Kapitals für unterschiedliche Subgruppen. So untersuchte bereits Di Maggio (1982) zwei kontrastierende Hypothesen über die Wirkung kulturellen Kapitals, nämlich die Reproduktionshypothese, die sich an die bourdieuschen Annahmen anlehnt, und die Mobilitätshypothese, die davon ausgeht, das kulturelles Kapital nicht primär von den oberen Statusgruppen zur sozialen Schließung benutzt wird, sondern den unteren Schichten als Mittel zum sozialen Aufstieg dient. Di Maggio konnte einen Reproduktionseffekt nur für die weibliche Subgruppe nachweisen. Die oben erwähnten niederländischen Autoren kommen in zwei unterschiedlichen Analysen zu dem Ergebnis, dass kulturelles Kapital, gemessen als Lesepraxis, seine Wirkung vor allem im Sinne der Mobilitätshypothese entfaltet (De Graaf et al. 2000, 2002). Aufgrund dieser Ergebnisse ist zu fragen, ob es bei der Wirkung kulturellen Kapitals

geschlechts- und schichtspezifisch unterschiedliche Effekte gibt und wenn ja, welche.

Drittens ist zu klären, mit welchem Erhebungsdesign Reproduktionseffekte realistisch erfasst werden können und welche statistische Modellierung hierfür angemessen ist. Bei einigen Studien ist die Messung kulturellen Kapitals problematisch, weil von Elternangaben auf das Kapital der Kinder geschlossen wird (De Graaf 1986) bzw. umgekehrt (Sullivan 2001, Rössel & Beckert-Zieglschmid 2002). Angemessener ist es stattdessen, unabhängige Informationen von Eltern und Kindern zu erheben, was jedoch eine relativ aufwendige Feldarbeit voraussetzt. Fast alle bisherigen Studien beruhen auf Querschnittsdaten, was insofern Probleme aufwirft, als die Wirkung kulturellen Kapitals nur für einen Zeitpunkt im Lebenslauf (abhängige Variablen sind zumeist eine oder mehrere Schulnoten oder der Schulabschluss) modelliert werden kann. Von Bedeutung ist jedoch die Frage, ob, wie Bourdieu dies unterstellt, ein Hysteresis-Effekt des Habitus nachgewiesen werden kann, d. h. dass das inkorporierte kulturelle Kapital auch zu unterschiedlichen Zeitpunkten im Lebenslauf wirksam ist. Für die statistische Modellierung wäre es wünschenswert, die vollständige Kausalkette, von der Erfassung kulturellen Kapitals im Elternhaus über dessen Transmission auf das Kind bis hin zu dem Effekt auf unterschiedliche Maße des schulischen oder beruflichen Erfolgs simultan darzustellen.

3 Fragestellung

In der PISA-Studie 2000 gingen Watermann & Baumert (2006, S. 66ff.) von einem Struktur- und Prozessmodell aus, in dem der sozio-ökonomische und Bildungshintergrund der Familie über konsumtives Verhalten sowie kulturelle und soziale Praxis auf den Kompetenzerwerb wirkt. Zentrale Annahme in diesem Mediationsmodell ist, ausgehend vom bourdieuschen Konzept kulturellen Kapitals und Colemans (1988) Sozialkapitaltheorie, dass unterschiedliche Schichtungsmilieus über eine differenzielle Ausstattung dieser Kapitalarten verfügen, die wiederum den Kompetenzerwerb beeinflussen. Im Folgenden soll ein Ausschnitt aus diesem Modell mit Hilfe der Latent-Class-Analyse (LCA) untersucht werden. Bereits die Befunde von Watermann & Baumert (2006, S. 82ff.) im Rahmen eines Strukturgleichungsmodells, das in allen 32 beteiligten Länder getestet wurde, zeigten überall bis auf Dänemark, die Niederlande und Portugal einen dominanten Effekt des kulturellen Kapitals auf die Lesefähigkeit. Somit ist, auch im internationalen Vergleich, mit dem kulturellen Kapital einer Familie die zentrale Transmissionskategorie für die Kompetenzentwicklung identifiziert.

Die bisherigen zahlreichen Analysen zum Effekt kulturellen Kapitals auf die Entwicklung sozialer Ungleichheit im Bildungssystem (Aschaffenburg & Maas 1997, De Graaf 1986, De Graaf et al. 2000, DiMaggio 1982, DiMaggio & Mohr 1985, Dumais 2002, Eitle & Eitle 2002, Farkas et al. 1990, Ganzeboom et al. 1990, Georg 2004, 2005, 2006; Kalmijn & Kraaykamp 1996, Robinson & Garnier 1985, Roscigno

& Ainsworth-Darnell 1999, Sullivan 2001) sind entweder auf manifeste Variablen in Regressionsmodellen oder auf kontinuierliche latente Variablen (Faktoren) beschränkt. Diese Analysen auf Variablenebene vermitteln zwar wichtige Einblicke in den Zusammenhang von kulturellem Kapital und sozialer Ungleichheit im Bildungssystem, jedoch eröffnen personenzentrierte Verfahren wie die LCA einen realistischeren Zugang zur Prozesslogik familialen Sozialisationsverhaltens: Einerseits gehen nicht nur bivariate Zusammenhangsmaße wie Kovarianzen in die Analyse ein, sondern alle Interaktionen auch höchster Ordnung; andererseits werden auch nicht-lineare Zusammenhänge berücksichtigt. Meines Wissens gibt es, bis auf Ehmke (2008), die sich in der PISA-Studie 2006 mit lernfördernden Prozessen in der Familie und naturwissenschaftsbezogenen Einstellungen beschäftigte, noch keine Anwendungen der LCA in diesem Bereich. Ausgehend von den Befunden von Watermann & Baumert (2006) soll es im Folgenden darum gehen, mittels einer LCA den mediierenden Effekt von kulturellem Kapital zwischen vertikalen Merkmalen sozialer Ungleichheit (sozio-ökonomischer und Bildungsstatus) und dem Kompetenzerwerb bei der Lesefähigkeit vergleichend zwischen Finnland und Deutschland mit den PISA 2000-Daten zu untersuchen. Dieser Vergleich ist insbesondere deshalb interessant, weil Finnland das beste Ergebnis erzielte, während Deutschland weit abgeschlagen nur auf dem 21. Platz lag.

4 Exkurs: Das finnische Schulsystem

Das finnische Schulsystem unterscheidet sich in einigen Punkten deutlich vom deutschen. Die Pflichtschule beginnt im Alter von 7 Jahren und erstreckt sich im Rahmen einer Integrierten Gesamtschule bis zur neunten Klasse (Prucha 2001). In einer Reihe von Reformschritten wurden in den 80er- und 90er-Jahren des 20. Jahrhunderts die wichtigsten Kompetenzen zunehmend von zentralen Institutionen auf die Schulebene verlagert (www.ibe.unesco.org – country dossiers). So wird zwar die minimale Zahl der zu unterrichtenden Fächerstunden vorgegeben, jedoch kann die Gemeinde bzw. die Schule selbst entscheiden, wie viele Stunden eines spezifischen Faches in einem Jahrgang erteilt werden.

Die obere Sekundarbildung beginnt nach der neunten Klasse und bietet zwei Wege an:

- Eine dreijährige allgemeinbildende Schule des gymnasialen Typs (Lukio), die zu der Studienberechtigung an einer Universität oder einem Polytechnikum (entspricht der deutschen Fachhochschule) führt. Etwa 60% eines Jahrgangs besuchen diesen Schultyp.
- Den Besuch einer sekundären Berufsschule, die entweder zwei- bis dreijährig (berufliche Grundbildung) oder drei- bis vierjährig (spezialisierte Fachausbildung) ist, wobei letztere zu einem Bachelor-Abschluss führt.

Sozio-kulturelle Benachteiligungen und regionale Unterschiede werden relativ effizient ausgeglichen. Bereits in der Reading Literacy Study 1990/91 wurden Unter-

schiede im Testergebnis einzelner Schulen in verschiedenen Ländern untersucht. Ein Homogentitätskoeffizient, der bei hoher Homogenität der Testergebnisse aller Schulen einen niedrigen und bei großen Unterschieden einen hohen Wert annimmt (Prucha 2001, S. 5), lag für Finnland bei einem Wert von .03 und für Deutschland bei .49.

Generell zeigte die PISA 2000-Studie, dass Länder mit Gesamtschulsystemen einen geringeren Anteil von Schülern auf der untersten Leistungsstufe entließen, während dieser Anteil in gestuften Schulsystemen wie dem deutschen verhältnismäßig hoch war (Finnland: 2%: Deutschland: 10%; ebd, S. 6).

Für das finnische Bildungssystem ist weiterhin typisch, dass der Lehrerbildung eine hohe Bedeutung beigemessen wird und alle Lehrer, auch auf der Vorschulebene, über einen Universitätsabschluss verfügen.

5 Daten und Maße

Grundlage der folgenden Analysen bildet der Datensatz der PISA 2000-Studie (Deutsches PISA-Konsortium 2001). Die Stichprobe umfasst in Deutschland 4284 und in Finnland 4377 Schüler im Alter von 15 Jahren, wobei es sich um eine mehrfach geschichtete Stichprobe handelt. Folgende Variablen gingen für die Operationalisierung kulturellen Kapitals in die Analyse ein:

- die Häufigkeit des Besuchs von Museen oder Kunstgalerien (von 1 = bis zweimal jährlich bis 4 = mehr als viermal jährlich),
- die Häufigkeit des Besuchs von Oper, Ballett oder klassischen Konzerten (Kategorien siehe oben),
- die Häufigkeit von Theaterbesuchen (Kategorien siehe oben),
- die Häufigkeit von Gesprächen über politische oder soziale Themen (von 1 = so gut wie nie bis 5 = mehrmals wöchentlich),
- die Häufigkeit des Hörens klassischer Musik mit den Eltern (Kategorien siehe oben),
- der Besitz von klassischer Literatur (0 = nein, 1 = ja),
- der Besitz von Gedichtbänden (0 = nein, 1 = ja),
- der Besitz von Kunstwerken (0 = nein, 1 = ja),
- die Anzahl der Bücher im Haushalt (0 = 0 Bücher bis 7 = 500 Bücher und mehr).

Für die Beschreibung des ungleichheitsbezogenen Hintergrunds der Familien wurden folgende Maße verwendet:

Der Berufsstatus einer Familie wurde mit Hilfe des höchsten International Socio Economic Index (ISEI) (Ganzeboom et al. 1992) operationalisiert, der auf der 1968 vom Internationalen Arbeitsamt in Genf entwickelten International Standard Classification of Occuoation (ISCO) fußt. Der ISEI geht von der Annahme aus, dass Schul- und Berufsbildung über Berufe in Einkommen umgesetzt werden, und umfasst in der PISA 2000-Studie eine Spannweite von 74 Skalenpunkten (von 16 bis 90).

Die Bildung des Vaters wurde mit dem höchsten erreichten Schulabschluss nach Maßgabe der International Standard Classification of Education (ISCED, 5 Skalenstufen) erfasst.

Als abhängige Variable diente die auf einen Mittelwert von 500 standardisierte Testleistung im Lesefähigkeitstest von PISA 2000.

6 Methode

Das Grundmodell der LCA wurde von Lazarsfeld & Henry (1968) entwickelt und hat für dichotome Items die Form

$$\text{prob}(x_i) = \sum_g \pi_g \, \pi_{ig} \,, \tag{1}$$

d. h. die Antwortwahrscheinlichkeit des Items x ($\text{prob}(x_i)$) wird zerlegt in die relative Größe (Klassenwahrscheinlichkeit) der Klasse g (π_g) und die Antwortwahrscheinlichkeit (π_{ig}) des Items in dieser Klasse. Inhaltlich bedeutet dies, dass alle Personen in einer Klasse über die gleiche Antwortwahrscheinlichkeit bei allen Items verfügen. Das Ziel einer LCA besteht also darin, die Personen in latente Klassen einzuteilen, die dadurch definiert sind, dass sie innerhalb der Klassen dieselben und zwischen den Klassen verschiedene Antwortwahrscheinlichkeiten haben. Im Unterschied zu deterministischen Clusteranalysen handelt es sich um keine manifeste Klassifikation, sondern um eine latente, da die Personen jeder Klasse mit einer bestimmten Wahrscheinlichkeit angehören. Sollen Personen aufgrund einer LCA tatsächlich einer Klasse zugeteilt werden, so ordnet man sie der Klasse zu, der sie mit der größten Wahrscheinlichkeit angehören (Modalwert). Die mittlere Höhe dieser Modalwerte kann gleichzeitig als Indikator für die Eindeutigkeit des Analyseergebnisses gewertet werden. Bei der LCA handelt es sich um ein Modell, in dessen Rahmen multivariate Abhängigkeiten mit einer mehrdimensionalen Kontingenztabelle erklärt werden sollen, wobei nicht, wie etwa bei Strukturgleichungsmodellen, von bivariaten Assoziationen zwischen den manifesten Variablen ausgegangen wird (Kovarianzen bzw. Korrelationen), sondern von den vollständigen Häufigkeiten der Kontingenztafel, in denen noch die Wechselwirkungen höchster Ordnung zwischen den manifesten Variablen enthalten sind. Dieser theoretische Vorteil wird jedoch mit praktischen Problemen erkauft, die z. B. in der Anwendbarkeit auf große Itemmengen liegen. Als Resultat der Berücksichtigung aller Interaktionen zwischen den analysierten Items sind diese bei gegebener Klassenzugehörigkeit voneinander unabhängig (man spricht auch von lokaler stochastischer Unabhängigkeit). Ein Vorteil von LCA-Modellen besteht darin, dass für die Auswahl der Klassenanzahl unterschiedliche inferenzstatistische und informationstheoretische Kriterien zur Verfügung stehen. In Simulationsstudien hat sich das Bayesian Information Criterion (BIC) als bestes informationstheoretisches Maß für die Bestimmung der korrekten Klassenanzahl erwiesen (Nylund et al. 2006, S. 31; Roeder & Wasserman 1997):

$$BIC = -2\ln(L) + \ln(n)n(P), \quad (2)$$

wobei ln(L) = natürlicher Logarithmus der marginalen Likelihood der Datenmatrix, ln(n) = natürlicher Logarithmus der Stichprobengröße und n(P) = Anzahl der unanhängigen Modellparameter (Bozdogan 1987) sind.

Der kleinste BIC-Wert zeigt die optimale Klassenanzahl an. Für ordinale Variablen wurde dieser Ansatz von Jürgen Rost in einem Schwellenwert-Modell weiterentwickelt (vgl. z. B. Rost 1988) und im Programm LACORD (Rost 1990) umgesetzt.

7 Ergebnisse

Schon seit einiger Zeit ist es durch den Einsatz von LCA-Programmen möglich, den Einfluss von Kovariaten auf latente Klassen zu modellieren (vgl. etwa Vermunt & Magidson 2005, Collins & Lanza 2010). Bisher erlaubt es jedoch meines Wissens nur das Programm Mplus (Muthen & Muthen 1998-2010), gleichzeitig latente Klassen auch als Prädiktor für spezifische Outcome-Variablen zu verwenden (distal outcomes), sie also als Mediationsvariablen zu behandeln. Daher wurde das folgende Modell mit Mplus, Version 6.1 geschätzt:

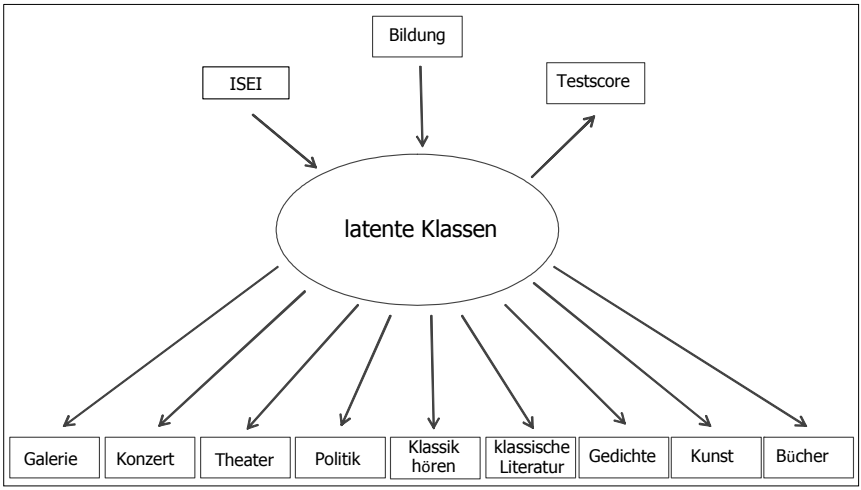


Abbildung 1: LCA-Modell für kulturelles Kapital

Da einige Ausprägungen nur sehr gering besetzt waren und demzufolge Schwierigkeiten bei der Schätzung entstanden, wurden die Indikatoren des Modells so nahe wie möglich am Median dichotomisiert. Um den Einfluss kulturellen Kapitals in Deutschland und Finnland zu vergleichen, wurde ein gemeinsames Messmodell für beide Länder angepasst. Ursprünglich sollten alle Lösungen von 1 bis 5 latenten Klassen berechnet werden, jedoch traten bei der Lösung mit 4 und 5 Klassen

Schätzprobleme auf, so dass die Anpassungsmaße nur für 1 bis 3 latente Klassen geschätzt werden konnten. Auch war die Modellierung von landesspezifischen Regressionen der LCA-Klassen auf den Berufsstaus und den Bildungsabschluss nicht möglich, weil die Informationsmatrix für die entsprechenden Koeffizienten singulär war. Um lokale Maxima bei der Likelihood-Schätzung zu vermeiden, wurden 500 Zufallsstartwerte verwendet.

Anzahl Klassen	BIC	ln(L)	n(P)
1	347726.492	-173774.808	19
2	197135.641	-98440.857	28
3	195864.059	-97779.751	41

Tabelle 1: BIC für 1 bis 3 latente Klassen

Nach Maßgabe des BIC repräsentiert die 3-Klassen-Lösung die Daten am besten. Tabelle 2 gibt die Responsewahrscheinlichkeiten dieser Lösung wieder.

Variable		Klasse 1	Klasse 2	Klasse 3
Kunstgalerien		.040	.544	.064
Klassisches Konzert		.064	.647	.096
Theater		.326	.819	.427
Diskussion Politik		.259	.787	.485
Klassik hören		.095	.483	.174
Besitz: klass. Literatur		.116	.916	.822
Besitz: Gedichtbände		.323	.945	.855
Besitz: Kunstwerke		.552	.884	.853
Anzahl Bücher >100		.211	.896	.675
Klassengröße	FI	.445	.125	.429
	DE	.460	.224	.316

Tabelle 2: Responsewahrscheinlichkeiten für die 3-Klassen-Lösung. FI = Finnland; DE = Deutschland

Klasse 1 ist in Finnland (.445) und in Deutschland (.460) die größte. Sowohl im Bereich der kulturellen Aktivitäten (Besuch von Kunstgalerien: .040, Besuch klassischer Konzerte: .064, Theaterbesuch: .326) als auch bei gemeinsamen familiären Aktivitäten (Diskussionen über Politik: .259, gemeinsames Hören klassischer Musik: .095) und bezüglich des Besitzes von kulturellen Gütern (klassische Literatur: .116, Gedichtbände: .323, Kunstwerke: .552, Bücherbesitz: .211) präsentiert sich diese Klasse als diejenige mit den geringsten kulturellen Ressourcen.

Die zweite latente Klasse, die in Finnland 13% und in Deutschland 22% der Stichprobe repräsentiert, besteht aus Familien mit durchgängig hohem kulturellen Ka-

pital: Dies fällt insbesondere beim Besuch von Museen und Kunstgalerien (.544) sowie klassischen Konzerten (.647) und Theatern (.819) auf. Bedeutsam sind ebenso das Diskutieren über Politik (.787) und das Hören klassischer Musik gemeinsam mit den Eltern (.483). Diese Gruppe ist auch mit objektiviertem kulturellen Kapital mit Abstand am besten ausgestattet, denn nahezu alle Familien besitzen klassische Literatur (.916), Gedichtbände (.945), Kunstwerke (.884) und mehr als 100 Bücher (.896)

Klasse 3 schließlich, die in Finnland 43% und in Deutschland 32% der Befragten umfasst, ist zwischen den beiden ersten Klassen positioniert. Während sie im Bereich der kulturellen Aktivitäten und des gemeinsamen Hörens klassischer Musik näher bei der ersten Klasse liegt (Kunstgalerien: .064, klassische Konzerte: .096, Theater: .427, Klassik hören: .174), ist sie beim objektivierten Kulturkapital näher an der zweiten Klasse mit hohen kulturellen Ressourcen positioniert (klassische Literatur: .822, Gedichtbände: .855, Kunstwerke: .853, Bücherbesitz: .675).

Während die Klasse 1 mit niedrigem kulturellen Kapital in beiden Ländern mit 45% gleich groß ist, repräsentiert die kulturaffine Klasse 2 in Deutschland fast doppelt so viele Schüler wie in Finnland. Demgegenüber verfügen in Finnland deutlich mehr Befragte über mittlere kulturelle Ressourcen als in Deutschland.

Da das kulturelle Kapital in dieser Analyse als intervenierende Variable behandelt wird, sind zwei weitere Fragen von Interesse:

- Wie hängt die Verfügung über kulturelles Kapital von Merkmalen sozialer Ungleichheit wie Bildungs- und Berufsstatus der Familie ab? Wie bereits weiter oben beschrieben, ließ sich wegen einer singulären Informationsmatrix hier nur ein gemeinsames Modell für beide Länder schätzen
- In welchem Umfang beeinflusst das kulturelle Kapital der Familie die Lesefähigkeit in beiden Ländern?

Variable	Klasse 1 vs. Klasse 3	Klasse 2 vs. Klasse 3
Bildung Vater	.770	1.408
ISEI	.096	1.026

Signifikanzniveau für alle Koeffizienten <.001

Tabelle 3: Odds Ratios für die Regression der latenten Klassen auf den höchsten ISEI in der Familie und den Schulabschluss des Vaters

Pro Einheit der Bildungsvariablen sinkt die Wahrscheinlichkeit, zu Klasse 1 (niedriges kulturelles Kapital) statt zu Klasse 3 (mittleres kulturelles Kapital) zu gehören, um 23%. Umgekehrt steigt die Wahrscheinlichkeit, zur bildungsaffinen Klasse 2 statt zu Klasse 3 zu gehören, um 41%. In eine ähnliche Richtung gehen die Koeffizienten beim Berufsstatus (Klasse 1 vs. Klasse 3: -3.4%; Klasse 2 vs. Klasse 3: +2.6%). Bei der Interpretation der Effekte des ISEI ist zu beachten, dass dieser nicht, wie die Bildungsvariable, 5 Skalenpunkte umfasst, sondern eine Spannweite von 74 Einheiten (von 16 bis 90). Zusammenfassend kann festgestellt werden,

dass zwischen dem Bildungs- und Berufsstatus (und zwar unabhängig voneinander) einerseits und dem Umfang kulturellen Kapitals ein Zusammenhang von beachtlicher Stärke besteht.

Wie nun sind die Unterschiede in der Leseleistung, die sich bei gleichem Umfang kulturellen Kapitals in beiden Ländern bilden?

LCA-Klasse	Finnland	Deutschland
1. Niedriges Kulturkapital	520.500 (-25.5)	462.531 (-21.47)
2. Hohes Kulturkapital	608.621 (62.62)	575.968 (91.97)
3. Mittleres Kulturkapital	565.792 (19.79)	532.798 (48.80)
M	546	484
RD	88.121	113.458

Signifikanzniveau für alle Koeffizienten <.001

Tabelle 4: Leseleistung nach Klassenzugehörigkeit und Land (Abweichung vom Mittelwert in Klammern). M = Mittelwert, RD = Spannweite der Differenz zwischen Klassen

Zunächst fällt auf, dass die Abweichungen vom Landesmittelwert bei der Klasse mit dem niedrigsten kulturellen Kapital in beiden Ländern etwa gleich groß sind. Dies ist nachvollziehbar, weil die hypothetischen Annahmen erst ab einem höheren Niveau einen Effekt auf die Kompetenzentwicklung unterstellen und ein niedriges Niveau nur als Kontrastfolie hierzu dient. Anders stellt sich der Zusammenhang jedoch bei den Klassen mit mittleren und hohen kulturellen Ressourcen dar: Hier sind die Abstände zum Landesmittelwert in Finnland etwa 30 Testpunkte geringer als in Deutschland. Zudem ist die Spannweite in den Testscores, die in Deutschland durch die Differenzen im kulturellem Kapital des Elternhauses induziert wird (113), um 25 Einheiten (oder eine viertel Standardabeichung) höher als in Finnland. Beide Befunde sprechen dafür, dass im finnischen Schulsystem Unterschiede im kulturellen Kapital des Elternhauses weniger stark auf die Kompetenzentwicklung bei der Lesefähigkeit zurückwirken als im gestuften deutschen Schulsystem.

8 Diskussion

Ausgangspunkt dieses Beitrags waren Befunde, die trotz deutlicher Mobilitätseffekte (Breen et al. 2009, 2010) weiterhin ein stärkeres Fortbestehen sozialer Ungleichheit im deutschen Bildungssystem im Vergleich zu anderen, insbesondere skandinavischen Ländern belegen. So konnte in der PISA 2000-E-Studie nachgewiesen werden, dass die Chance auf einen Gymnasialbesuch bei Kontrolle der kognitiven Grundfähigkeiten und der Lesefähigkeit für einen Schüler aus der oberen Dienstklasse immer noch dreimal so hoch war wie für das Kind eines Facharbeiters (Baumert & Schümer 2001, S. 323 ff.). Dafür gibt es zwei Erklärungsansätze, nämlich einerseits die Rational-Choice-Theorie, in deren Zentrum eine Abwägung von Bildungskosten und Erfolgswahrscheinlichkeit für die Erzielung von Bildungsrendi-

ten steht, und andererseits der konflikttheoretische Ansatz von Pierre Bourdieu, der vor allem eine differenzielle Verteilung von kulturellem Kapital in den verschiedenen Klassenfraktionen fokussiert. Ein Überblick über die bisherige Forschung erbrachte Hinweise auf einige Desiderata: Dies war einerseits die teilweise unvollständige Modellierung der unterstellten Kausalkette; andererseits wurde die Frage aufgeworfen, ob, wie Paul DiMaggio (1982) dies formulierte, kulturelles Kapital eher von oberen Klassen monopolisiert wird und somit deren Statusreproduktion dient (Reproduktionshypothese), oder ob es in einer auch kulturell demokratisierten Gesellschaft auch als Mobilitätsmittel für untere Klassen eingesetzt werden kann (Mobilitätshypothese). Vor diesem Hintergrund zielte die vorliegende Analyse auf zweierlei Fragestellungen:

- Einerseits sollte, vorwiegend unter methodologischen Gesichtspunkten, in Abgrenzung von den bisherigen variablenzentrierten Verfahren (Faktorenanalyse bzw. Strukturgleichungsmodelle) mit Indikatoren für kulturelles Kapital eine auf Personengruppen fußende Typologie konstruiert werden, auf deren Grundlage die sozialisationsbezogenen Prozesse realistischer analysiert werden können.
- Andererseits stellt sich die inhaltliche Frage, in welchem Umfang diese auf der Grundlage einer LCA entwickelte Typologie als familiales Prozessmerkmal in der Lage ist, die black box zwischen sozialstrukturellen Merkmalen der Situiertheit im Gefüge sozialer Ungleichheit (Bildungs- und Berufsstatus der Familie) und den lageabhängig differenziellen Testergebnissen bei der Lesefähigkeit zu erklären und wie sich diese Prozessmerkmale in einem relativ egalitären Land wie Finnland im Vergleich zur schärfer kontrastierten Ungleichheit im deutschen Bildungssystem darstellen.

Die LCA identifizierte für beide Länder drei trennscharfe Gruppen, die in niedrigem, mittlerem und hohem Maß über kulturelles Kapital verfügen. Weiterhin ließ sich eine deutliche Abhängigkeit dieser Gruppen vom Bildungs- und Berufsstatus der Familie nachweisen. Da in dieser LCA, im Gegensatz zu den meisten anderen, die Klassen als intervenierende Variablen zwischen Merkmalen der sozialen Ungleichheit und der Lesefähigkeit dienten, wurde in einem zweiten Analyseschritt ihre Auswirkung auf diese untersucht.

Als Ergebnis konnte festgehalten werden, dass in einem integrierten Schulsystem wie dem finnischen, in dem die Schüler während der ersten neun Schuljahre gemeinsam unterrichtet werden, die Effekte des kulturellen Hintergrunds des Elternhauses deutlich weniger prononciert sind als im gestuften deutschen Bildungssystem. Sowohl bei mittlerem als auch bei hohem kulturellem Ressourcenniveau des Elternhauses waren die Abstände zum Landesmittelwert in Finnland um etwa 30 Testpunkte (oder fast eine Drittel Standardabweichung) geringer als in Deutschland. Die durch Unterschiede in kulturellem Kapital abgedeckte Spannweite war in Finnland ebenfalls um 25 Testpunkte geringer.

Die vorliegende Analyse hat gezeigt, dass die Anwendung eines personenzentrierten Ansatzes bei der Analyse sozialer Ungleichheit im Bildungssystem über ein beachtliches Innovationspotenzial verfügt. Nur so war es möglich, die Auswirkungen

der gleichen familialen Ausgangslage in unterschiedlichen Landeskontexten zu untersuchen und deren differenzielle Effekte zu beleuchten. Wünschenswert für zukünftige Analysen wäre die Verwendung von Längsschnittsdaten, mit deren Hilfe es dann auch möglich wäre, mittels einer Latent Transition Analysis (Muthen & Muthen 1998-2010, S. 225) Wanderungsbewegungen zwischen verschiedenen latenten Klassen zu untersuchen.

Latent-Class-Modelle von Spracheinstellungen in der berliner Sprachgemeinschaft nach 1989

Irena Regener

1 Sozialpsychologische und soziolinguistische Grundlagen

Die Frage einer gesamtdeutschen bzw. der ost-westdeutschen Identität(en) oder auch Mentalität(en) hatte im Gefolge des Anschlusses der DDR an die BRD eine neue Dimension erhalten. In der soziologischen und sozialpsychologischen Forschung zeichnete sich der Trend ab, nicht grundlegende Ost-West-Unterschiede zu statuieren, sondern eher die Gemeinsamkeiten von Ost- und Westdeutschen bzw. die gemeinsame Basis für nach 1989/90 einsetzende, zunehmend divergente Entwicklungen in Mentalitätsstrukturen und Einstellungsmustern in Ost und West zu akzentuieren (z. B. Bauer 1991, Braun 1993, Piontkowski et al. 1997, Meulemann 1998). Diese Betrachtungsweise kann auch aus der soziolinguistischen Perspektive gestärkt werden.

1.1 Identität und Einstellung

Die zur Ausbildung einer sozialen Identität nötigen Identifikationsprozesse sind Grundmechanismen der Persönlichkeitsbildung in der Soziogenese, die durch multiple Gruppenzugehörigkeiten des Individuums gesteuert werden. Identifikation erfolgt wesentlich über Einstellungsbildung und -veränderung zu identitätsstiftenden Wertorientierungen und unter Mitwirkung sozial vermittelter Vorurteile, Stereotype und Typisierungen. Unter Einstellungen wie auch unter Wertorientierungen gibt es jedoch sowohl mehr längerfristig wirksame als auch mehr situationsabhängige (Pollack 1998). Erstere sind die im Laufe der Sozialisation erworbenen, die persistent sind und aus diesen beiden Gründen relativ änderungsresistent. Letztere sind soziostrukturell bedingt und ändern sich mit der Modifikation der entsprechenden soziostrukturellen Merkmale.

Unter Einstellungen kommt den Spracheinstellungen eine zentrale Rolle zu, weil sie als soziale Einstellungen Identifikationsmuster und Identitätszuweisungen indizieren: Über Spracheinstellungen ordnet sich das Individuum selbst in das soziale Gefüge ein, bewertet aus dieser Perspektive die anderen (Giles et al. 1987) und nimmt auf diese Weise Selbst- und Fremdkategorisierungen vor. Spracheinstellungen dienen folglich der sozialen Orientierung und spiegeln letztlich Haltungen ge-

genüber Sprechern und deren kulturellen Werten wider. Soziale Konnotationen gegenüber Sprechern werden auf deren Sprache projiziert.

Aus der soziolinguistischen Perspektive gründen sich Spracheinstellungen auf subjektive Wahrnehmungen, die bezogen sind auf die innersprachlichen, d. h. die Sprachvariationen *innerhalb* einer Sprache – *wer* spricht *wie* mit *wem* unter *welchen Kontextbedingungen*, und *wie erfolgreich* ist er dabei.

Da das sprachlich handelnde, Sprache produzierende und rezipierende Individuum jedoch in seinen Erwartungen einerseits und in seinen Verhaltensweisen andererseits geprägt ist durch die sozialen Muster seines Lebenskontextes, sind subjektive Spracheinstellungen letztlich Funktionen der Zugehörigkeit der Individuen zu ihrer jeweiligen sozialen Gruppe. Es ist die soziale Gruppe, die das Individuum über einen – gruppenspezifisch differenzierten – Kriterienkatalog zur Bewertung der verschiedenen Varietäten[1] einer natürlichen Sprache verfügen lässt. Diese Bewertung erfolgt über die der Sprecher, die die jeweilige Varietät aktualisiert haben. Folglich werden Varietäten, deren Sprecher hohes Ansehen genießen, positiver beurteilt als Varietäten, deren Sprecher über ein geringeres Sozialprestige verfügen.

Der sprachliche Markt (Bourdieu 1982, 1993) erfordert es nun, dass die Angehörigen aller sozialen Gruppen eine Strategie verfolgen, die darauf gerichtet ist, Kapital an Legitimität und Prestige zur Selbstdefinition ihrer jeweiligen In-Group und zur Erhöhung ihres Wertes sowohl in der Selbst- als auch in der Fremdbetrachtung zu akkumulieren. Die Sprache wird als das entscheidende Medium zur Erlangung von Sozialprestige und damit potenziell auch eines höheren Marktwertes betrachtet. Jedes Individuum ist mit einer Sprachkompetenz bzw. mit sprachlichem Kapital ausgestattet, das dem seiner In-Group entspricht.

In den standardorientierten bürgerlichen Gesellschaften ist es in der Regel so, dass von Sprechern von der Mittelschicht an aufwärts erwartet wird, dass sie aus dem Spektrum von standardnahen vs. standardfernen Varietäten eine mindestens sehr standardnahe Varietät auswählen. Das hat letztlich dazu geführt, dass die Nonstandardvarietäten per se als weniger prestigeträchtig betrachtet werden und es damit auch definitiv sind. Konterkariert werden kann diese Tatsache jedoch in spezifischen sozialen Situationen, in denen eine andere Werteskala die allgemein bildungsgeprägte und damit standardorientierte dominiert, wenn z. B. das Bekunden von Solidarität, Integrität, Wohlwollen o. ä. die eigentlichen Sozialziele des Diskurses sind (Ros 1984). Spracheinstellungen und -bewertungen unterliegen folglich der kontextuellen Variation, die in jedem Falle extralinguistisch und damit sozial determiniert ist.

1 Varietäten sind Subsysteme einer natürlichen Sprache mit spezifischen eigenen Regelsystemen und Funktionsbereichen, die sich in Standard- und Nonstandardvarietäten untergliedern lassen. Zu den Nonstandardvarietäten gehören z. B. die regional begrenzten Dialekte – schwäbisch, thüringisch etc. – und Umgangssprachen mit größerer regionaler und funktionaler Reichweite als die Dialekte – brandenburgisch z. B. oder auch berlinisch. Die Standardvarietät ist Unterrichtsgegenstand und via Kodifizierung für ein in der Regel politisch definiertes Sprachgebiet mindestens für den öffentlichen Schriftverkehr verbindlich.

Das bedeutet letztlich, dass die Einstellung einer Varietät gegenüber und der Gebrauch einer Varietät zwar ineinander verschränkt sind, ohne jedoch unmittelbar aufeinander abbildbar zu sein. Die Sprecher können sich konvergent oder divergent verhalten, d. h. sie passen ihr Sprachverhalten an normative Kontexte an – oder auch nicht.

1.2 Ost-westdeutsche Identität

Die Transformationsforschung hat empirisch belegt, dass Identitätsfindungsprozesse im Deutschland der 1990er-Jahre von ostwestunterschiedener Spezifik waren. Die grundlegende These besteht darin, dass der westliche Teil recht ungebrochen die diesbezügliche bundesdeutsche Tradition kontinuierte, während sich der Ostteil genötigt sah, eine spezifisch ostdeutsche Identität als Akt der Selbstbehauptung gegenüber dem Westen überhaupt erst auszubilden (Pollack 1998). Diese ostdeutsche Identität hat sich aufgrund der zunehmenden Selbstkategorisierung der Ostdeutschen in den 90er-Jahren entwickelt. Weil Selbstkategorisierungen ein wesentliches Element nicht nur individueller, sondern auch sozialer Identitätsfindungsprozesse sind, wirken sie gleichzeitig sowohl konstitutiv für Gruppenbildungen als auch prägend für Beziehungen zwischen Gruppen, werden durch Vorurteile und Stereotypisierungen verstärkt und waren in diesem Falle verbunden mit z. T. zunehmender Distanzierung der Ostdeutschen von den Westdeutschen (wie die Befunde der in Meulemann [1998] vereinigten Studien belegen).

Da die gemeinsame Sprache zu den wesentlich identitätsstiftenden Werten sozialer Gruppen gehört, ist nach ihrer Funktion bzw. ihrer Stellung bei dieser neuen, spezifisch ostdeutschen Identitätsfindung zu fragen ebenso wie nach der Bewertung des als fremd empfundenen Sprachgebrauchs der Out-Group. Aus sozialpsychologischer Perspektive haben sich Piontkowski et al. (1997) mit den kommunikativen Manifestationen ost-westdeutscher Identität(en) beschäftigt. Das wesentliche Ergebnis ihrer Studie besteht darin, belegt zu haben, dass sich zwar hoch signifikante Unterschiede zwischen West- und Ostdeutschen in ihrem kommunikativen Verhalten nachweisen lassen, dass deren Annäherungs- und Abgrenzungsverhalten aber nicht 'ost'- bzw. 'west'spezifisch festzumachen ist, sondern dieselben Wirkungsmechanismen zeigt wie jede andere hoch akzessible soziale Kategorie. Das bedeutet m. E. unter anderem, dass in Sprachgebrauch und Spracheinstellungen nachweisbare Unterschiede zwischen Ost und West nicht sozialisatorisch, sondern soziostrukturell und somit situativ bedingt sind und deshalb eher weniger änderungsresistent.

Es gibt keine zweite Stadt in Deutschland, in der in 40 Jahren entstandene Ost-West-Unterschiede so unmittelbar aufeinanderprallten wie in Berlin. Auf die Sprache bezogen bedeutet das, dass die berliner Sprachgemeinschaft mehrfach inhomogen war und ist: Es gibt die Inhomogenität, die sie mit allen anderen Sprachgemeinschaften teilt und die letztlich auf die soziale Stratifikation rückführbar ist. Es gab und gibt aber außerdem, auf einer Metaebene, einen deutlichen Unter-

schied zwischen der ostberliner Sprachgemeinschaft auf der einen und der westberliner Sprachgemeinschaft auf der anderen Seite. Dies betrifft den Stellenwert der städtischen Nonstandardvarietät, also des Berlinischen, als Kommunikationsmittel, und, unmittelbar damit verbunden, Einstellungen der Stadtvarietät und ihren Sprechern gegenüber.

2 Ergebnisse von Vorgängeruntersuchungen

Die Auswertung einer von Dittmar et al. (1986) 1982 in Westberlin durchgeführten Fragebogenerhebung hatte ergeben, dass die berliner Sprachgemeinschaft in zwei signifikant unterschiedliche Kommunikationsgemeinschaften zerfalle, zu deren Beschreibung das Prestigemodell für den westlichen und das Solidaritätsmodell für den östlichen Teil entwickelt wurden (Dittmar & Schlobinski 1988). Im Westteil der Stadt war der Gebrauch des Berlinischen deutlich negativ markiert; er signalisierte die soziale Stratifikation und wurde nicht nur von erfolgs- und aufstiegsorientierten Sprechern vermieden. Mit den eher negativen Einstellungen der westlichen Kommunikationsgemeinschaft der Stadtvarietät gegenüber korrespondierten deren restringierte Verwendung und auch die geringere Dialekttiefe des Berlinischen. Umgekehrt sicherte der Gebrauch der Standardvarietät Sozialprestige und fungierte als soziales Abgrenzungsmittel.

Positiv besetzt dagegen war der Gebrauch des Berlinischen im Ostteil der Stadt, worauf bereits 1981 die DDR-Linguisten H. Schönfeld und M. Peine hingewiesen hatten ("mit hauptstädtischem Prestige ausgestattet"). Die Verwendung des Berlinischen war nicht schichtenspezifisch begrenzt geschweige denn sozial abgrenzend, sondern fungierte im Gegenteil als wesentlich identifikationsstiftendes Moment für die Hauptstädter in Abgrenzung vom "Rest" der Republik.

Ein Großteil dessen, was die positive Besetzung des Berlinischen im Ostteil der Stadt ausmachte bzw. hervorrief, existiert heute nicht mehr. Die Lebenssituation der Ostberliner hat sich mit dem Anschluss der DDR an die BRD grundlegend verändert, aber auch die der Westberliner ist nicht mehr die von 1982. Daraus ergab sich als naheliegende Ausgangshypothese für jegliche weiteren Untersuchungen, dass die östliche Kommunikationsgemeinschaft unter einen immensen Anpassungsdruck an das westliche Prestigemodell geraten sei. Gleichzeitig war aber auch zu fragen, ob Ausgleichs- oder Anpassungsprozesse ebenso in umgekehrter Richtung stattfinden, und überhaupt: Was eint, was trennt die Ostberliner und die Westberliner in ihrem Sprachbewusstsein, in ihren Spracheinstellungen und in ihrem Sprachverhalten?

3 Zum Untersuchungsdesign

1994 habe ich die erste Erhebungsstaffel einer Befragung mittels standardisierter Fragebogen durchgeführt, die sich mit genau diesen Fragestellungen auseinandersetzt. Weitere Querschnitte erfolgten 1996 und 1998.

Um eine partielle Vergleichbarkeit der Ergebnisse mit denen der '82er Studie zu gewährleisten, wurde für diese Befragung der damalige Fragebogen modifiziert und den veränderten Kommunikationsverhältnissen in der Stadt angepasst. Die gravierendsten Veränderungen ergaben sich daraus, dass es nun möglich war, vergleichende Untersuchungen in ganz Berlin durchzuführen und genauer zu fragen,

1. welcher Art die Unterschiede im Sprachverhalten und in den Spracheinstellungen von Ost- und Westberlinern sind,
2. wie Ost- und Westberliner den Sprachgebrauch in der jeweils anderen Stadthälfte wahrnehmen und beurteilen, und
3. ob und wenn ja, in welcher Weise sich, von den politischen Ereignissen von 1989 initiiert, Veränderungen sowohl im Verhalten als auch in den Einstellungen vollzogen haben und vollziehen.

Befragt wurden insgesamt 876 Berlinerinnen und Berliner in den Stadtbezirken Zehlendorf, Wedding, Pankow und Prenzlauer Berg im Alter von 25-45 Jahren. Diese Altersgruppe verfügt einerseits bereits über relativ gefestigte Einstellungsstrukturen und besitzt andererseits außerdem eine hohe Flexibilität und Anpassungsfähigkeit, die sie bewusst trainieren muss, um auf dem Arbeitsmarkt bestehen zu können, der in vielen Bereichen – und das war für die Ostberliner neu – zunehmend über den sprachlichen Markt reguliert wird. Somit war zu vermuten, dass Veränderungen in Einstellungen, so sie denn überhaupt schon fassbar waren, am ehesten in dieser Altersgruppe gefunden werden konnten. Die Auswahl der Informanten erfolgte nach dem Zufallsprinzip. Ihre Verteilung auf die verschiedenen Stadtbezirke ist Tabelle 1 zu entnehmen.

	Gesamt	Ost	Pb	Pa	West	We	Ze
1994	246	84	42	42	162	81	81
1996	318	146	91	55	172	74	98
1998	312	168	84	84	144	72	72
Gesamt	876	398	217	181	478	227	251

Tabelle 1: Gesamtstatistik der Befragten. Pb = Prenzlauer Berg, Pa = Pankow, We = Wedding, Ze = Zehlendorf

Im Westteil wurden wieder die Bezirke der 82er Befragung, Wedding und Zehlendorf, ausgewählt; im Ostteil Prenzlauer Berg und Pankow, um eine Vergleichbarkeit der Soziallage zu gewährleisten: Bei Pankow handelt es sich um ein (gut)bürgerliches, mit Zehlendorf vergleichbares Wohnviertel; Prenzlauer Berg war, wie der Wedding, ein traditioneller Arbeiterbezirk, wovon in den 1990er-Jahren noch genug zu spüren war (auch wenn nicht erst nach 1989 ein sozialer Umstrukturierungsprozess eingesetzt hatte) – inzwischen jedoch ist die Soziallage völlig gekippt.

Ausgewertet wurden die Daten mittels verschiedener quantitativer und qualitativer statistischer Verfahren als Grundlage der soziolinguistisch-sozialwissenschaftli-

chen Interpretation. Dabei wurde davon ausgegangen, dass die Verteilung der Einstellungsindikatoren in der Untersuchungspopulation (hier: der berliner Bevölkerung) nicht homogen ist, sondern eine Mischverteilung verschiedener, in sich homogener Klassen darstellt. Diese Klassen sind durch eine je typische Verteilung der Indikatoren charakterisiert, die mit manifesten Subpopulationen (hier: den Stadtbezirken in Ost- bzw. Westberlin) korreliert sein können, aber nicht notwendigerweise mit ihnen identisch sind. Man spricht deshalb von latenten Klassen. Latent-Class-Analysen (im Folgenden: LCA) und darauf aufbauende statistische Verfahren erlauben es, diese Klassen zu identifizieren und ihre Korrelation mit manifesten Merkmalen der Untersuchungspersonen sowie allfällige Entwicklungstendenzen über die verschiedenen Erhebungszeitpunkte hinweg empirisch darzustellen (methodisch im Detail dazu: Rost 1996; zur Anwendung in der Soziolinguistik: Schmidt-Regener 1998a).

4 Sprachverhalten in Ost- und Westberlin

4.1 Selbsteinschätzung des Sprachgebrauchs

Hinsichtlich der Selbsteinschätzung des Sprachgebrauchs ("Sprechen Sie normalerweise Berliner Dialekt oder Hochdeutsch?"; 0 = Berlinisch [Buv] / 1 = Standardvarietät) ergeben sich (bei $\alpha = 1\%$) keine signifikanten Unterschiede zwischen den Erhebungsstaffeln (χ^2-Quadrat = 6.0683, df = 2, $p > 0.025$), aber, erwartungsgemäß und hoch signifikant, zwischen Ost- und Westberlinern (χ^2-Quadrat = 47.1557, df = 1, $p < 0.001$) und zwischen den einzelnen Stadtbezirken (χ^2-Quadrat = 75.3301, df = 1, $p < 0.001$).

Der über alle Staffeln vergleichsweise hohe Anteil von bekennenden Berlinisch-Sprechern in Ostberlin (vgl. Tab. 2) ist ein erstes Indiz dafür, dass die These vom Anpassungsdruck an das Prestigemodell, unter den die Ostberliner nach '89 geraten seien, für die zweite Hälfte der 90er-Jahre nicht mehr aufrechtzuerhalten ist: Der Gebrauch des Berlinischen bleibt stabil. Gleichermaßen ist der über die Staffeln konstant bleibende Anteil von Berlinisch-Sprechern im Westteil der Stadt ein Indiz dafür, dass eine zweite These, und zwar die vom tendenziellen Dialektverlust der Westberliner, ebenfalls obsolet ist.

	1994	1996	1998
Ost	45,2%	48,6%	54,2%
West	26,5%	25,6%	31,3%
Gesamt	32,9%	35,8%	42.6%

Tabelle 2: Anteil der bekennenden Berlinisch-Sprecher

Als Berlinisch-Sprecher ordnet sich die Hälfte der Ostberliner, aber nur ein gutes Viertel der Westberliner ein (vgl. Tab. 3). Der Westteil der Stadt hat die aufgrund der Sozialprestigehypothese erwarteten Antworten gegeben: 38.8% Berlinisch-

Sprecher im Wedding, aber nur 16.7% in Zehlendorf. Für den Ostteil hingegen bestätigt sich die Hypothese, dass das Berlinische unabhängig von der sozialen Stratifikation verwendet wird, eben weil es dort *nicht* möglich ist, eine Zuordnung zu regional definierbaren Soziallagen via Varietätengebrauch vorzunehmen: In Pankow finden sich sogar etwas mehr Berlinisch-Sprecher als in Prenzlauer Berg.

	Gesamt	Ost	Pb	Pa	West	We	Ze
Buv	37.4%	49.7%	45.6%	54.7%	27.2%	38.8%	16.7%
Standard	62.6%	50.3%	54.4%	45.3%	72.8%	61.2%	83.3%

Tabelle 3: Antworten auf die Frage "Sprechen Sie normalerweise Berliner Dialekt oder Hochdeutsch?"

4.2 Kenntnis und Gebrauch ausgewählter berlinischer Lexik

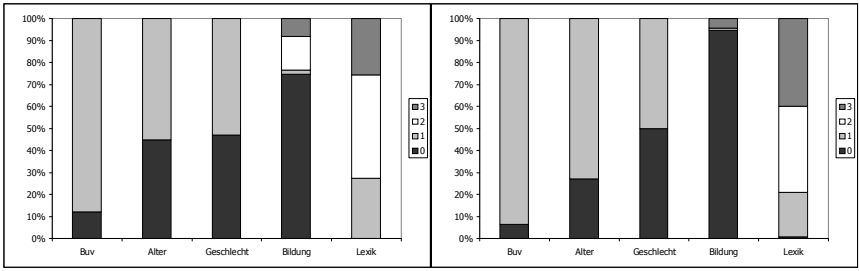
Hinsichtlich der Kenntnisse ausgewählter berlinischer Lexik ergaben sich keine signifikanten Ost-West-Unterschiede; sehr wohl aber hinsichtlich der Gebrauchsintensität. Der Prozentsatz der Ostberliner, die berlinische Lexik hoch frequent einsetzen, ist fast vier Mal so hoch wie der der Westberliner (7.3%/2%). Im Gegenzug verwenden nahezu doppelt so viele West- wie Ostberliner keine der abgefragten nichtstandardsprachlichen Lexik (41.8%/23.6%). Stadteinheitlich ist der Gebrauch nichtberlinspezifischer umgangssprachlicher Lexik (je ein Drittel der Befragten in Ost und West).

Spezifischer gefasst werden können Unterschiede und Gemeinsamkeiten durch Latent-Class-Analysen, in die die Selbsteinschätzung des Sprachgebrauchs (0 = Berlinisch [Buv] / 1 = Standardvarietät) sowie Alter (0 = 36-45 Jahre / 1 = 25-35 Jahre), Geschlecht (0 = weiblich / 1 = männlich) und Bildungsgrad der Informanten (0 = Abitur / 1 = Hauptschule / 2 = Realschule / 3 = Fachoberschule) sowie ihr Gebrauch berlinischer Lexik (0 = Berlinisch-User / 1 = gelegentlicher Gebrauch berlinischer Lexik / 2 = Sprecher der nichtberlinspezifischen Umgangssprache / 3 = Standardvarietätensprecher) als unabhängige Variablen einbezogen wurden. Für den Ostteil und den Westteil Berlins getrennt berechnet, ergaben die LCA jeweils eine 3-Klassen-Lösung als beste Beschreibung der Daten. Die Goodness-of-Fit-Statistiken der Latent-Class-Analysen sind in Tabelle 4 dokumentiert.

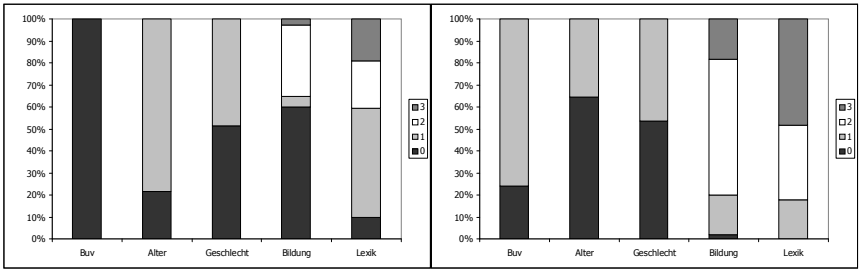
Anzahl der Klassen	Ost			West		
	ln(L)	n(P)	AIC	ln(L)	n(P)	AIC
1	-1344.895	9	2707.790	-1243.238	9	2504.476
2	-1317.956	19	2673.912	-1211.053	19	2460.106
3	-1303.187	29	2664.374	-1194.400	29	2446.800
4	-1300.227	39	2678.454	-1187.012	39	2452.024

Tabelle 4: Goodness-of-Fit-Statistiken. ln(L) = Logarithmus der Likelihood-Funktion, n(P) = Anzahl der unabhängigen Modellparameter, AIC = Akaikes Informationskriterium

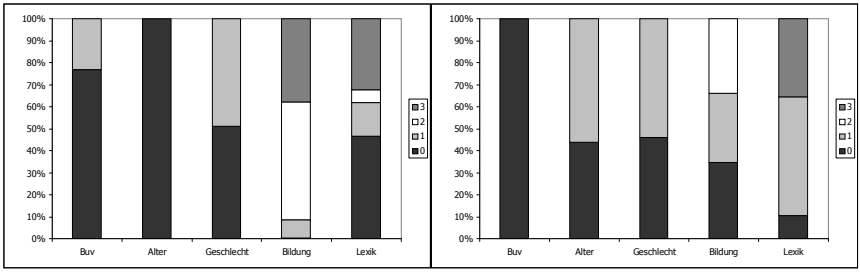
Die Ergebnisse der LCA zeichnen folgendes Bild von den Nutzern der Stadt- bzw. der Standardvarietät (vgl. Abb. 1):



Ost-Klasse 1 (53.4%) West-Klasse 1 (52.5%)



Ost-Klasse 2 (39.2%) West-Klasse 2 (31.5%)



Ost-Klasse 3 (7.4%) West-Klasse 3 (16.0%)

Abbildung 1: Latente Klassen des Gebrauchs berlinischer Lexik. Der Unterschied zwischen den beiden LCA-Modellen (Ost vs. West) ist statistisch hoch signifikant ($\chi^2 = 122.726$, df = 29, $p < 0.001$).

1. In Ost wie auch in West sind es die jüngeren Leute, die weniger prestigebewusst sind und eher zu Verletzungen des standardorientierten Sprachkodexes neigen, wenn auch bzw. aber auf jeweils unterschiedlichen Ebenen:
   - Es sind die 25-35-jährigen, recht gebildeten Ostberliner, die die Majorität derer ausmachen, die sich selbst als Berlinisch-Sprecher einordnen und auch tatsächlich berlinische Lexik verwenden (Ost-Klasse 2). Die Mehrheit der 36-45-jährigen Ostberliner bezeichnet sich zwar als Hochdeutsch-Sprecher (Ost-Klasse 1), benutzt aber auch nichtberlinspezifische umgangssprachliche

Lexik (z. B. "Kneipe").
- In Westberlin hingegen machen die Jüngeren die Majorität dieser zwar bekennenden, in ihrem Sprachverhalten jedoch davon abweichenden Hochdeutsch-Sprecher aus (West-Klasse 1), während die älteren Westberliner eine ausgesprochene Standardorientierung haben (West-Klasse 2).

2. Der Bildungsgrad der Ostberliner, die sich als Berlinisch-Sprecher einordnen und berlinische Lexik verwenden (Ost-Klasse 2), ist weitaus höher als der der Westberliner (West-Klasse 3). Das belegt die These von der nicht sozial abgrenzenden, sondern eher identitätsstiftenden Funktion der Stadtvarietät im Ostteil der Stadt.

3. In den ostberliner Bezirken gibt es keine quantitativ oder gar sozial fassbare Gruppierung, die sich über den Gebrauch der Standardvarietät definieren ließe. Im Westteil hingegen gibt es eine solche Gruppe, in die ein Drittel aller Westberliner gehört (West-Klasse 2), die mehrheitlich älter sind und über eine mittlere Bildung verfügen. Diese große Gruppe mit ausgeprägter Standardorientierung ist insofern für den Westteil der Stadt typisch, als ein Pendant unter den anderen sozialen Bedingungen im Ostteil gar nicht entstehen *konnte:* Zum einen wurde die Notwendigkeit standardsprachlicher Präsentation in der DDR eher unter- als überbewertet. Zum anderen, eng damit verbunden, fand auch die Konstruktion sozialer Scheinrealitäten nicht via standardnahen Sprachgebrauchs statt. Und zum dritten schließlich wurde standardnaher Sprachgebrauch in seinen sozialen Wirkungsmöglichkeiten folglich auch nicht so maßlos überschätzt wie in Teilen der westlichen Sprachgemeinschaft.

5 Einstellungsanalysen

Ost-west-spezifische Gemeinsamkeiten und Unterschiede lassen sich nicht nur auf der unmittelbar objektsprachlichen Ebene nachweisen, sondern gleichwohl in Einstellungen der Stadtvarietät gegenüber.

Zu dieser Einschätzung komme ich aufgrund der Analyse zweier thematisch geordneter Fragenkomplexe, zu denen ich ebenfalls LCA durchgeführt habe:

1. Fragen zu Spracheinstellungen dem Berlinischen generell gegenüber und
2. Fragen zur wechselseitigen sprachlichen Beurteilung der Ost- und Westberliner.

5.1 Einstellungen der Stadtvarietät generell gegenüber

Zur Elizitierung dieser Einstellungen habe ich für die verschiedenen Erhebungsstaffeln getrennte Latent-Class-Analysen durchgeführt, in die neben der Variable Ost-West (0 = Ost / 1 = West) die folgenden Fragebogenitems einbezogen wurden: "Sprechen Sie normalerweise Berliner Dialekt oder Hochdeutsch?" (0 = Berlinisch / 1 = nicht), "Hören Sie den berliner Dialekt gern?" (0 = kommt auf die Situation an / 1 = ja, immer / 2 = nein / 3 = weiß nicht), "Würden Sie Ihren Kindern den berliner Dialekt beibringen?" (0 = k. A. / 1 = ja / 2 = nein / 3 = weiß nicht)

und "In den Programmen der berliner Radiosender sprechen die Moderatorinnen und Moderatoren so gut wie kein Berlinisch. Das ist ..." (0 = in Ordnung so / 1 = manchmal irgendwie schade / 2 = sehr schade / 3 = mir egal).

Zwischen den Erhebungsstaffeln bestehen signifikante Unterschiede (χ^2 = 148.838, df = 82, p < 0.001). Für alle drei Querschnitte ergab sich eine 4-Klassen-Lösung als adäquates Beschreibungsmodell (vgl. Tab. 5-7).

Anzahl der Klassen	ln(L)	n(P)	AIC
1	-1057.875	11	2137.749
2	-975.936	23	1997.872
3	-950.680	35	1971.359
4	-936.866	47	1967.732
5	-931.942	59	1981.884
6	-926.654	71	1995.307

Tabelle 5: Erhebungsstaffel 1994. Goodness-of-Fit-Statistiken. ln(L) = Logarithmus der Likelihood-Funktion, n(P) = Anzahl der geschätzten Modellparameter, AIC = Akaikes Informationskriterium

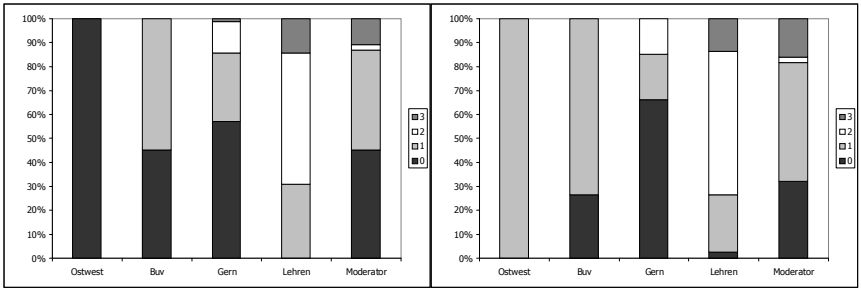
Anzahl der Klassen	ln(L)	n(P)	AIC
1	-1375.203	11	2772.406
2	-1306.280	23	2658.560
3	-1276.477	35	2622.953
4	-1262.332	47	2618.664
5	-1253.645	59	2625.290
6	-1247.251	71	2636.502

Tabelle 6: Erhebungsstaffel 1996. Goodness-of-Fit-Statistiken. (Legende vgl. Tab. 5)

Anzahl der Klassen	ln(L)	n(P)	AIC
1	-1444.188	11	2910.377
2	-1360.927	23	2767.853
3	-1327.253	35	2724.507
4	-1309.037	47	2712.073
5	-1297.044	59	2712.088
6	-1296.003	71	2734.006

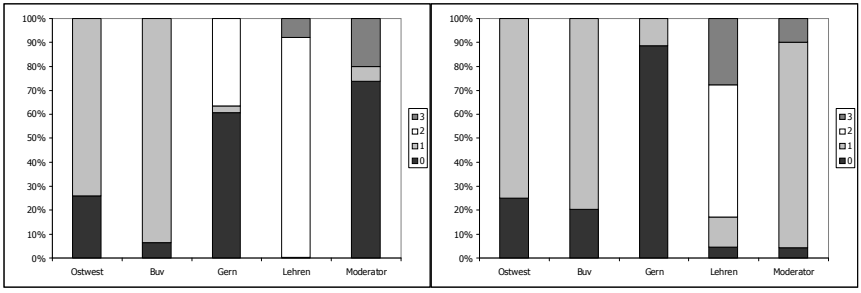
Tabelle 7: Erhebungsstaffel 1998. Goodness-of-Fit-Statistiken. (Legende vgl. Tab. 5)

Ich beschränke mich aus Platzgründen auf die statistische Dokumentation der Gesamtverteilungen (vgl. Abb. 2, 4 und 6) und der Ergebnisse der LCA (vgl. Abb. 3, 5 und 7), ohne sie näher zu beschreiben (dazu vgl. Regener 2000, S. 17-27), weil die interessantesten Ergebnisse hier aus der Analyse der Entwicklungen zwischen den Untersuchungsstaffeln resultieren.

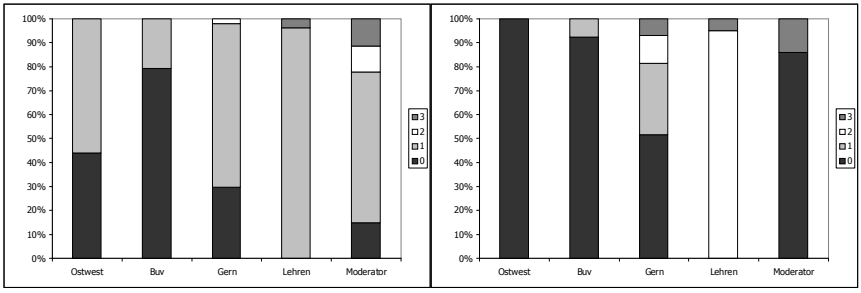


Gesamtverteilung Ost 1994 Gesamtverteilung West 1994

Abbildung 2: Antworten auf die Fragen zu Spracheinstellungen dem Berlinischen gegenüber



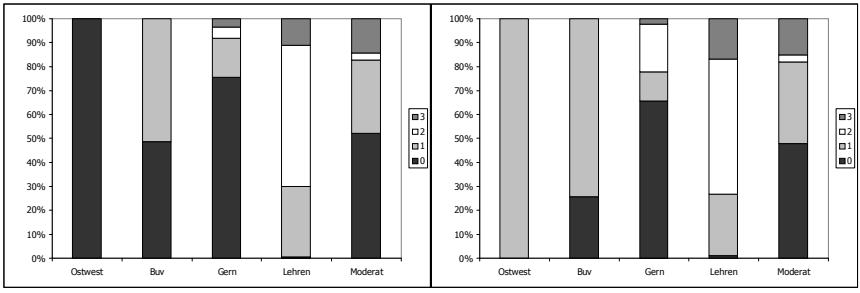
Klasse 1 (36.0%) Klasse 2 (35.4%)



Klasse 3 (22.7%) Klasse 4 (5.9%)

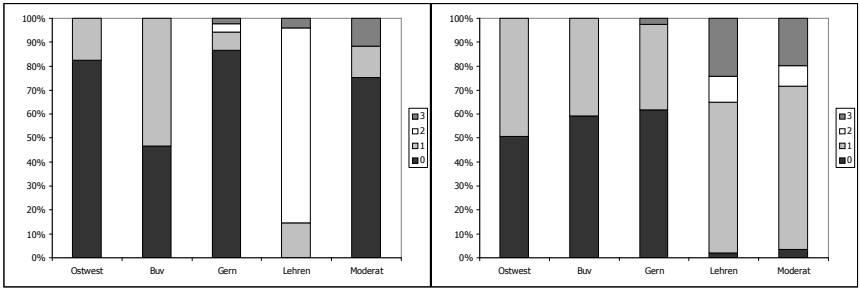
Abbildung 3: Ergebnisse der LCA (1994)

Eine Klassencharakteristik, die sich in allen drei Staffeln substanziell unverändert nachweisen ließ, ist die der dem Berlinischen eindeutig ablehnend gegenüberstehenden. Diese Klasse verändert ihre Größe jeweils ziemlich dramatisch, und im Zusammenhang damit auch ein wenig ihr Profil. In der '94er Staffel ist sie mit 36% die größte Klasse, in der zwar im Besonderen Westberliner, aber auch viele Ostberliner vertreten sind. 1996 gehören nur noch knapp 16% des Samples in diese Klasse, die gleichzeitig jedoch klar westdominiert geworden ist. Mit der Verkleine-

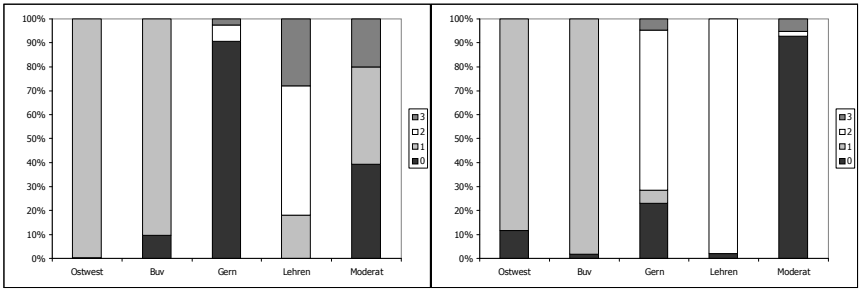


Gesamtverteilung Ost 1996 Gesamtverteilung West 1996

Abbildung 4: Antworten auf die Fragen zu Spracheinstellungen dem Berlinischen gegenüber



Klasse 1 (35.1%) Klasse 2 (29.8%)

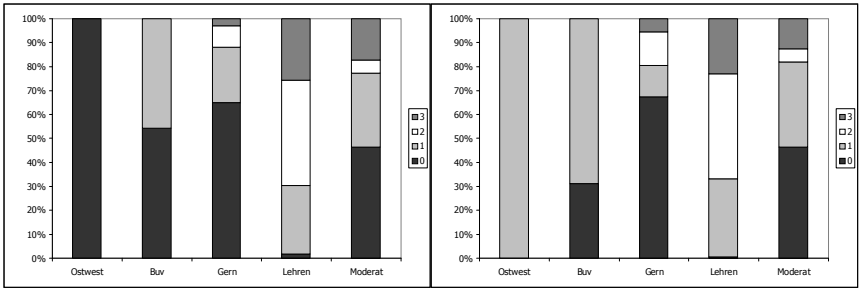


Klasse 3 (19.5%) Klasse 4 (15.6%)

Abbildung 5: Ergebnisse der LCA (1996)

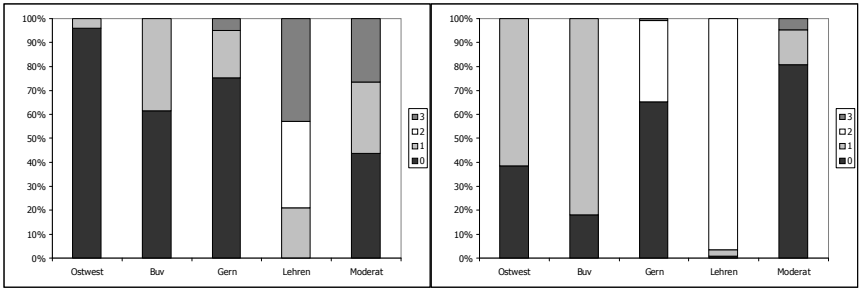
rung einhergegangen ist eine Radikalisierung der Ablehnung. Nach 1998 hin verdoppelt sich die Klasse wieder, ist immer noch westdominiert, aber nicht mehr so eindeutig wie 1996 und von zurückgenommenerer Reserviertheit. In allen drei Staffeln ist dies eine Klasse der Nichtberlinisch-Sprecher.

1994 gibt es eine der Stadtvarietät positiv gegenüberstehende Klasse (22.7%), die eher von Ostberlinern dominiert wird. 1996 gibt es eine mit 29.8% etwas größere, gemäßigter berlinischfreundlich eingestellte Ost-West-Klasse, die 1998 in Größe

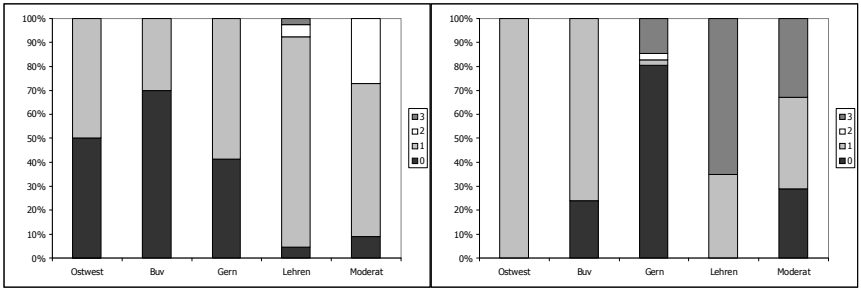


Gesamtverteilung Ost 1998 Gesamtverteilung West 1998

Abbildung 6: Antworten auf die Fragen zu Spracheinstellungen dem Berlinischen gegenüber



Klasse 1 (32.7%) Klasse 2 (32.2%)



Klasse 3 (20.0%) Klasse 4 (15.1%)

Abbildung 7: Ergebnisse der LCA (1998)

und Profil wieder der '94er vergleichbar ist. Es ist nicht überraschend, dass diese Klassen immer vorrangig von Berlinisch-Sprechern besetzt sind.

Die zweitgrößte Klasse der '94er Befragung (35.4%) ist eher eine Westklasse und in ihrer Positionierung nicht klar entschieden, aber mit leicht positiver Tendenz. Nicht klar entschieden ist auch die drittgrößte Klasse der '96er Befragung (19.5%), die ebenfalls eher westdominiert ist, aber eine eher negative Positionierungstendenz aufweist. Jedoch ist sie nur noch nahezu halb so groß wie 1994. Daraus sind

1998 zwei Klassen geworden: eine große Ost- und eine kleine Westklasse (32.7%/ 15.1%). Die positive Tendenz tritt in beiden wieder stärker auf, ganz besonders jedoch in der Westklasse. In diesen Klassen finden sich in erster Linie Nichtberlinisch-Sprecher.

Die kleinste Klasse der '94er Befragung (5.9%) ist eine reine Ostklasse von ganz besonderem Profil: In dieser Klasse befinden sich nahezu ausschließlich Berlinisch-Sprecher, deren Antworten jedoch ein nicht mehr gewagtes Bekenntnis zur angestammten Varietät indizieren und die letztlich eher dem ablehnenden Spektrum zuzuordnen ist. Diese verunsicherte Ostklasse explodiert nach '96 regelrecht, und zwar durch Liberalisierung. In der '98er Staffel verschwindet sie völlig.

Insgesamt ist somit in Ost und West von '94 nach '96 eine Verunsicherung ablehnender wie zustimmender Positionierungen zu beobachten, von '96 nach '98 hingegen eine zunehmend positive Haltung der Stadtvarietät gegenüber. Dieser stadteinheitliche Trend zu einer berlinischfreundlicheren Einstellung ist je ost- bzw. westspezifisch festzumachen. Dass die Westberliner ein positiveres Verhältnis zur städtischen Umgangssprache gewinnen, äußert sich vorrangig in einer Verunsicherung der ablehnenden Position, aber auch in der Zunahme der Akzeptanz. Die Ostberliner beginnen nach '96, ihre Unsicherheit zu überwinden und ihre Fähigkeit zu klarer Positionsbestimmung zurückzugewinnen. Dass das namentlich beim Vergleich der '96er mit der '98er Staffel auffällt, ist gut kompatibel mit Befunden der Sozialforschung, nach denen in der deutsch-deutschen Annäherungs- vs. Abgrenzungsgeschichte nach 1995 "etwas passiert" ist; d. h. die Annäherung wurde aus der Ostperspektive durch Distanzierung ersetzt (Brunner & Walz 1998, Kaase & Bauer-Kaase 1998). Im konkreten Falle bedeutet das die Wiederentdeckung des Selbstidentifikationsmittels Stadtvarietät durch die Ostberliner mit der Möglichkeit (1) sowohl der positiven als auch der negativen Positionierung zu diesem Identitätsmarker und (2) der *selbstbestimmten* Abgrenzung gegenüber denen, die nicht als dessen Träger wahrgenommen werden. Die Entwicklung, die die Westberliner durchliefen, spiegelt hingegen eher einen ganz normalen Rückbesinnungsprozess auf positive Aspekte von ehemals bereits Vertrautem. Und dies ist sicher dem Kontakt mit den der Stadtvarietät objektsprachlich wie auch von den Einstellungen her aufgeschlossener gegenüberstehenden Ostberlinern geschuldet.

5.2 Wechselseitige Beurteilung des Sprachverhaltens in Ost- bzw. Westberlin

Ost-west-spezifische Gemeinsamkeiten und Unterschiede lassen sich nicht nur auf der unmittelbar objektsprachlichen Ebene nachweisen, sondern gleichwohl in Einstellungen der Stadtvarietät gegenüber.

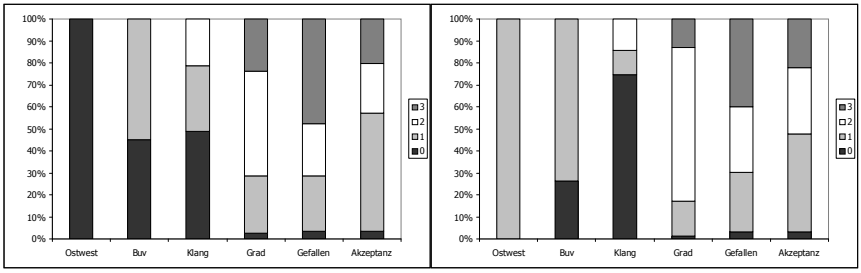
Zu dieser Einschätzung komme ich einerseits aufgrund der bereits in Kapitel 5.1 skizzierten Analyse der Spracheinstellungen dem Berlinischen generell gegenüber und andererseits aufgrund der Analyse von Fragen zur wechselseitigen sprachlichen Beurteilung der Ost- und Westberliner. Zu diesem Fragenkomplex habe ich

für die verschiedenen Erhebungsstaffeln getrennte Latent-Class-Analysen durchgeführt, in die neben der Variable Ost-West (0 = Ost / 1 = West) die folgenden Fragebogenitems einbezogen wurden: "Sprechen Sie normalerweise Berliner Dialekt oder Hochdeutsch?" (0 = Berlinisch / 1 = nicht), "Klingt der Berliner Dialekt bei Ostberlinern anders als bei Westberlinern?" (0 = ja / 1 = nein / 2 = weiß nicht), "Finden Sie, dass die Leute in dem für Sie 'anderen Teil' Berlins mehr oder weniger berlinern als in 'Ihrer' Stadthälfte" (0 = k. A. / 1 = weiß nicht / 2 = mehr bzw. weniger / 3 = gleich stark), "Gefällt es Ihnen, wie die Leute im 'anderen' Teil der Stadt sprechen?" (0 = k. A. / 1 = ja / 2 = nein / 3 = weiß nicht) und "Glauben Sie, dass die Akzeptanz des Berlinischen nach '89 in ganz Berlin eher zu- oder abgenommen hat?" (0 = k. A. / 1 = weiß nicht / 2 = eher zugenommen / 3 = eher abgenommen).

Der Unterschied zwischen den LCA-Modellen der verschiedenen Staffeln ist bei Zugrundelegung eines 5-Klassen-Modells, wie es vom AIC für die Gesamtverteilung favorisiert wird, statistisch hoch signifikant ($\chi^2 = 229.964$, df = 138, $p < 0.001$).[2]

5.2.1 Ergebnisse der ersten Erhebungsstaffel 1994

Analyse der Gesamtverteilungen



Gesamtverteilung Ost 1994 Gesamtverteilung West 1994

Abbildung 8: Antworten auf die Fragen zur wechselseitigen Beurteilung des Sprachverhaltens in Ost- und Westberlin

74.7% der befragten Westberliner, aber nur 48.8% der Ostberliner nehmen wahr, dass die Stadtvarietät in der jeweils anderen Stadthälfte anders klingt. Im Gegenzug finden 29.8% der Ostberliner dies nicht, aber nur 11.1% der Westberliner. 14.2% der Ost- und 21.4% der Westberliner haben dazu keine Meinung. Dass in der anderen Stadthälfte merklich mehr bzw. weniger berlinert wird, finden 69.8% der West- und 47.6% der Ostberliner, d. h. sie stellen graduelle Unterschiede in der Dialekttiefe fest. 26.2% (Ost) bzw. 16% (West) wissen dazu nichts zu sagen, und 23.8% der Ost- und 13% der Westberliner können keine Unterschiede feststellen. 47.6% der Ostberliner und 40.1% der Westberliner sind sich nicht darüber

2 Da der AIC für die ersten beiden Erhebungsstaffeln (1994 und 1996) jedoch eine 4-Klassen-Lösung favorisiert, wird für diese beiden Staffeln im Folgenden das sparsamere 4-Klassen-Modell dargestellt.

im Klaren, ob ihnen der Sprachgebrauch in der anderen Stadthälfte zusagt. In Ostberlin bejahen 25% die Frage, und 23.8% verneinen sie. In Westberlin sagen 27.2% 'ja', und 29.6% 'nein'. 53.6% der Ost- und 44.4% der Westberliner wissen auch nicht zu sagen, ob die Akzeptanz des Berlinischen nach '89 gestiegen oder gesunken ist. 30.2% (West) bzw. 22.6% (Ost) meinen, sie habe zugenommen, und 22.2% (West) bzw. 20.2% (Ost) sagen das Gegenteil.

Beim Vergleich der Gesamtverteilungen fällt auf, dass die Antworten der Ostberliner in jedem Falle nivellierender, weniger Unterschiede wahrnehmend und ausweichender sind als die der Westberliner. Stadteinheitlich ist der Anteil der Weißnicht-Antworten auf die Fragen nach Gefallen und Akzeptanz vergleichsweise hoch. In Ostberlin ist es außerdem so, dass sich die Positiv-negativ-Antworten zu einem Nullsummenergebnis neutralisieren, in Westberlin trifft dies nur für die Gefallensfrage zu. Bei diesen Positionierungs- bzw. direkten Einstellungsfragen ist also eine stadteinheitliche Zurückhaltung feststellbar.

Latent-Class-Analyse

Für 1994 ist laut AIC-Kriterium (vgl. Tab. 8) das 4-Klassen-Modell das die Daten adäquat beschreibende.

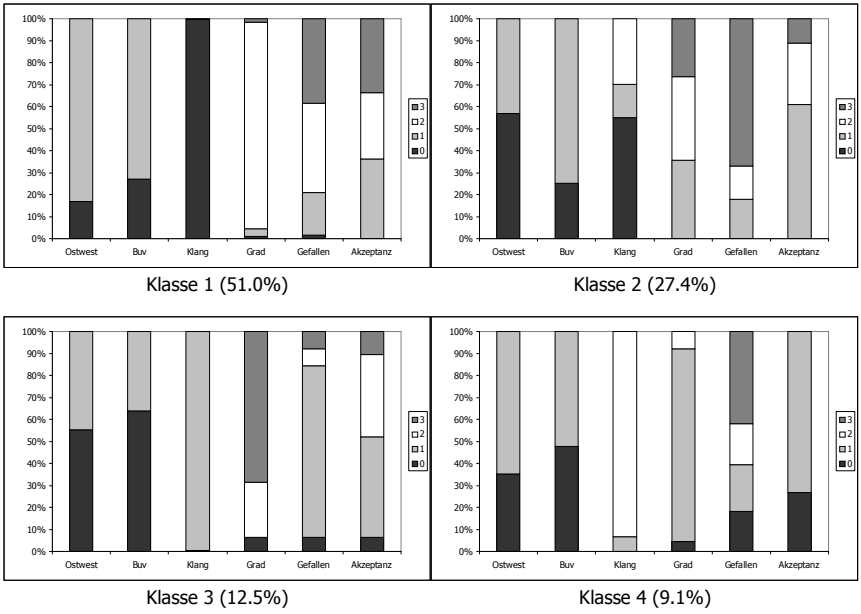
Anzahl der Klassen	ln(L)	n(P)	AIC
1	-1344.879	13	2715.758
2	-1264.866	27	2583.732
3	-1231.844	41	2545.688
4	-1216.483	55	2542.965
5	-1209.058	69	2556.115
6	-1190.390	83	2546.780

Tabelle 8: Goodness-of-Fit-Statistiken. ln(L) = Logarithmus der Likelihood-Funktion, n(P) = Anzahl der geschätzten Modellparameter, AIC = Akaikes Informationskriterium

Die größte Klasse, in die 51% aller Befragten gehören, ist von Westberlinern und Nichtberlinisch-Sprechern dominiert (64.5% aller Westberliner sind in dieser Klasse und 24.9% der Ostberliner), die ganz klar die Unterschiede im Klang des Berlinischen wie auch im Grad des Berlinerns zwischen Ost und West wahrnimmt. Ebenso klar drückt diese Klasse ihre Ablehnung des 'fremden' Sprachgebrauchs aus. Hinsichtlich der Akzeptanz des Berlinischen herrscht eher der Meinung, sie habe abgenommen. Die Klasse ist recht entschieden, und zwar entschieden reserviert.

Die zweite Klasse konstituieren 27.4% der Befragten. Sie ist von Ostberlinern und Nichtberlinisch-Sprechern dominiert (45.5% aller Ostberliner gehören in diese Klasse). In den entschiedenen Antworten auf die Sachfragen (Klang, Grad) ist diese Klasse der ersten ähnlich; Unterschiede liegen im Ausgeprägtheitsgrad der vertretenen Meinung. Die Antworten auf die Frage nach dem Gefallen haben die höchste Weiß-nicht-Quote aller Klassen; die zweithöchste haben die Antworten auf

die Akzeptanzfrage. Auch diese Klasse zeigt eine gewisse Entschiedenheit, und zwar bei den Sachfragen, ist aber ansonsten durch eine Tendenz zu nicht eindeutigen Antworten charakterisiert.



Klasse 1 (51.0%) Klasse 2 (27.4%)

Klasse 3 (12.5%) Klasse 4 (9.1%)

Abbildung 9: Ergebnisse der LCA (1994)

Die beiden kleinen Klassen spezifizieren dieses Nichtpositionieren jeweils. Die drittgrößte Klasse versammelt 12.5% aller Befragten. Es ist eher eine Ostklasse (20.2% der Ostberliner sind darin im Unterschied zu 8.5% der Westberliner) und eine der Berlinisch-Sprecher. Dies ist die Klasse der in den Sachfragen keine Unterschiede Wahrnehmenden mit positiver Grundeinstellung, in der es die größte Akzeptanz des Sprachverhaltens im anderen Stadtteil bzw. eine große Toleranz diesem gegenüber gibt. Hier findet sich der höchste Prozentsatz derjenigen, die eine steigende Akzeptanz der Stadtvarietät bemerkt zu haben glauben.

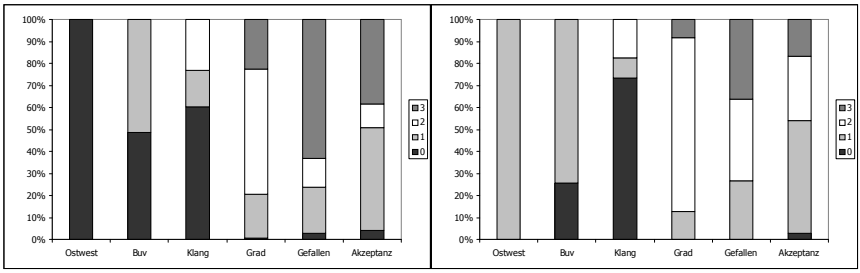
Die kleinste Klasse umfasst 9.1% der Befragten und weist ebenso eine ausgewogene Ost-West-Verteilung wie auch der Berlinisch- und Nichtberlinisch-Sprecher auf. Diese Klasse ist die der Indifferenten, Desinteressierten: Bei jeder Variable (außer bei der Gefallensfrage, die das Nullsummenergebnis der Gesamtverteilungen spiegelt) hat die Weiß-nicht-Antwort Werte von 73% an aufwärts.

1994 gibt es eine sehr entschiedene, reserviert-ablehnende Haltung dem in der jeweils anderen Stadthälfte beobachteten Sprachverhalten gegenüber, ganz besonders in Westberlin. Diese Ablehnung resultiert aus der klaren Erkenntnis der Ost-West-Unterschiede. Im Ostteil ist es eher so, dass Sachunterschiede nicht of-

fen zur Abgrenzung herangezogen, sondern abwartend neutral zurückhaltend beurteilt werden. In Ost wie in West gibt es eine nicht sehr weit verbreitete gleichgültige Einstellung, und die positivste Haltung findet sich vorrangig unter Ostberlinern, die im wesentlichen Ost-West-Gleichheit in den Sachfragen festzustellen glauben.

5.2.2 Ergebnisse der zweiten Erhebungsstaffel 1996

Analyse der Gesamtverteilungen



Gesamtverteilung Ost 1996 Gesamtverteilung West 1996

Abbildung 10: Antworten auf die Fragen zur wechselseitigen Beurteilung des Sprachverhaltens in Ost- und Westberlin

73.3% der Westberliner und 60.3% der Ostberliner nehmen 1996 Unterschiede im Klang des Berlinischen in Ost und West wahr, 16.3% (Ost) bzw. 9.3% (West) tun dies nicht, und 23.3% (Ost) bzw. 17.4% (West) können sich nicht entscheiden. 79.1% der West- und 56.8% der Ostberliner bemerken auch Unterschiede in der Dialekttiefe zwischen beiden Stadthälften, 22.6% (Ost) bzw. 8.1% (West) sehen keine Unterschiede, und 19.9% (Ost) bzw. 12.8% (West) wissen es nicht. 63% der Ostberliner und 36% der Westberliner wissen auch nicht, ob ihnen der Sprachgebrauch in der anderen Stadthälfte gefällt. 37.2% der Westberliner geben zu, den Sprachgebrauch im Osten nicht zu mögen, aber nur 13% der Ostberliner tun gleiches für den Westteil. 21.2% der Ostberliner und 26.7% der Westberliner beantworten die Frage mit 'ja'. 51.2% (West) bzw. 46.6% (Ost) ziehen sich bei der Frage nach der Entwicklung der Akzeptanz des Berlinischen nach '89 auf die Weiß-nicht-Antwort zurück; 38.4% der Ost- bzw. 16.9% der Westberliner meinen, sie habe abgenommen; 29.1% der West- und 11% der Ostberliner, sie habe zugenommen. Die hohe Quote der Ostberliner, die glauben, sinkende Akzeptanz der Stadtvarietät festgestellt zu haben, deutet auf einen durchaus wahrgenommenen Prestigeverlust der einstigen 'Hauptstadtvarietät' hin.

Überhaupt hat im Vergleich zu 1994 die Entschiedenheit in den Antworten der Ostberliner deutlich zugenommen mit einer Ausnahme: Die Ja-Antworten auf die Gefallensfrage sind um fast 11% zurückgegangen und werden durch den Anstieg der Weiß-nicht-Antworten kompensiert.

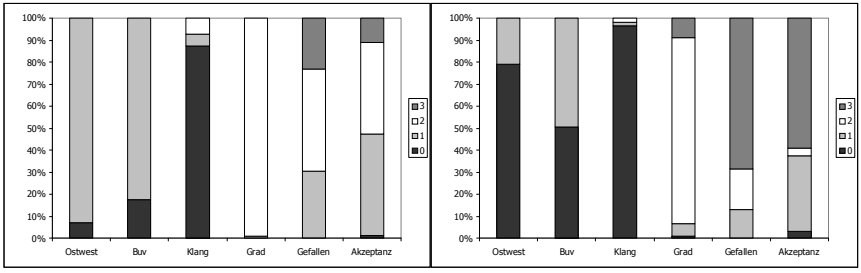
Auch im Westteil der Stadt sind die unentschiedenen Antworten leicht rückläufig, wenn auch nicht in dem Maße und mit ebenfalls einer Ausnahme: Die Unsicherheitsquote in den Antworten auf die Frage nach der Akzeptanz nimmt zu zulasten derjenigen, die sinkende Akzeptanz der Stadtvarietät festgestellt haben. Und: Der Anteil derjenigen, denen das Sprachverhalten der Ostberliner nicht gefällt, steigt um fast 8%.

Latent-Class-Analyse

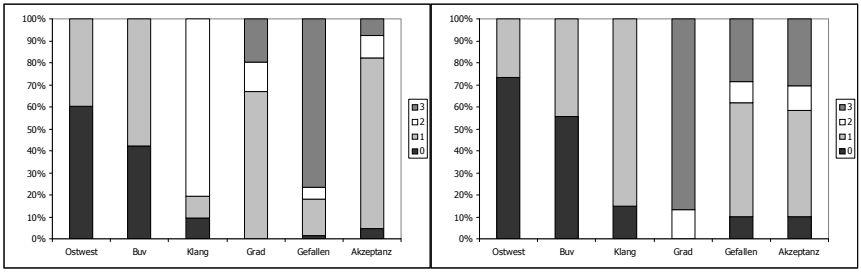
Auch die '96er Daten werden durch ein 4-Klassen-Modell optimal beschrieben (vgl. Tab. 9).

Anzahl der Klassen	ln(L)	n(P)	AIC
1	-1682.271	13	3390.543
2	-1580.327	27	3214.645
3	-1534.049	41	3150.099
4	-1509.306	55	3128.611
5	-1501.876	69	3141.752
6	-1493.256	83	3152.513

Tabelle 9: Goodness-of-Fit-Statistiken. ln(L) = Logarithmus der Likelihood-Funktion, n(P) = Anzahl der geschätzten Modellparameter, AIC = Akaikes Informationskriterium



Klasse 1 (39.9%) Klasse 2 (30.3%)



Klasse 3 (20.6%) Klasse 4 (9.2%)

Abbildung 11: Ergebnisse der LCA (1996)

Die größte Klasse der '96er Staffel wird von 39.9% der Befragten gebildet. Es ist *die* Westklasse, in die 68.7% aller Westberliner gehören und die von Nichtberlinisch-Sprechern dominiert wird. Sie reflektiert die Ost-West-Unterschiede in den Sachfragen, gibt bipolare Antworten auf die Gefallensfrage und ist mehrheitlich der Auffassung, dass die Akzeptanz des Berlinischen zugenommen habe. Die Klasse ist also mit einer festen Meinung ausgestattet, ohne diese mit Reserviertheit oder gar Ablehnung zu verbinden.

Die zweitgrößte Klasse umfasst 30.3% des Samples und 52.2% aller Ostberliner. Eine Dominanz von Berlinisch- oder Nichtberlinisch-Sprechern gibt es nicht. Bei den Sachfragen sind die Ergebnisse ähnlich wie in der ersten Klasse; deutliche Unterschiede aber gibt es bei den Einstellungs- bzw. Positionierungsfragen. Bei der Gefallensfrage zieht sich die Klasse auf die Weiß-nicht-Antwort zurück; was die Akzeptanz der Stadtvarietät betrifft, so ist sie ganz eindeutig der Meinung, sie habe abgenommen.

Die drittgrößte Klasse bilden 20.6% der Befragten. Es ist eher eine Ostklasse ohne Dominanz von Berlinisch- oder Nichtberlinisch-Sprechern. Es ist dies die Weißnicht-Klasse des '96er Querschnitts.

Die kleinste Klasse, die von 9.2% der Befragten konstituiert wird, ist eindeutig eine Ostklasse ohne Dominanz einer Varietät. Sie nimmt auf der objektsprachlichen Ebene keine Ost-West-Unterschiede wahr, hat den höchsten Anteil derer, denen der Sprachgebrauch in der jeweils anderen Stadthälfte gefällt, und hat eher abnehmende Akzeptanz des Berlinischen festgestellt.

Die Profile der '96er Klassen zeigen ein relativ deutlich ausgeprägtes neutrales Spektrum in den Positionierungsfragen. Aus klarer Wahrnehmung der Differenzen in den Sachfragen baut sich keine Ablehnung des Sprachverhaltens in der jeweils anderen Stadthälfte auf: Der Westen positioniert sich klar, aber sowohl negativ als auch positiv; der Osten favorisiert Neutralität. Gleichermaßen in Ost wie in West gibt es eine gleichgültige Haltung, und der positive Pol des Spektrums ist von Ostberlinern belegt, die keine Unterschiede in den Sachfragen bemerken.

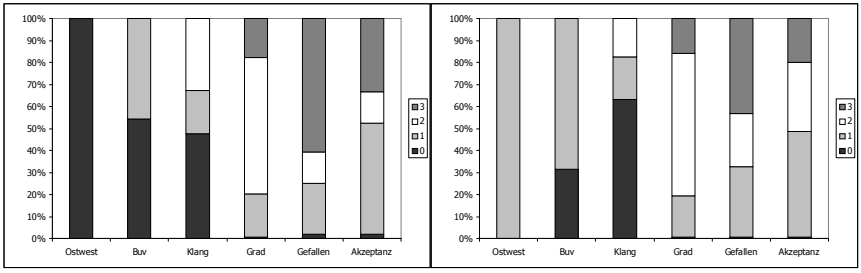
5.2.3 Ergebnisse der dritten Erhebungsstaffel 1998

Analyse der Gesamtverteilungen

Unterschiede im Klang des Berlinischen in Ost und West stellen 63.2% der 1998 befragten Westberliner fest, aber nur 47.6% der Ostberliner. 19.6% der Ost- bzw. 19.4% der Westberliner tun dies nicht, und 32.7% (Ost) bzw. 17.4% (West) haben keine Meinung dazu. Graduelle Unterschiede in der Dialekttiefe bemerkt ein nahezu identischer Prozentsatz in West (64.6%) und Ost (61.9%). 17.9% (Ost) und 16% (West) finden keine Unterschiede, 19.6% (Ost) und 18.6% (West) wissen es nicht. 60.7% der Ost- und 43.1% der Westberliner sind sich nicht darüber im Klaren, ob ihnen der Sprachgebrauch in der jeweils anderen Stadthälfte zusagt; 23.2% der Ostberliner bejahen die Frage, und 14.3% verneinen sie; in Westberlin

sagen 31.9% 'ja', und 24.3% 'nein'. 50.6% der Ost- und 47.9% der Westberliner können auch nicht sagen, ob die Akzeptanz der Stadtvarietät nach '89 gestiegen oder gesunken ist; 31.2% der West- und 14.3% der Ostberliner meinen ersteres, 33.3% der Ost- und 20.1% der Westberliner letzteres.

Außer bei den Antworten auf die Frage nach Unterschieden im Klang (Zunahme der unentschiedenen Antworten) zeigt sich in dem ostberliner Ergebnis Kontinuität zu dem der '96er Staffel.



Gesamtverteilung Ost 1998 Gesamtverteilung West 1998

Abbildung 12: Antworten auf die Fragen zur wechselseitigen Beurteilung des Sprachverhaltens in Ost- und Westberlin

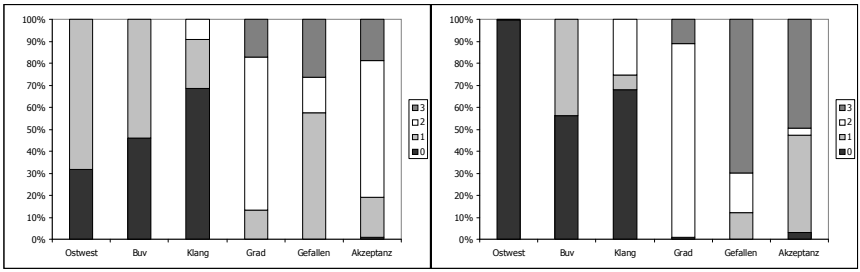
Auch für die Antworten der Westberliner ist im Vergleich mit den '96er Ergebnissen eine generelle Tendenz mit einer Ausnahme festzumachen: Die entschiedenen Antworten gehen zurück (bei der Gefallensfrage nimmt die negative Antwort um 13% ab); lediglich bei der Frage nach der Akzeptanz der Stadtvarietät gibt es keinen wesentlichen Unterschied zu 1996.

Latent-Class-Analyse

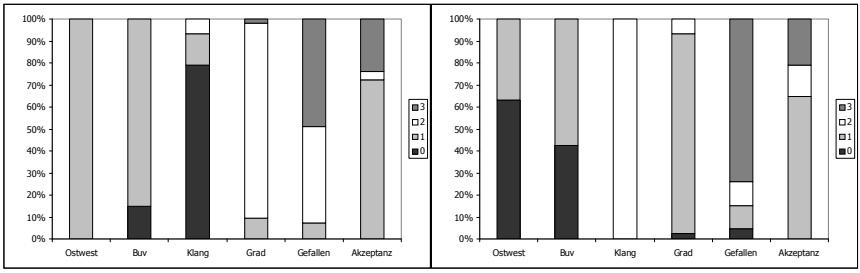
Für 1998 ist ein 5-Klassen-Modell angemessen zur Beschreibung der Daten (vgl. Tab. 10).

Anzahl der Klassen	ln(L)	n(P)	AIC
1	-1706.345	13	3438.691
2	-1656.500	27	3366.999
3	-1628.491	41	3338.981
4	-1607.361	55	3324.722
5	-1592.917	69	3323.833
6	-1586.297	83	3338.594

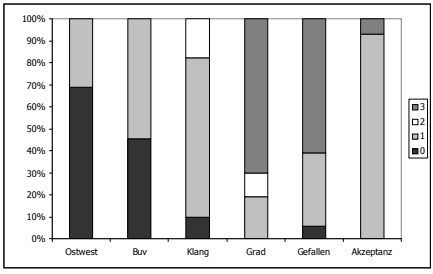
Tabelle 10: Goodness-of-Fit-Statistiken. ln(L) = Logarithmus der Likelihood-Funktion, n(P) = Anzahl der geschätzten Modellparameter, AIC = Akaikes Informationskriterium



Klasse 1 (30.0%) Klasse 2 (28.4%)



Klasse 3 (17.3%) Klasse 4 (12.4%)



Klasse 5 (11.8%)

Abbildung 13: Ergebnisse der LCA (1998)

Die größte Klasse der '98er Staffel wird von 30% der Befragten gebildet. Es ist eine Westklasse, in die 44.5% der Westberliner gehören und die weder von Berlinisch- noch von Nichtberlinisch-Sprechern dominiert wird. Sie nimmt mehrheitlich Ost-West-Unterschiede im Klang und beim Grad des Berlinerns wahr. In dieser Klasse findet sich der höchste Prozentsatz derer, denen der Sprachgebrauch in der anderen Stadthälfte gefällt. Es ist somit die Klasse, die dem anderen Stadtteil am aufgeschlossensten und positivsten gegenübersteht, womit auch das Ergebnis gut kompatibel ist, dass sie den höchsten Prozentsatz derjenigen aufweist, die eine steigende Akzeptanz der Stadtvarietät nach '89 konstatieren.

Die zweitgrößte Klasse, in die 28.4% der Befragten gehören, ist in den Antworten auf die Sachfragen das Ostpendant zu Klasse 1 (mit 52.6% aller Ostberliner). Bei

der Gefallensfrage jedoch zieht sie sich wieder auf die Weiß-nicht-Antwort zurück und ist mehrheitlich der Meinung, die Akzeptanz der Stadtvarietät habe abgenommen.

Die drittgrößte Klasse vereinigt 17.3% der Befragten, die ausschließlich in Westberlin wohnen und mit sehr großer Mehrheit Nichtberlinisch-Sprecher sind. Diese Klasse nimmt die Unterschiede zwischen Ost und West sehr klar wahr, mag den Sprachgebrauch in Ostberlin nicht und wagt kein Urteil über die Entwicklung der Akzeptanz des Berlinischen nach '89.

Die vierte Klasse (12.4% der Befragten) ist eher eine Ostklasse ohne Dominanz einer Varietät. Es ist die Weiß-nicht-Klasse der '98er Befragung.

In die kleinste Klasse schließlich gehören 11.8% des Samples. In dieser Ostklasse befinden sich Berlinisch- und Nichtberlinisch-Sprecher gleichermaßen, die in den Sachfragen keine Unterschiede zwischen Ost- und Westsprache feststellen können, denen der Sprachgebrauch in der anderen Stadthälfte eher gefällt und die die Akzeptanzfrage nahezu ausschließlich ratlos zurücklässt.

1998 ist es so, dass eine aus klaren Sachurteilen gespeiste Ablehnung des Sprachgebrauchs in der anderen Stadthälfte nur in Westberlin zu finden ist. Die große Mehrheit der Berliner nimmt die Sachdifferenzen sehr wohl wahr. Die Westberliner verbinden dies mit einer ausgesprochen positiven Haltung, die Ostberliner wahren Neutralität. Die Materialisierung der generell unentschiedenen Antworten findet sich im Profil einer Ost(west)klasse; und diejenigen, die keine Unterschiede wahrnehmen, sind auch vorrangig im Ostteil zu suchen.

5.2.4 Vergleich der Erhebungsstaffeln

Im Wesentlichen gibt es vier verschiedene Klassenprofile mit mehr oder minder großen Modifikationen in allen drei Erhebungsstaffeln.

Die größte Klasse des '94er Querschnitts ist in ihren Antworten klar entschieden und steht dem Sprachgebrauch in der jeweils anderen Stadthälfte ablehnend gegenüber. In dieser Klasse befindet sich die Majorität der Westberliner, aber auch ein Viertel der Ostberliner. 1996 gibt es eine gleichfalls klar entschiedene Klasse in gleicher Größe, die jedoch eine nahezu reine Westklasse geworden ist und die dem Sprachgebrauch in Ostberlin sowohl negativ als auch positiv gegenübersteht. 1998 sind aus dieser Klasse zwei geworden: Es gibt wieder eine entschieden-reservierte Westklasse, die aber etwas verunsichert und nicht einmal mehr halb so groß ist; und eine sehr große Westklasse, die dem Sprachgebrauch in der anderen Stadthälfte ausgesprochen positiv gegenübersteht (am positivsten in der '98er Staffel überhaupt) und die ihren Zuwachs dem Schrumpfen der Weiß-nicht-Klasse verdankt.

Die zweitgrößte Klasse der '94er Erhebungsstaffel ist eine Ostklasse mit einer zwar eindeutigen, aber nicht ganz so entschieden vorgetragenen Meinung in den Sach-

fragen, die sich bei den Positionierungsfragen auf die Weiß-nicht-Antwort zurückzieht. Ihre Größe bleibt nach '96 hin konstant, aber sie verändert ihr Profil erheblich: Ihre Antworten werden ebenso entschieden wie die der Westklasse. Lediglich bei der Gefallensfrage bleibt es bei der Weiß-nicht-Antwort. Nach 1998 kontinuiert sie Charakter und Klassengröße.

Die Klasse, die in allen drei Staffeln keine Unterschiede zwischen Ost und West in den Sachfragen bemerkt, ist eine kleine, ihre Größe leicht variierende Ostklasse, die in den positiven Antworten auf die Gefallensfrage immer über dem Prozentsatz der Gesamtverteilung liegt, aber mit eindeutig fallender Tendenz. 1994 und 1996 ist sie die der anderen Stadthälfte am positivsten gegenüberstehende.

Und schließlich gibt es noch die Weiß-nicht-Klasse, die 1994 eine kleine gemischte Ost-West-Klasse ist, die 1996 sehr großen Zulauf bekommt, und zwar nicht nur, aber in erster Linie von Ostberlinern, die die entschieden-reservierte Klasse verlassen. Nach 1998 hin wird sie wieder deutlich kleiner.

Wenn auch stadteinheitlich ein Trend zu größerer Zurückhaltung bei den Einstellungs- bzw. Positionierungsfragen feststellbar ist, so ist es doch über alle Staffeln so, dass die Westberliner immer die klarer entschiedenen Antworten geben. Das ist ein Ergebnis, das ins gängige Ost-West-Klischee passt, das die Ostdeutschen generell als zaghafter, ausgleichender und unentschlossener als die Westdeutschen klassifiziert. Jedoch zeichnet sich auch bei den Ostberlinern eine Tendenz zu bipolaren Antworten ab.

Die die ganze Stadt '94 einende eher reservierte Grundeinstellung der anderen Stadthälfte gegenüber wird nach '96 zunehmend aufgegeben. Die Westberliner ergänzen ihre ablehnende Haltung um positive Positionierungen, die nach '98 deutlich ausgebaut werden. Das impliziert, dass klar wahrgenommene Differenzen nicht mehr ablehnend funktionalisiert werden. Die Ostberliner werden entschiedener in den Sachantworten, behalten aber ihre Zurückhaltung bei der Akzeptanz des Sprachverhaltens der Westberliner bei (was sicher auch damit zusammenhängt, dass sie mehrheitlich der Meinung sind, die Akzeptanz ihrer identitätsstiftenden Stadtvarietät habe nach '89 abgenommen). Sie gestehen (sich) also zunehmend Ost-West-Unterschiede ein, ohne diese jedoch bewerten zu wollen.

Eine positive Einstellung dem in der anderen Stadthälfte wahrgenommenen Sprachverhalten gegenüber ist 1994/96 eher im Ostteil zu finden, 1998 dagegen ziemlich eindeutig im Westteil.

Dieser ostwestunterschiedene Befund, der durch die Resultate einer LCA zu den Sozialkontakten zwischen Ost- und Westberlinern soziologisch untermauert wird (Regener 2000, S. 48-64), bestätigt die Existenz einer Abgrenzungsidentität als eine der Dimensionen der in den 1990er-Jahren entwickelten eigenen ostdeutschen Identität aus der soziolinguistischen Perspektive.

6 Resümee

In der Gesamtschau aller Ergebnisse ergibt sich folgendes Bild von sprachlich gestützten Identitätsfindungsprozessen in Berlin am Ende der 1990er-Jahre:

1. Die These vom sinkenden Kommunikationswert der Stadtvarietät ist obsolet. Im Ostteil der Stadt blieb der Gebrauch des Berlinischen auf relativ hohem Niveau stabil; im Westteil wurde der tendenzielle Dialektverlust der letzten Jahrzehnte gestoppt, der Gebrauch des Berlinischen stabilisierte sich auf niedrigerem Niveau.

2. Damit einher ging ein stadteinheitlicher Trend zur positiveren Einstellung der Stadtvarietät generell gegenüber, der jedoch für Ost und West unterschiedlich zu spezifizieren ist: Die westberliner Sprachgemeinschaft gab ihre extremer ablehnenden Positionen z. T. auf, und zwar zugunsten vorrangig neutraler, aber auch direkt positiver Positionierungen. Die ostberliner Sprachgemeinschaft verließ genau diese für die Nach-Anschlusszeit charakteristischen neutralen Positionen, um sich entschiedener zu artikulieren.

3. Es existierte nach wie vor ein deutliches und auch deutlich wertendes Differenzbewusstsein zwischen Ost und West hinsichtlich des Gebrauchs und der Bewertung der Stadtvarietät. Darin spiegeln sich m. E. die differenten soziokulturellen Entwicklungen, die die beiden Stadthälften in der Zeit ihrer Teilung durchlaufen haben. Wenn dies richtig ist, dann resultieren Spracheinstellungen eher aus situativen als aus allgemeinen Wertorientierungen sozialer Gruppen.

4. Die im Sprachverhalten wahrgenommenen Unterschiede wurden mehrheitlich auf die Sprecher in der jeweils anderen Stadthälfte distinktiv projiziert.

5. Generell beobachtbar sind zwei Tendenzen:
   Zum einen existierte ein stadteinheitlicher Hang zu Unbestimmtheit, Nicht-Festlegen-Wollen, Unentschiedenheit; vor allem bei Bewertungsfragen.
   Zum anderen nahmen auch eindeutig positive oder negative Positionierungen zu.

6. Auch diese beiden Tendenzen sind für Ost bzw. West spezifisch zu differenzieren:
   Die westberliner Sprachgemeinschaft schien zwar gefestigter in ihren Haltungen, Einstellungen, Werten und Urteilen als die ostberliner und deshalb änderungsresistenter oder auch konservativer. Dennoch gab es eine Bewegung hin zu positiverer Interessiertheit am Ostpendant. Diese westberliner Entwicklung reflektierte die Normalität eines an seinem Beginn von Vorurteilen und Stereotypen belasteten Kennenlern- und Annäherungsprozesses.
   Die ostberliner Sprachgemeinschaft war einerseits in ihrem Selbstverständnis relativ verunsichert und irritiert worden. Das materialisierte sich in einer hohen Unsicherheitsquote in den Antworten. Andererseits gewann sie aber zunehmend ihre Entscheidungsfähigkeit zurück, wobei beide Pole des Positiv-negativ-Spektrums bedient wurden. Diese ostberliner Entwicklung reflektierte die

Normalität eines an seinem Beginn von übergroßen Erwartungen belasteten Kennenlern- und Annäherungsprozesses, der u. a. wegen dieser Überfrachtung partiell in Abgrenzung umschlug.

7. Die empirisch nachweisbaren Unterschiede sowohl im Sprachverhalten als auch in den Spracheinstellungen indizieren deutlich, dass die berliner Sprachgemeinschaft am Ende der '90er-Jahre nach wie vor inhomogen war.
   Aufgrund einer Zwischenauswertung von Ergebnissen der beiden ersten Querschnitte (Schmidt-Regener 1998a, 1998b) hatte der Schluss nahegelegen, dass die Beschreibung über Solidaritäts- vs. Prestigemodell die veränderten Bedingungen nicht mehr adäquat abbildet, dass die Ostberliner dem über den sprachlichen Markt ausgeübten Anpassungsdruck – notwendigerweise selbstverständlich – zunehmend nachgeben und dass die Westberliner jedenfalls etwas von der Stigmatisierung des Berlinischen abrücken; dass so etwas wie ein softes Prestigemodell die gemeinsame sprachliche Zukunft der Stadt beschreiben könnte.
   Diese Annahme wurde durch die dritte Erhebungsstaffel nicht bestätigt. Am Ende der 1990er-Jahre hatten die Ostberliner einen Teil ihres Selbstbewusstseins und damit auch ihrer sprachlichen Autonomie zurückgewonnen, weil sie ihre eigene Sprache – zunehmend wieder – als einen der zentral identitätsstiftenden Werte wahrnahmen bzw. entdeckten und auch dringend brauchten.
   Die Westberliner gewannen tatsächlich ein entspannteres, aber dennoch von dem der Ostberliner um Welten entferntes Verhältnis zum nichtprestigebesetzen Sprachgebrauch und, unmittelbar damit verbunden, zur nichtprestigeträchtigen Stadtvarietät.

Modellierung von längsschnittlichen Daten am Beispiel einer quasi-experimentellen Studie zur Erfassung von naturwissenschaftlichen Kompetenzen im Kindergartenalter

Claus H. Carstensen, Eva-Maria Lankes & Mirjam Steffensky

1 Theoretischer Hintergrund

1.1 Naturwissenschaftliche Kompetenz im Kindergartenalter

Frühkindlichen Bildungsprozessen im Bereich Naturwissenschaften wird seit einiger Zeit große Bedeutung beigemessen (KMK 2004, 2009); über die Ausprägung naturwissenschaftlicher Kompetenz im frühen Alter sowie über Möglichkeiten der Beeinflussung und Förderung ist jedoch noch wenig bekannt. Im DFG-Projekt Studie zur naturwissenschaftlichen Kompetenzentwicklung im Elementarbereich (SNaKE: Carstensen et al. 2011, Steffensky et al. 2012a) wird in einer quasi-experimentellen Studie untersucht, was fünfjährige Kinder über Naturwissenschaften wissen, wie sich dieses Wissen über einen Zeitraum von neun Monaten entwickelt und wie die Entwicklung durch verschiedene Lernangebote – als quasi-experimentelle Variation - unterschiedlich gefördert werden kann.

Naturwissenschaftliche Kompetenz wird dabei im Sinne von Konzeptionen zu scientific literacy (KMK 2004, Bybee et al. 2009, Bos et al. 2008, NSES 1996, QCA 2000, GDSU 2002, Osborne & Dillon 2008) verstanden: Wissen, als 'verstandenes' Wissen, wird differenziert in naturwissenschaftliches Wissen und Wissen über Naturwissenschaften. Im Zentrum des naturwissenschaftlichen Wissens steht das Verständnis zentraler Konzepte und Theorien. Zum Wissen über Naturwissenschaften gehören ein Verständnis naturwissenschaftlicher Denk- und Arbeitsweisen sowie ein Verständnis der Naturwissenschaft als Disziplin (Nature of Science), das Wissen über die Rolle der Naturwissenschaften in unserer Gesellschaft und über das Verhältnis zwischen Naturwissenschaften und Technik einschließt (Norris & Phillips 2003). Naturwissenschaftliche Kompetenz umfasst neben Wissenskomponenten und der Fähigkeit zur Anwendung dieses Wissens in lebensnahen Kontexten auch nicht-kognitive Komponenten, wie beispielsweise motivationale oder affektive Aspekte naturwissenschaftlicher Bildung (Bybee 1997, Bybee et al. 2009, Norris & Phillips 2003). Die Entwicklung von scientific literacy ist ein kumulativer Prozess, der in der frühen Kindheit beginnt und im Verlauf der Kindergarten- (in

Ansätzen) und Schulzeit (stärker systematisch) weiterentwickelt wird (Prenzel 2000).

Ziele der Studie

Die Studie SNaKE untersucht, inwieweit sich naturwissenschaftliche Kompetenzen von fünfjährigen Kindern erfassen und durch eine Intervention verändern lassen. In einer quasi-experimentellen Variation erhielten die Kinder unterschiedliche Lernangebote, in denen Experimente durchgeführt und alltagsnahe Erfahrungen aufgegriffen wurden. Die Studie implementiert drei unterschiedliche Lernangebote, das Reflektieren von Erfahrungen aus dem Alltag, das Durchführen von der Altersstufe angemessenen naturwissenschaftlichen Experimenten und eine Kombination aus beiden Bedingungen (Steffensky et al. 2012b). Mit der Studie sollen die Auswirkungen der Lernangebote auf die naturwissenschaftliche Kompetenz der Kinder und ihr Interesse an Naturwissenschaften untersucht werden. Dazu wurden geeignete Erhebungsinstrumente für fünfjährige Kinder entwickelt. Der Kompetenztest (vgl. Carstensen et al. 2011) wird in Kapitel 2 näher erläutert; zur Erfassung des Interesses an Naturwissenschaften wurde eine Befragung entwickelt (Nölke et al. 2011). In diesem Beitrag betrachten wir die Frage der längsschnittlichen Modellierung der Kompetenztestergebnisse aus einem Vortest, einem Nachtest nach etwa drei Monaten und einem Follow-up-Test nach weiteren vier Monaten.

1.2 Modellierung von längsschnittlichen Daten

In der probabilistischen Testtheorie werden zahlreiche Modelle für Messwiederholungen diskutiert, eine Übersicht gibt Meiser (2007). In Messwiederholungen erhobene Datenpunkte sind neben einer Zuordnung zu einer antwortenden Person und zu einem Item weiterhin durch die Zuordnung zu einem Erhebungszeitpunkt gekennzeichnet. Im Vergleich zu Modellen für Antwortdaten aus Einmal-Messungen müssen solche für wiederholte Messungen neben den Eigenschaften der untersuchten Personen und der verwendeten Items auch die Messzeitpunkte einbeziehen. Daher diskutiert Rost (2004) dreifaktorielle Verallgemeinerungen des Rasch-Modells (Rasch 1960) zur Modellierung von Antwortdaten auf Messwiederholungen, in welchen die beiden Faktoren Personen und Items um den dritten Faktor Messzeitpunkt ergänzt werden. Für dichotome Antworten spezifiziert Rost ein System von acht dreifaktoriellen Modellen, mit denen die Wahrscheinlichkeit einer Person v, das Item i zum Zeitpunkt t zu lösen,

$$\text{prob}(X_{vit}) = \frac{\exp(\alpha_{vit})}{1+\exp(\alpha_{vit})} \tag{1a}$$

durch einen dreifaktoriell konstruierten Exponenten α_{vit} spezifiziert wird. Da im vorliegenden Beitrag ein Modell für ordinale Antwortdaten abgeleitet werden soll, werden hier zusätzlich ordinale Antwortkategorien j unterschieden und es resultiert das Modell

$$\text{prob}(X_{vijt}) = \frac{\exp(\alpha_{vijt})}{1+\exp(\alpha_{vijt})}. \quad (1b)$$

Auf eine geeignete Restriktion der Kategorienparameter wird in Kapitel 2 eingegangen. Das restriktivste dieser dreifaktoriellen Modelle (entsprechend dem Modell (1) in der Systematik bei Rost)

$$\alpha_{vijt} = \theta_v + \sigma_{ij} + \delta_t \quad (2)$$

ergibt sich aus der Annahme, dass alle Effekte des Messzeitpunkts δ unabhängig sowohl von den Itemschwierigkeiten σ als auch von den Personenfähigkeiten θ sind. Damit lassen sich Veränderungen zwischen Zeitpunkten beschreiben, die für alle Personen identisch sind. Diese Annahme ist für die vorliegende Studie, wie auch für die allermeisten Studien, unrealistisch. Soll etwa untersucht werden, ob sich die Veränderungen in der Kompetenz der Kinder nach Experimentalgruppen unterscheiden, ist ein Modell für personengruppenspezifische Veränderungen über die Zeit erforderlich. Für die Studie SNaKE ist es von Interesse, personenspezifische Änderungen über die Zeit abbilden zu können, um Unterschiede in den Kompetenzzuwächsen beschreiben zu können. Ein entsprechendes dreifaktorielles Antwortmodell muss deshalb Interaktionen zwischen Zeitpunkten und Personen spezifizieren, in der Notation von Rost wird dies durch einen doppelt indizierten Parameter δ_{vt} gekennzeichnet. Mit der Annahme von personenspezifischen Veränderungen folgt das Modell (entsprechend Modell (3) bei Rost)

$$\alpha_{vijt} = \sigma_{ij} + \delta_{vt}, \quad (3)$$

das eine eindimensionale latente Fähigkeit definiert, die für jede Person zu jedem Zeitpunkt eine andere Ausprägung annehmen kann. Weiterhin wird angenommen, dass sich die Itemschwierigkeiten über die Zeit nicht verändern, was die Invarianz des Messmodells bedeutet.

Mit Blick auf unterschiedliche Itemtypen, etwa Aufgaben zu alltagsnahen versus weniger alltagsnahen Phänomenen, ließe sich stattdessen annehmen, dass auch Itemschwierigkeiten sich über die Zeit verändern. Damit werden Veränderungen des Messmodells zwischen den Zeitpunkten eingeführt, in der Notation von Rost wird dies durch eine Interaktion von Zeitpunkt und Item durch einen doppelt indizierten Parameter σ_{it} ausgedrückt. Ohne die Annahme von Messinvarianz über die Zeitpunkte folgt das Modell (entsprechend Modell (5) bei Rost)

$$\alpha_{vijt} = \sigma_{ijt} + \delta_{vt}. \quad (4)$$

Grundsätzlich sind beide Annahmen für die vorliegende Studie plausibel. Zum einen werden personenspezifische Veränderungen der untersuchten Kompetenzen erwartet; in den Verteilungen der Parameter δ_{vt} zu jedem Zeitpunkt bilden sich Veränderungen in der mittleren Kompetenz und in der Varianz der Kompetenzen ab. Zum anderen kann nicht ausgeschlossen werden, dass sich die Schwierigkeiten

der Aufgaben über die Zeit in unterschiedlichem Ausmaß verändern. Ohne eine weitere Restriktion wie die Annahme von Messinvarianz (3) sind die Ergebnisse dieses Modells im Grunde nicht interpretierbar. Jedoch eignet es sich zum Vergleich mit (3). Lässt sich die Geltung von (3) gegenüber (4) belegen, so können unter der Annahme von Messinvarianz personenspezifische Veränderungen derselben Kompetenz über die Zeit analysiert werden.

Die Modelle (3) und (4) spezifizieren eine Personeneigenschaftsausprägung für jeden Zeitpunkt und lassen sich auch als restringierte mehrdimensionale Testmodelle darstellen. Am Beispiel einer wiederholten Befragung von Erwachsenen mit demselben Instrument stellt Andersen (1985) ein mehrdimensionales Rasch-Modell vor, welches mit seinen Dimensionen die Eigenschaften der Personen zu unterschiedlichen Messzeitpunkten abbildet. Durch eine Gleichsetzung der Itemschwierigkeiten aus den Messwiederholungen wird die Messinvarianz des Instruments angenommen, so dass Unterschiede in den Dimensionen der Personenfähigkeit allein auf personenspezifische Veränderungen über die Zeit zurückgehen. Dies entspricht Modell (3) in der Notation von Rost für den Fall eines einzigen Testinstruments zu allen Zeitpunkten und einer mehrdimensionalen Modellierung der Personeneigenschaften. Das Modell erlaubt es, unterschiedliche Personenausprägungen zu jedem Zeitpunkt zu analysieren. Zur Schätzung der Parameter des Modells stellt Andersen einen Algorithmus zur Maximierung der bedingten Likelihood (CML) vor.

Zur Abbildung von Lerneffekten und Veränderungen stellt Embretson (1991) ein Modell vor, welches die Veränderungen der Personen zwischen je zwei Messzeitpunkten latent modelliert. Neben einer Ausgangsfähigkeit θ_v hängt die Lösungswahrscheinlichkeit von diesen Veränderungen (modifiabilities) δ_{vt} ab. In der Notation von Rost ergibt sich die Wahrscheinlichkeit einer richtigen Lösung zum ersten Zeitpunkt (t = 0) als

$$\alpha_{vijt} = \theta_v + \sigma_{ij}\,, \tag{5}$$

und zum Zeitpunkt t als

$$\alpha_{vijt} = \theta_v + \sum_{u=1}^{t} \delta_{vu} + \sigma_{ij}\,. \tag{6}$$

Auch dies ist ein mehrdimensionales Rasch-Modell, welches aus einer Ausgangsfähigkeit und den eindimensionalen Modifiabilities zusammengesetzt ist.

Im Gegensatz zu Andersen (1985) verwendet Embretson unterschiedliche Aufgaben bei jeder Messwiederholung, um Wiederholungseffekte zu vermeiden. Sie präsentiert ein Erhebungsdesign für mehrere Personengruppen, wobei jede Gruppe einen Block gemeinsamer Ankeraufgaben (zum ersten Testzeitpunkt) bearbeitet und daneben zu jedem Testzeitpunkt einen weiteren Aufgabenblock vorgelegt bekommt, so dass keine Aufgaben wiederholt und alle Aufgabenblöcke zu jedem Testzeitpunkt eingesetzt werden. Zur Schätzung der Modellparameter verwendet

Embretson ein CML-Verfahren: Unter der Annahme von Messinvarianz für jeden Aufgabenblock über die Messzeitpunkte werden die Itemparameter über die verschiedenen Testzeitpunkte gleichgesetzt und können mit Schätzalgorithmen für mehrere Gruppen (vgl. das LLTM von Fischer & Formann 1972 oder Loglin von Whitely & Nieh 1981) ermittelt werden. Für Personenparameter werden ML-Schätzer unter Berücksichtigung der mehrdimensionalen Fähigkeitsverteilung angegeben.

2 Fragestellung und Methode

Über naturwissenschaftliche Kompetenzen von fünfjährigen Kindern liegen bislang wenige Studien vor; im Projekt konnte gezeigt werden, dass sich der zur Erhebung von naturwissenschaftlicher Kompetenz entwickelte Test zum ersten Messzeitpunkt als homogen im Sinne eines Rasch-Modells für ordinale Antworten zeigt (Carstensen et al. 2011). Für das Projekt SNaKE stehen Fragen nach Veränderungen der naturwissenschaftlichen Kompetenz der Kinder über drei Messzeitpunkte im Zeitraum von sieben Monaten sowie nach der Abhängigkeit der Veränderungen von den Gruppen der quasi-experimentellen Variation im Vordergrund. Folgende Fragen werden untersucht:

1. Stellen auch die Tests zum zweiten und dritten Messzeitpunkt homogene Skalen gemäß eines Ordered-Partition-Modells dar, wie es für den Test zum ersten Messzeitpunkt in Carstensen et al. (2011) gezeigt wurde?
2. Gelingt es, mit den Tests zu den drei Messzeitpunkten, in der naturwissenschaftlichen Kompetenz der Kinder Kompetenzen und deren Veränderungen auf einer Rasch-homogenen Skala abzubilden?
3. Wie verändern sich die Kompetenzen der Kinder über die drei Messzeitpunkte?

Im weiteren Verlauf des Projekts wird es darum gehen, welche Unterschiede es in den Kompetenzveränderungen der Kinder nach den Experimentalgruppen gibt und welche Wirkung die verschiedenen Lernangebote in der Interventionsphase haben. Weiter soll untersucht werden, auf welche Faktoren sich differentielle Veränderungen der Aufgabenschwierigkeiten möglicherweise zurückführen lassen und ob Zusammenhänge zwischen Veränderungen und den Lernangeboten bestehen. Diese Fragen werden in weiterführenden Arbeiten untersucht (vgl. Steffensky et al. 2012b), während der vorliegende Beitrag die längsschnittliche Modellierung fokussiert.

Kompetenztest

Im Rahmen des Projekts wurde ein Test zur Erfassung naturwissenschaftlicher Kompetenzen für ausgewählte Inhalte im Elementarbereich entwickelt, mit dem sowohl die Fähigkeiten von Kindern dieser Altersstufe abgebildet werden als auch Entwicklungen der Kompetenz nach den Treatments und über einen Zeitraum von sieben Monaten verfolgt werden können. Der Test wurde zum Thema Wasser konstruiert; die Aufgaben erfassen zum einen Kenntnisse naturwissenschaftlicher Be-

griffe und Konzepte sowie zum anderen Wissen über naturwissenschaftliche Denk- und Arbeitsweisen. Auf der Ebene von naturwissenschaftlichen Konzepten werden Schmelzen und Gefrieren, Verdunsten/Verdampfen und Kondensieren sowie Lösen unterschieden.

Als Format für den Test wurde ein strukturiertes Interview gewählt. Zunächst wurden über die Kompetenz- und Inhaltsbereiche verteilt 62 Multiple-Choice-Aufgaben, 32 offene Aufgaben, 17 dichotome MC-Aufgaben und 11 Zuordnungsaufgaben entwickelt, wobei 79 der 122 Aufgaben Bilder zur Veranschaulichung und Fokussierung enthalten (Steffensky et al. 2012a). Die Auswertung der Aufgaben erfolgte durch die Interviewer vor Ort und konnte anhand eines Tonbandmitschnitts der Interviews überprüft werden. Die Aufgaben wurden mit wenigen Kindern in cognitive labs präpilotiert und nach Überarbeitung in einer Erhebung mit etwa 30 Kindern je Aufgabe pilotiert. Für den ersten Messzeitpunkt wurde ein Test mit 32 Items zusammengestellt, von denen letztlich 29 in eine Kompetenzskalierung eingingen (vgl. Carstensen et al. 2011). Der Test für den zweiten Messzeitpunkt enthielt 18 Items vom ersten Messzeitpunkt und 14 im Mittel etwas schwerere Aufgaben; der für den dritten umfasste 18 Aufgaben vom zweiten Messzeitpunkt und 14 neue Items. Die Tests vom ersten und dritten Messzeitpunkt haben 15 Aufgaben, alle drei Tests haben 13 Aufgaben gemeinsam. Eine Übersicht über die Zahl der gemeinsamen und der messzeitpunktspezifischen Aufgaben gibt Tabelle 1.

	Aufgaben	t1	t2	t3
Gemeinsam zu t1, t2 & t3	13	13	13	13
Gemeinsam zu t1 & t2	4	4	4	
Gemeinsam zu t1 & t3	2	2		2
Gemeinsam zu t2 & t3	5		5	5
Spezifisch zu t1	10	10		
Spezifisch zu t2	7		7	
Spezifisch zu t3	11			11
Gesamtzahlen	52	29	29	31

Tabelle 1: Anzahl der Aufgaben insgesamt sowie gemeinsamer und messzeitpunktspezifischer Aufgaben in den drei Testungen (bezeichnet mit t1, t2 und t3)

Die wenigen Studien, die sich mit der naturwissenschaftlichen Kompetenz von Vorschulkindern befassen, basieren überwiegend auf zeitintensiven Interviews oder Beobachtungen (vgl. Carstensen et al. 2011); effiziente und ökonomische Erhebungsinstrumente, wie sie in der Studie SNaKE entwickelt wurden, liegen kaum vor (in Entwicklung ist ein Test für Scientific Literacy im NEPS, vgl. Weinert et al. 2011). Mit einem solchen Verfahren können nun größere Stichproben untersucht und komplexe statistische Verfahren, etwa zur längsschnittlichen Modellierung, eingesetzt werden. Damit entstehen aber neue Herausforderungen: Aus der Psychologie der kognitiven Entwicklung und der Forschung zur Entwicklung mathe-

matischer Kompetenzen ist bekannt, dass sich der Erwerb von Kompetenzen in dieser Altersstufe als ein Erwerb von Teil- und Vorläuferkompetenzen beschreiben lässt, der keineswegs linear und integriert verläuft (Weinert et al. 2008). Im Rahmen der SNaKE-Studie soll deshalb geklärt werden, ob sich die naturwissenschaftlichen Kompetenzen im Kindergartenalter auf (Rasch-)homogenen Skalen messen lassen und ob sich Entwicklungsverläufe auf einer eindimensionalen latenten Variable abbilden lassen.

Modellierung der Antwortdaten

Die Bewertung der Antworten erfolgte bei einigen einfachen Multiple-Choice-Aufgaben und bei einigen offenen Antwortformaten dichotom. Bei anderen offenen Aufgaben wurde ein Teilpunkt für ansatzweise richtige Lösungen gegeben; komplexe Multiple-Choice- und Auswahlaufgaben wurden aus der Summe der richtigen Antworten in drei Kategorien anhand beobachteter Häufigkeiten bewertet. Zur Skalierung des Tests vom ersten Messzeitpunkt wurde ein Ordered-Partition-Modell (OPM, Wilson 1992) spezifiziert, welches für die richtige Antwort bei dichotomen Items ebenso wie für die höchste Antwort bei Items mit drei Antwortkategorien einen Score von zwei Punkten annimmt. Für falsche, aber anschlussfähige Lösungen bei dreikategoriellen Items wird ein Punkt vergeben (Carstensen et al. 2011). Auf diese Weise werden richtige Lösungen in dichotomen und mehrkategoriellen Items gleich gewichtet, während falsche, aber anschlussfähige Lösungen niedriger gewichtet werden. Diese Modelle werden auch für den zweiten und dritten Messzeitpunkt geschätzt; anhand dieser Modelle werden ggf. Items ausgewählt und die Homogenität der Tests wird überprüft. Zur Kontrolle der globalen Modellgeltung werden ebenso wie für den ersten Messzeitpunkt jeweils für Geschlecht, hohe oder niedrige Scores im Naturwissenschaftstest und in einem Intelligenztest Zweigruppenmodelle mit einem Modell für alle Kinder verglichen.

In einem zweiten Schritt wird ein mehrdimensionales OPM spezifiziert, welches wie das Andersen-Modell die Tests zu den drei Messzeitpunkten als unterschiedliche Dimensionen spezifiziert. Jedoch werden nicht die Itemparameter der wiederholt eingesetzten Items auf Gleichheit restringiert, sondern es werden lediglich die Summen der Schwierigkeiten der gemeinsamen Items zwischen dem ersten und zweiten sowie dem zweiten und dritten Messzeitpunkt gleichgesetzt. Das entspricht einem Equating der Testschwierigkeiten, ohne jedoch dasselbe Messmodell für jeden Test anzunehmen (vgl. Kolen & Brennan 2004). Zur Identifizierung des Modells wird weiterhin der Mittelwert der ersten Dimension auf Null fixiert. Dieses Ordered-Partition(OP)-Equating-Modell entspricht Modell (4) und damit dem fünften der dreifaktoriellen Rasch-Modelle nach Rost (2004) mit zeitspezifischen Veränderungen der Personen und der Items. Aufgrund der beiden Restriktionen der Summen der Schwierigkeiten der gemeinsamen Items hat dieses OP-Equating-Modell genauso viele Parameter wie Modell (4). Damit ist auch die Gültigkeit des Equating-Modells bereits nachgewiesen, sobald für jeden Zeitpunkt die Geltung eines OPM gezeigt werden kann.

Das OP-Equating-Modell dient in einem dritten Schritt dem Vergleich mit einem verallgemeinerten Andersen- und einem verallgemeinerten Embretson-Modell. Für das OP-Andersen- und das OP-Embretson-Modell werden die Parameter der jeweils zwei oder drei Tests gemeinsamen Items gleichgesetzt. Damit bilden die drei Dimensionen jeweils Veränderungen der Personenfähigkeiten über die Zeit auf derselben latenten Variablen ab. Zur Identifizierung wird die Schwierigkeit eines der 13 allen Zeitpunkten gemeinsamen Items auf Null fixiert.

Alle erwähnten Modelle, das Equating-, das Andersen- und das Embretson-Modell, jeweils als Verallgemeinerung des Ordered-Partition-Modell, lassen sich als Spezialfälle des Multidimensional Random Coefficient Multinomial Logit Modell (MRCML) (Adams et al. 1997) darstellen und werden mit der Software Conquest (Wu et al. 2007) durch Maximierung einer marginalen Likelihood (MML) geschätzt. In der Parametrisierung des MRCML werden für ordinale Antworten Itemlokationsparameter σ_i^* und Distanzparameter τ_{ij}, für die Distanzen der Schwellen zum Lokationsparameter geschätzt, die sich bei dreikategoriellen Antworten durch folgende Beziehung in Schwellenparameter des Partial-Credit-Modells σ_{ij} (Masters 1982) überführen lassen:

$$\sigma_{ij} = \sigma_i^* + \tau_{ij} \quad \text{mit} \quad \tau_{i2} = -\tau_{i1} \quad \text{für} \quad j \in \{1,2\}. \tag{7}$$

Im OP-Andersen- und im OP-Embretson-Modell werden lediglich Lokationsparameter σ_i^* der allen Zeitpunkten gemeinsamen Items gleichgesetzt, nicht aber die Distanzparameter τ_{ij}, um Veränderungen im Antwortformat zuzulassen.

Die zu schätzenden Modelle sind insofern komplex, als dass an drei Testzeitpunkten insgesamt 51 Aufgaben teils ein Mal und teils mehrfach vorgegeben wurden, so dass von jedem Kind 89 Aufgabenbearbeitungen beobachtet wurden. Damit gleicht das Erhebungsdesign weder dem von Andersen (1985) noch dem von Embretson (1991). Da die Aufgaben wiederholt eingesetzt wurden, kann unser Erhebungsdesign als eine Erweiterung des Designs von Andersen um Aufgaben gesehen werden, die nicht allen Zeitpunkten gemeinsam beziehungsweise zeitpunktspezifisch sind.

Die hier spezifizierten verallgemeinerten Modelle nach Andersen und Embretson definieren dieselben Itemparamter und dieselbe Zahl von Populationsparametern, je einen Mittelwert für die zweite und dritte Dimension, drei Varianzen und drei Kovarianzen. Die Dimensionen beider Modelle unterscheiden sich in ihrer Bedeutung, als Kompetenzen zu jedem Messzeitpunkt im Andersen-Modell und als Ausgangskompetenz und Veränderungen gegenüber dem vorangegangenen Messzeitpunkt im Embretson-Modell. Da sich ihre latenten Variablen durch Addition ineinander überführen lassen, sind sie mathematisch äquivalent und müssen in der Parameterschätzung dieselbe Likelihood aufweisen, obwohl es empirisch wegen sehr kleiner Varianzen von Veränderungen zum Beispiel zu unterschiedlichen Likelihoods kommen könnte. Embretson (1991) merkt an, dass die Varianz der Modifiabilities abhängig von der Diskrimination der Items für die Veränderungen zwischen zwei Testungen ist und das Modell entsprechend unterschiedliche Veränderungssensitivitäten der Tests zwischen den Messzeitpunkten abbilden kann. Im Andersen-Mo-

dell wird hingegen angenommen, dass die Diskrimination der Items zu allen Testzeitpunkten gleich ist, was im Erhebungsdesign bei Andersen (1985) und in der vorliegenden Studie durch den Einsatz von Ankeritems zu allen Testzeitpunkten gewährleistet ist. Da sich aus den Varianzen und Kovarianzen im Andersen-Modell die Varianzen der Veränderungen im Embretson-Modell zurückrechnen lassen, kann auch das Andersen-Modell unterschiedliche Änderungssensitivitäten zwischen den Messzeitpunkten abbilden.

3 Ergebnisse

3.1 Homogenität der Tests

Bereits für den ersten Messzeitpunkt wurden Modellvergleiche zur Überprüfung der Personenhomogenität berichtet. Entsprechend dem Rational eines Andersen-Tests (Andersen 1973) werden die Kinder nach ihrer naturwissenschaftlichen Kompetenz in potenziell heterogene Gruppen unterteilt. Im Unterschied zum Andersen-Test werden die Parameter nicht durch eine Maximierung der bedingten Likelihood geschätzt, sondern durch die Maximierung einer marginalen Likelihood (MML). Wegen der großen Zahl von unterschiedlichen Antwortmustern bei 29 beziehungsweise 31 zwei- oder dreikategoriellen Items ergeben sich für die saturierten Modelle Parameterzahlen, die Signifikanzprüfungen von Likelihoodunterschieden unmöglich machen. Zum Vergleich der Modelle wird daher das Bayes-Informationskriterium (BIC; Schwarz 1978) herangezogen. In den Tabellen 2 und 3 werden für den zweiten und den dritten Messzeitpunkt Vergleiche zwischen Jungen und Mädchen, zwischen Kindern mit hohem vs. niedrigen Score im Naturwissenschaftstest sowie in einem Untertest des CFT-1 (Cattell et al. 1997) zu figuralen Analogien dargestellt. Für jeden Vergleich wird ein Modell mit der Gruppenunterscheidung als Haupteffekt (M2, M4 und M6) mit einem Modell mit dem Haupteffekt sowie den Interaktionen zwischen Gruppe und Itemschwierigkeiten kontrastiert.

	Modell	Deviance	n(P)	BIC
M1	Gesamtmodell	9282.7	35	9476.9
M2	Gesamtmodell & Geschlecht	9279.3	36	9479.1
M3	… & geschlechtsspezifische Itemschwierigkeiten	9252.3	65	9613.0
M4	Gesamtmodell & NaWi-Gruppen	9031.4	36	9231.2
M5	… & NaWi-gruppenspezifische Itemschwierigkeiten	8953.7	65	9314.4
M6	Gesamtmodell & CFT-Gruppen	9256.7	36	9456.4
M7	… & CFT-gruppenspezifische Itemschwierigkeiten	9227.8	65	9588.5

Anmerkung: Deviance = -2*ln(L), BIC = Deviance + ln(n)*n(P), n = 257 für alle Modelle.
ln(L) = Log.Likelihood, n(P) = Anzahl Parameter, n = Anzahl Vpn

Tabelle 2: Modellvergleiche für den zweiten Messzeitpunkt anhand des BIC

	Modell	Deviance	n(P)	BIC
M1	Gesamtmodell	10058.2	37	10263.5
M2	Gesamtmodell & Geschlecht	10057.1	38	10268.0
M3	… & geschlechtsspezifische Itemschwierigkeiten	10004.4	69	10387.3
M4	Gesamtmodell & NaWi-Gruppen	9778.6	38	9989.4
M5	… & NaWi-gruppenspezifische Itemschwierigkeiten	9672.0	69	10054.9
M6	Gesamtmodell & CFT-Gruppen	10036.0	38	10246.8
M7	… & CFT-gruppenspezifische Itemschwierigkeiten	10005.2	69	10388.1

Anmerkung: Deviance = -2*ln(L), BIC = Deviance + ln(n)*n(P), n = 257 für alle Modelle.
ln(L) = Log.Likelihood, n(P) = Anzahl Parameter, n = Anzahl Vpn

Tabelle 3: Modellvergleiche für den dritten Messzeitpunkt anhand des BIC

Die Modellierung von geschlechtsspezifischen Itemparametern zum zweiten Messzeitpunkt (M3 in Tab. 2) erfordert 65 Parameter, ein Modell ohne geschlechtsspezifische Parameter (M2 in Tab. 2) nur 36. Der Unterschied in der logarithmierten Likelihood beträgt 13.5 beziehungsweise 27.0 in der Testgröße für Likelihoodvergleiche -2*ln(L), in der Tabelle mit Deviance bezeichnet. Nach dem BIC rechtfertigt der Likelihoodunterschied die zusätzlichen Parameter nicht, so dass die Annahme geschlechtsspezifischer Itemschwierigkeiten verworfen werden kann. Entsprechendes gilt für die Kinder mit hohem versus niedrigem Score in der naturwissenschaftlichen Kompetenz sowie für die Gruppen nach dem Untertest 4 des CFT-1. Vergleichbare Befunde zeigen sich für den dritten Messzeitpunkt und wurden bereits für den ersten berichtet (Carstensen et al. 2011). Damit kann für alle drei Tests die Gültigkeit des OPM und Rasch-Homogenität angenommen werden.

Am Vergleich der Modelle M2, M4 und M6 mit M1 lässt sich ablesen, ob die Gruppenunterscheidungen im Sinne von explanativen Antwortmodellen (DeBoeck & Wilson 2004) zur Erklärung der Kompetenzverteilungen beitragen. Eine Unterscheidung nach den Geschlechtern zeigt keine bessere Modellgeltung als das Gesamtmodell (M1 mit M2, in den Tab. 2 und 3), d. h. das Geschlecht klärt keine Varianz in den naturwissenschaftlichen Kompetenzen der Kinder auf. Die Unterscheidung nach hohem oder niedrigem Score im Naturwissenschaftstest (Vergleich M1 mit M4) und im CFT-1 (Vergleich M1 mit M6) stellt eine Modellverbesserung dar, d. h. diese Gruppen unterscheiden sich in ihren naturwissenschaftlichen Kompetenzen. Für Kinder mit hohem versus niedrigem Score im Naturwissenschaftstest ist dies zu erwarten. Aus dem Ergebnis, dass die Gruppen nach dem CFT-1 sich in ihren naturwissenschaftlichen Kompetenzen unterscheiden, folgt ferner, dass ein Zusammenhang zwischen dem Kompetenztest und dem CFT-1 besteht. Dies gilt für alle drei Messzeitpunkte.

3.2 Passung des längsschnittlichen Modells

Zur Modellierung der Kompetenzen über die Zeit soll geprüft werden, ob diese Veränderungen durch ein Andersen- oder ein Embretson-Modell auf der Grundlage

der Ordered-Partition-Modelle abgebildet werden können. Zur Prüfung der beiden Modelle wird deren Modellanpassung mit der des Equating-Modells verglichen, von dessen Geltung ausgegangen wird, da die Geltung der Messmodelle zu jedem Zeitpunkt bereits gezeigt wurde.

Zur MML-Schätzung der Modellparameter mit der Software Conquest (Wu et al. 2007) wurde ein Quadratur-Verfahren zur Approximation der Personenfähigkeitsverteilung verwendet. Die Schätzung war mit einer Zahl von 25 Knotenpunkten je Dimension, also mit 15625 Knotenpunkten für die gemeinsame multivariate Normalverteilung der drei Dimensionen, erfolgreich. Mit der Standardeinstellung von 15 Knotenpunkten je Dimension ließ sich für das OP-Andersen-Modell hingegen keine Konvergenz beobachten. Für eine hilfsweise eingesetzte weniger restriktive Modellvariante, in welcher nur die 13 allen drei Zeitpunkten gemeinsamen Items gleichgesetzt wurden, waren bei 15 Knotenpunkten deutliche Unterschiede zwischen dem Andersen- und dem Embretson-Modell in Likelihood und Fähigkeitsmittelwertparametern zu beobachten.

Modellvergleich	n	n(P)	Deviance	BIC
OP-Equating-Modell	257	115	29208.7	29846.9
OP-Andersen-Modell (13 Items restr.)	257	91	29285.5	29790.4
OP-Embretson-Modell (13 Items restr.)	257	91	29285.7	29790.6
OP-Andersen-Modell (24 Items restr.)	257	80	29312.3	29756.3
OP-Embretson-Modell (24 Items restr.)	257	80	29312.5	29756.4

Tabelle 4: Modellvergleich des Equating- und des Andersen-Modells für drei Messzeitpunkte und des Embretson-Modells für Veränderungen über drei Messzeitpunkte anhand des BIC

In Tabelle 4 sind die Stichprobengröße, Parameterzahlen, Deviance und BIC-Statistik für alle drei Modelle angegeben. Im Modellvergleich werden beide Varianten von Gleichsetzungen aufgegriffen, zum einen unter Gleichsetzung der Schwierigkeiten der oben genannten 13 Items, zum anderen unter Gleichsetzung aller 24 zu zwei oder mehr Zeitpunkten eingesetzten Items (vgl. Tab. 1). Das Equating-Modell spezifiziert mit 115 Parametern 24 Parameter mehr als das Andersen- und das Embretson-Modell unter Restriktion der Schwierigkeiten von 13 Items. Diese Differenz ergibt sich durch die Gleichsetzungen der Schwierigkeiten von 39 Items (die zu drei Zeitpunkten verwendet wurden) auf 13 Parameter und die Schätzung zweier zusätzlicher Mittelwerte für den zweiten und dritten Messzeitpunkt. Durch die zusätzliche Restriktion der Schwierigkeiten von 11 weiteren Items reduziert sich die Zahl der Parameter auf 80. Die Geltung des Equating-Modells ist mit der Geltung der drei Modelle für die drei Testzeitpunkte nachgewiesen, da es über die Geltung dieser drei Modelle hinaus keine weiteren Annahmen formuliert. Im Folgenden werden daher relative Modellvergleiche anhand des BIC vorgenommen. Nach dem BIC zeigen beide Modelle (Andersen- und Embretson-Modell), die 13 Parameter gleichsetzen, eine bessere Passung zu den Daten als das Equating-Modell. Die beiden Modelle mit Gleichsetzung der Schwierigkeiten von 24 Items zeigen wiederum eine bessere Passung als die beiden mit Restriktionen auf 13 Items.

Der Unterschied zwischen dem Andersen- und dem Embretson-Modell ist in der Deviance sehr klein, so dass sie als gleichwertig gelten können. Demnach können die Schwierigkeiten aller mehrfach vorgegebenen Aufgaben gleichgesetzt werden, und es bestätigt sich die Annahme, dass mit den drei Tests eine einzige latente Variable gemessen wird.

Item		t2-t1	t3-t2
1	B1Ko130	-1.24	-0.51
7	A2Ao220	-0.90	0.47
6	C2Ao210	-0.85	-0.03
3	C1Km150	-0.75	0.26
10	C3Am330	-0.64	-0.23
5	B1Km190	-0.46	0.23
9	B2Kr290	-0.46	-0.48
4	A1Km180	-0.38	-0.29
12	B3Km370	-0.34	-0.10
11	A3Am340	-0.07	0.18
2	C1Ko140	-0.04	-0.61
13	B3Km390	0.08	-0.27
8	B2Km280	0.14	-0.45

Tabelle 5: Veränderungen der Aufgabenschwierigkeiten zwischen den Messzeitpunkten in Logits

Außer dem Vergleich werden zur Deskription der Modelle die Differenzen der Itemschwierigkeiten im unrestringierten Equating-Modell und die Itemfitwerte für die restringierten Items berichtet. In den Differenzen der Schwierigkeitsparameter der den Messzeitpunkten gemeinsamen Items zeigt sich eine gewisse Heterogenität: Die wenigsten von ihnen liegen nahe der mittleren Veränderungen über alle Aufgaben (von -0.49 und -0.14, vgl. Tab. 6 unten). In Tabelle 5 sind die Veränderungen der Aufgabenschwierigkeiten zwischen den Messzeitpunkten in Logits angegeben, die Standardabweichungen der Differenzen von diesen Mittelwerten betragen $SD_{12} = 0.42$ zwischen Vor- und Nachtest sowie $SD_{23} = 0.34$ zwischen Nach- und Follow-up-Test.

In Abbildung 1 sind Weighted-Mean-Square-Fit-Werte (Wright & Masters 1982) für die 13 allen drei Zeitpunkten gemeinsamen Items nach dem Equating-Modell zu drei Zeitpunkten und nach dem Andersen-Modell dargestellt. Im Vergleich zeigt sich ein deutlicher Unterschied: Im OP-Equating-Modell sind die Fitwerte für alle 89 Items unauffällig, d. h. nahe ihrem Erwartungswert von Eins. Im OP-Andersen-Modell mit Restriktion von 13 Items (ebenso wie im entsprechenden OP-Embretson-Modell) sind die Fitwerte der zwei und ein Mal eingesetzten Items, deren Pa-

rameter in diesem Modell nicht restringiert sind, ebenfalls unauffällig. Der Fit dagegen ist bei den 13 gemeinsamen Items mit restringierten Parametern teilweise deutlich erhöht, für 10 von 13 Items sind die Fitwerte statistisch bedeutsam höher als Eins. Das weist darauf hin, dass für diese Items zu viele unerwartete Antworten beobachtet wurden bzw. dass das Andersen-Modell für diese Items nicht passt.

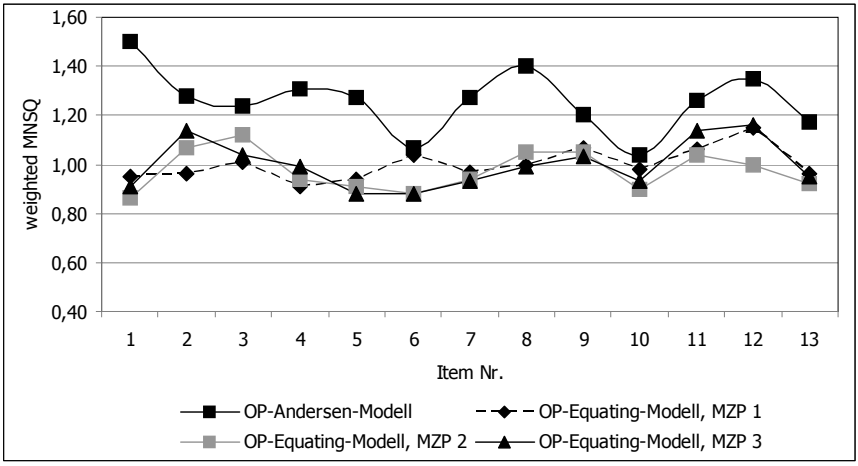


Abbildung 1: Itemfit für die 13 allen drei Zeitpunkten gemeinsamen Items nach dem OP-Andersen-Modell und nach dem OP-Equating-Modell über je drei Werte für die drei Zeitpunkte eines Items gemittelt

Zur weiteren Kontrolle des Fit für diese Parameter aus dem OP-Andersen-Modell wurde ein modifiziertes OP-Equating-Modell geschätzt, wofür die Itemparameter aller 89 Itembearbeitungen auf die Parameter aus dem OP-Andersen-Modell fixiert wurden. Das neu geschätzte Modell unterscheidet sich in den Itemparametern somit nicht vom OP-Andersen-Modell, jedoch unterscheidet es sich in der Zuordnung der beobachteten Antworten zu den latenten Dimensionen: Im Equating-Modell mit (und auch ohne) fixierten Itemparametern werden (Itemparameter und) Itemfitwerte anhand der Antworten eines Messzeitpunkts bestimmt, während im OP-Andersen-Modell die Daten aller drei Testzeitpunkte zur Berechnung von deren Fitwerten herangezogen werden. Als Ergebnis des Equating-Modells mit fixierten Parametern finden sich unauffällige Fitwerte für die 13 allen Zeitpunkten gemeinsamen Items.

3.3 Veränderungen in den naturwissenschaftlichen Kompetenzen

Sowohl mit dem OP-Andersen- als auch mit dem OP-Embretson-Modell können Veränderungen direkt an den Kompetenzverteilungen abgelesen werden. In Tabelle 6 sind die Mittelwerte und Standardabweichungen der Kompetenzveränderungen nach dem Embretson-Modell dargestellt; in Tabelle 7 die Reliabilitäten der drei Tests, die mittleren Kompetenzen, deren Standardabweichungen sowie stan-

dardisierte Kompetenzmittelwerte nach dem Andersen-Modell. Die Standardisierung bezieht sich erstens auf die Varianz der Kompetenz zum ersten Messzeitpunkt und zweitens auf die gepoolte Varianz aus allen drei Messzeitpunkten.

Kompetenzergebnisse (in logits)	t = 1	t2-t1	t3-t2
Mittelwert der Kompetenz und -veränderungen	0.00 gesetzt	.49	.13
Standardabweichung der Kompetenz und -veränderungen	0.84	0.50	0.20

Tabelle 6: Veränderungen in der naturwissenschaftlichen Kompetenz über die drei Zeitpunkte in Logits und in Standardabweichungen: OP-Embretson-Modell

Kompetenzergebnisse (in Logits)	t = 1	t = 2	t = 3
Reliabilität, Andersen-Modell (WLE)	.75	.80	.84
Mittelwert der Kompetenzen	0.00 gesetzt	.49	.61
Standardabweichung der Kompetenzen	0.84	0.93	1.04
Zuwachs in d (zur Varianz zu t1)	0.00 gesetzt	d = .59	d = .73
Zuwachs in d (zur gepoolten Var.)	0.00 gesetzt	d = .53	d = .66

Tabelle 7: Veränderungen in der naturwissenschaftlichen Kompetenz über die drei Zeitpunkte in Logits und in Standardabweichungen: OP-Andersen-Modell

In Tabelle 8 sind latente Korrelationen zwischen den Tests zu den drei Messzeitpunkten nach dem Andersen-Modell angegeben, die sich als Stabilitäten der Kinder über die Zeit interpretieren lassen; in Tabelle 9 latente Korrelationen der Veränderungen mit dem Ausgangswert nach dem OP-Embretson-Modell.

Latente Korrelationen	t1	t2
t2	.83	
t3	.84	.97

Tabelle 8: Korrelationen in der Kompetenz zu drei Zeitpunkten: OP-Andersen-Modell

Latente Korrelationen	t1	t2-t1
t2-t1	-.17	
t3-t2	.51	.02

Tabelle 9: Korrelationen in der Kompetenz zu drei Zeitpunkten: OP-Embretson-Modell

Die mittleren Zuwächse betragen 0.49 Logits über die Interventionsphase[1] und 0.13 Logits in der Follow-up-Phase. Dies ergibt sich direkt aus dem Embretson-Modell wie auch aus den entsprechenden Differenzen der mittleren Kompetenzen nach dem Andersen-Modell. In Standardabweichungen ausgedrückt entspricht diese Zunahme einer Effektstärke von d = .73, gemessen an der Varianz zum ersten Messzeitpunkt, oder d = .66, gemessen an der gemittelten Varianz aller drei Messzeitpunkte.

1 Interventionen wurden in 3 Experimentalgruppen durchgeführt, nicht in den beiden Kontrollgruppen (Steffensky et al. in Vorbereitung-b)

Die Standardabweichung der Veränderung vom ersten zum zweiten Messzeitpunkt beträgt 0.50 Logits und zwischen dem zweiten und dritten noch 0.20 Logits; die Varianzen der Kompetenzverteilungen nehmen über beide Phasen vergleichbar stark zu. In der Interventionsphase findet sich deutlich mehr individuelle Varianz in den Veränderungen als in der Follow-up-Phase. Dies lässt den Schluss zu, dass die Intervention eine differenzierende Wirkung auf die Kompetenzentwicklung der Kinder hatte.

Mit den zunehmenden Varianzen der Kompetenz steigt nicht nur die Diskriminationsfähigkeit der im Projekt SNaKE eingesetzten Tests über die Zeit, sondern auch deren Zuverlässigkeit, was die Reliabilitäten von .75, .80 und .84, berechnet anhand von Weighted Likelihood Estimates (WLE, Warm 1989), zeigen. In derselben Größenordnung liegen etwa die Reliabilitäten der Untertests des CFT-1 für Fünfjährige (Cattell et al. 1997).

Mit Koeffizienten zwischen r = 0.84 und r = 0.97 finden sich hohe Zusammenhänge zwischen den Tests und damit auch hohe Stabilitäten der Kompetenzen der Kinder. Während die Kompetenzen zum ersten Messzeitpunkt weniger mit denen vom zweiten und dritten kovariieren, ist die Stabilität zwischen dem zweiten und dritten Messzeitpunkt besonders hoch. Zwischen der zweiten und dritten Messung finden nur noch geringe Veränderungen statt, wie auch an der Standardabweichung dieser Veränderung deutlich wird. Jedoch kovariieren die Veränderungen in der Follow-up-Phase mittelstark (r = .51) mit der Kompetenz zum ersten Messzeitpunkt, so dass die Varianz zum dritten Messzeitpunkt noch einmal deutlich höher ausfällt als zum zweiten. Der Zuwachs zwischen den ersten beiden Messzeitpunkten ist mit r = -.17 dagegen schwach negativ mit der Ausgangkompetenz korreliert: Tendenziell zeigen am Anfang schwächere Kinder einen höheren Zuwachs und Kinder mit höheren Kompetenzen zum ersten Messzeitpunkt einen geringeren.

Im Gesamtbild zeigen sich also auffallende Veränderungen mit einem deutlichen mittleren Kompetenzanstieg in der Interventionsphase, die kaum mit der Kompetenz zum ersten Messzeitpunkt zusammenhängen und daher Raum für die Annahme lassen, dass die Intervention eine eigenständige Wirkung auf die Entwicklung der Kompetenzen der Kinder hat. Die Veränderungen in der Follow-up-Phase fallen in Varianz und mittlerem Zuwachs geringer aus und sind deutlich mit dem Ausgangswert korreliert. In dieser Phase hat die Intervention offenbar weniger Einfluss auf die Veränderungen, stattdessen ordnen sich die Kompetenzen der Kinder eher wieder so an wie vor der ersten Testung. Das lässt auf eine von individuellen Voraussetzungen abhängige starke Entwicklung von naturwissenschaftlicher Kompetenz in der untersuchten Altersstufe schließen, die jedoch durch die zeitlich begrenzt wirksame Intervention der Studie "unterbrochen" wurde.

4 Diskussion

Der vorliegende Beitrag geht der Frage nach, wie sich die Veränderungen über die drei Messzeitpunkte in den Itemantworten aus der SNaKE-Studie modellieren las-

sen. Die Überprüfung der drei für das Projekt entwickelten Tests erfolgt mit je einem Ordered-Partition-Modell (Wilson 1992), welches eine gleiche Gewichtung von dichotomen und mehrkategoriellen Antwortformaten umsetzt. Als mögliche Beschreibung für längsschnittliche Daten wurden dreifaktorielle Antwortmodelle nach Rost (2004) auf der Basis eines OPM diskutiert. Als restriktivste wurden ein OP-Andersen- und ein OP-Embretson-Modell entsprechend dem Modell (3) nach Rost (ebd.) abgeleitet. Beide wurden in den Varianten mit Gleichsetzung von 13 gemeinsamen und 24 mindestens zwei Zeitpunkten gemeinsamen Aufgaben geschätzt. Als weniger restriktive Variante wurde ein OP-Equating-Modell diskutiert, welches dem Modell (5) nach Rost entspricht. Dieses Modell setzt die Testschwierigkeiten der Tests zu den unterschiedlichen Messzeitpunkten gleich, ohne jedoch eine im Sinne von Rasch-Modellen eindimensionale gemeinsame latente Variable über alle Zeitpunkte anzunehmen, und dient dem Vergleich mit restriktiveren Modellen.

Für die drei Naturwissenschaftstests aus der Studie SNaKE wurde gezeigt, dass sie dieselbe Kompetenz erfassen. Das war zu erwarten, weil die Tests nach einer einheitlichen theoretischen Konzeption konstruiert waren und das Testdesign eine Mehrzahl von gemeinsamen Items in den aufeinanderfolgenden Tests enthielt, und wird durch die empirischen Ergebnisse der Studie bestätigt. Zunächst wurde festgestellt, dass jeder der drei Tests einer Überprüfung auf Homogenität im Sinne des OPM standhält. Die Varianzen der Kompetenzverteilungen der Kinder sind ausreichend groß, um Differenzierungen zwischen ihnen zuzulassen. Die hohen Reliabilitäten erlauben differenzierte Analysen auf Gruppenebene. Alle Aspekte der theoretischen Fundierung und empirischen Ergebnisse zusammenfassend ist eine valide und für die Altersstufe von Fünfjährigen sehr zuverlässige Testkonstruktion gelungen. Weiterhin ist es gelungen, eine längsschnittliche Modellierung der Kompetenzveränderungen vorzunehmen, so dass sich die beobachteten Veränderungen in der Studie anhand der Parameterschätzungen der beiden angepassten Modelle beschreiben lassen.

Als Ergebnis kann festgehalten werden, dass sich über den Zeitraum von sieben Monaten eine Entwicklung von Kompetenzen im Mittel um $d = 0.73$ Standardabweichungen gezeigt hat. Da für die Altersstufe von Fünfjährigen keine vergleichbaren Studien vorliegen, kann dieses Ergebnis nicht mit den Befunden anderer Untersuchungen verglichen werden. Bloom et al. (2008) berichten Effektgrößen für Lesen und Mathematik ab Klassenstufe 1 und für Naturwissenschaften ab Klasse 2 aus US-amerikanischen Vergleichsstudien. Daraus lassen sich zwar keine Erwartungen für eine Effektstärke für Naturwissenschaften im Kindergartenalter ableiten, jedoch fügt sich die beobachtete Effektstärke von $d = 0.73$ konsistent ins Bild. Ab 2013 werden im Nationalen Bildungspanel (NEPS: Weinert et al. 2011, Blossfeld et al. 2009) Ergebnisse zur Entwicklung der naturwissenschaftlichen Kompetenz im Alter zwischen vier und sechs Jahren vorliegen. Für Jugendliche zwischen ca. 15 und 17 Jahren wurden Zuwächse über ein Jahr zwischen $d = 0.24$ und $d = 0.45$, im Mittel jedoch kleiner als $d = 0.30$ gefunden (Walter et al. 2006). Im Vergleich dazu sind die Zuwächse der Fünfjährigen hoch. Aus entwicklungspsy-

chologischer Perspektive ist zu erwarten, dass Lernzuwächse über die Schullaufbahn generell abnehmen (vgl. Bloom et al. 2008); das vorliegende Ergebnis stellt damit einen Befund (und eine Benchmark) zur Höhe von Lernzuwächsen kurz vor dem Eintritt in das allgemein bildende Schulsystem dar.

Ein weiteres Ergebnis besteht darin, dass die Intervention der Studie in der Interventionsphase offenbar Kompetenzveränderungen bewirkt hat, in der folgenden Follow-up-Phase jedoch eher nicht. Daran sollten weitere Analysen anschließen, welche die Experimentalgruppen und die unterschiedlichen individuellen Voraussetzungen einbeziehen (vgl. Steffensky et al. 2012b).

Zu diskutieren ist, inwieweit mit den 51 eingesetzten Aufgaben eine homogene Entwicklung abgebildet werden kann, wie es mit der Annahme einer über die Zeit Rasch-homogenen latenten Variable in beiden Längsschnittmodellen gefordert wird. Aus fachdidaktischer und entwicklungspsychologischer Perspektive ist davon auszugehen, dass Veränderungen der Aufgabenschwierigkeiten über die Zeit heterogener sind als durch ein Rasch-Modell abbildbar, da sich die Kompetenzentwicklung der Kinder in viele Teilprozesse gliedert, die bei jedem Kind in unterschiedlichen Reihenfolgen und zeitlichen Abfolgen ablaufen können. Mit den Differenzen der Schwierigkeitsparameter der 13 den aufeinanderfolgenden Messzeitpunkten gemeinsamen Items liegt eine empirische Quantifizierung der Heterogenität in der Veränderung der Aufgabenschwierigkeiten vor. Die Standardabweichungen von $SD_{12} = 0.42$ und $SD_{23} = 0.34$ in Logits für den Unterschied zwischen Vor- und Nachtest bzw. Nach- und Follow-up-Test entsprechen beide etwa 0.45 Standardabweichungen der mittleren Kompetenzverteilung aus den beiden Messzeitpunkten. Gemäß der Berechnung des Linking-Fehlers in der OECD-Studie PISA (OECD 2005) wäre der Linkfehler zwischen Vor- und Nachtest sowie zwischen Nach- und Follow-up-Test $SE_L = 0.10$ Standardabweichungen der jeweiligen gemittelten Kompetenzverteilungen; auch für PISA werden Fehler in der Größenordnung von bis zu 0.10 Standardabweichungen der Kompetenzskala angegeben. Insofern ist die Heterogenität der SNaKE-Tests nicht größer als in anderen Studien.

Eine weitere Frage werfen die auffälligen Itemfitwerte der restringierten Items (vgl. Abb. 1) im Andersen- und im Embretson-Modell auf. Mit Ausnahme der restringierten Items werden alle Parameter bezogen auf eine Stichprobe zu einem Testzeitpunkt geschätzt. Durch die Restriktion werden Parameter geschätzt, die sich auf drei Stichproben von drei Testzeitpunkten beziehen. Genau diese Parameter fallen durch niedrige Fitwerte auf. Die Verteilung von Kompetenzen zu drei Messzeitpunkten, anhand der die Schwierigkeiten dieser Items geschätzt werden, ist in der vorliegenden Studie breiter als entsprechende Verteilungen zu einem Messzeitpunkt. Insofern scheint die Diskriminierungsleistung dieser Items geringer. Werden mit denselben Parametern Fitwerte in eindimensionalen Modellen geschätzt (wie oben berichtet), fallen diese hingegen unauffällig aus. Die unterschiedliche Spezifikation der Population wirkt sich in MML-geschätzten Modellen auch auf die Modellgüte aus. Insofern wäre hier ein Vergleich mit einem durch CML geschätzten Modell interessant.

Eine Modellalternative kann darin bestehen, für die restringierten Items einen zweiten logistischen Parameter gemäß dem 2PL-Modell (Birnbaum 1968) einzuführen. Von Davier et al. (2011) verallgemeinern die Modelle von Andersen und Embretson durch die Annahme von Trennschärfeparametern im Sinne des 2PL-Modells ebenfalls unter Annahme von multivariat normalverteilten Personenfähigkeiten zu mehrdimensionalen 2PL-Modellen wie dem Generalized Partial Credit Model (Muraki 1992) oder dem General Diagnostic Model (GDM; von Davier 2005). Zur Schätzung der Parameter dieser Modelle wird eine marginale Likelihood maximiert. Mit solchen Modellen können Trennschärfeparameter für die Veränderungen im Embretson-Modell spezifiziert werden, was die Untersuchung der Änderungssensitivität von Items ermöglicht. So kann es interessant sein, die unterschiedlichen Veränderungen der Itemschwierigkeiten in einem Modell mit Parametern zur Beschreibung der Änderungssensitivität von einzelnen Aufgaben neu zu analysieren. Weiterhin kann die durch die auffälligen Itemfitwerte aufgeworfene Frage durch alternative Modelle aufgegriffen werden: Betrachtet man diesen Befund etwa als Hinweis darauf, dass in Modellen mit über mehrere Zeitpunkte restringierten Items von unterschiedliche Diskriminationen auszugehen ist, so kann dies als ein 2PL-Embretson-Modell mit Gleichheitsrestriktionen in den Diskriminationsparametern spezifiziert werden.

Modelle der Multitrait-Multimethod-Analyse für Variablen mit geordneten Antwortkategorien

Michael Eid & Fridtjof W. Nussbeck

1 Die Multitrait-Multimethod-Analyse

Die Multitrait-Multimethod-Analyse gehört zu den einflussreichsten methodischen Entwicklungen der Psychologie. Sie geht zurück auf Campbell & Fiskes (1959) Arbeit zur konvergenten und diskriminanten Validierung, in der sie die Prinzipien der Validierung von Messinstrumenten mittels der Multitrait-Multimethod-Matrix vorstellen. Die Grundidee besteht darin, dass bei der Validierung eines Messinstruments sowohl mehrere Methoden zum Einsatz kommen sollen als auch mehrere Merkmale berücksichtigt werden müssen. Die Forderung nach Multimethodalität liegt darin begründet, dass sich ein Merkmal anhand mehrerer Methoden erheben lassen müsse und diese zu einem gleichen oder ähnlichen Ergebnis kommen müssten, sofern die verschiedenen Messinstrumente das erfassen, was sie messen sollen, nämlich das gemeinsame Merkmal. Konvergieren die verschiedenen Methoden, so liegt konvergente Validität vor. Entwickelt man ein neues Messinstrument, so sollte dieses zu bereits entwickelten Methoden der Erfassung desselben Merkmales hohe konvergente Validität aufweisen.

Die Berücksichtigung mehrerer Merkmale im Validierungsprozess ist bei der Einführung eines neuen Merkmals in die Wissenschaft wünschenswert, da dieses hinreichend von bereits vorhandenen verschieden sein sollte, um seine Einführung zu rechtfertigen. Ein Messinstrument zur Erfassung eines neuen Merkmals soll daher auch zu anderen Ergebnissen führen als Messinstrumente, die zur Erfassung anderer Merkmale eingesetzt werden. Mit anderen Worten: Es soll diskriminante Validität vorliegen, d. h. die Messwerte, die durch das neue Messinstrument gemessen werden, dürfen nicht hoch mit den Messwerten anderer Merkmale zusammenhängen, sofern diese voneinander unabhängig sind.

Bei der Validierung eines Messinstruments strebt man im Allgemeinen hohe konvergente und hohe diskriminante Validität an: Die Messergebnisse, die man durch den Einsatz eines Messinstruments gewonnen hat, sollen hoch mit denen korrelieren, die man anhand anderer geeigneter Messinstrumente erhalten hat. Auf der anderen Seite sollen die Messergebnisse nicht mit jenen zusammenhängen, die man anhand von Messinstrumenten erhalten hat, die zur Erfassung von Merkmalen geeignet sind, die unabhängig von dem Zielmerkmal sind.

Campbell & Fiske (1959) haben darüber hinaus deutlich gemacht, dass die Erfassung eines Merkmals (Traits) in der Psychologie immer nur mittels einer Methode möglich ist, jede Messung somit eine Trait-Methoden-Einheit repräsentiert. Das Messergebnis stellt nicht nur das zu erfassenden Merkmal dar, sondern zu einem Teil auch das dazu herangezogene Messinstrument. Inwieweit die Messung von der Methode abhängt oder methodenunspezifisch ist, ist eine empirische Frage. Liegt hohe konvergente Validität vor, so ist der methodenspezifische Anteil eher gering. Im umgekehrten Fall spiegelt die Messung eher die Messmethode wider als das zu messende Merkmal. Bei der Interpretation von Messergebnissen muss daher immer die Frage behandelt werden, ob das Messinstrument auch das misst, was es messen soll. Die Forderung nach konvergenter und diskriminanter Validität impliziert darüber hinaus, dass der Validierungsprozess mindestens zwei Methoden und mindestens zwei Merkmale umfassen muss.

2 Methoden zur Analyse der konvergenten und diskriminanten Validität

Zur Analyse der konvergenten und diskriminanten Validität haben Campbell & Fiske (1959) die Multitrait-Multimethod-Matrix (MTMM-Matrix) vorgeschlagen, in der die Korrelationen zwischen Traits angeordnet sind, die mittels verschiedener Methoden gemessen wurden. Sie haben darüber hinaus Kriterien empfohlen, anhand derer die Korrelationen im Hinblick auf die konvergente und diskriminante Validität bewertet werden können (Eid 2010, Eid et al. 2006b). Die Analyse der MTMM-Matrix anhand dieser Kriterien wird vor allem aus dreierlei Gründen kritisiert (Eid et al. 2006b, Millsap, 1995). 1. Zum einen beziehen sich die Korrelationen in der MTMM-Matrix auf den Zusammenhang manifester und somit messfehlerbehafteter Variablen. Die Bewertung der Korrelationen in Bezug auf die konvergente Validität (hohe Korrelationen von Verfahren, die dasselbe Merkmal messen) und die diskriminante Validität (geringe Korrelation von Verfahren, die verschiedene Merkmale messen) ist schwierig, wenn sich die verschiedenen Verfahren in ihren Reliabilitäten unterscheiden. Da die Höhe einer möglichen Korrelation zwischen zwei Variablen von deren Reliabilitäten abhängt, können Korrelationsunterschiede einfach nur reliabilitätsbedingt sein und die konvergente Validität kann unter- und die diskriminante Validität überschätzt werden. 2. Zum anderen basieren die Kriterien auf der visuellen Inspektion der MTMM-Matrix. Sie erlauben hingegen nicht die statistische Überprüfung spezifischer Hypothesen. 3. Schließlich werden die Zusammenhänge in der MTMM-Matrix nicht anhand eines statistischen Modells erklärt. Allerdings erlaubt erst der Rückgriff auf ein statistisches Modell die Trennung von Trait- und Methodeneinflüssen sowie die Definition von Variablen, die Trait- und Methodeneinflüsse repräsentieren, und damit die gezielte Erklärung von Methodeneffekten.

Aufgrund dieser Beschränkungen wurde in den letzten Jahrzehnten eine Reihe von Modellen zur Analyse von MTMM-Daten entwickelt. Insbesondere drei Klassen von Verfahren haben sich in der Praxis durchgesetzt: 1. Direct-Product-Modelle, 2. varianzanalytische Modelle sowie 3. Modelle der konfirmatorischen Faktorenanalyse

(Dumenci 2000). Die letztgenannten sind die am häufigsten eingesetzten (Eid et al. 2006a).

Die bisherigen Modellentwicklungen der MTMM-Analyse beziehen sich vor allem auf metrische Antwortvariablen. Sie können daher nur eingeschränkt zur Analyse von kategorialen Variablen und somit zur Itemselektion und Itemanalyse eingesetzt werden. Auch ermöglichen sie es nicht, auf Itemebene zu überprüfen, ob beispielsweise verschiedene Beurteiler (Rater), die als verschiedene Methoden angesehen werden können, die Itemkategorien in gleicher Weise nutzen. MTMM-Modelle für kategoriale Variablen hingegen sind rar. Rost & Walter (2006) haben gezeigt, wie Modelle zur Analyse multimethodaler Daten in der Item-Response-Theorie definiert werden können. Sie berücksichtigen Methodeneffekte einerseits durch die Einführung von methodenspezifischen Itemparametern, andererseits durch die Definition von methodenspezifischen latenten Traitvariablen. Anhand methodenspezifischer Itemparameter lässt sich untersuchen, ob sich Methodeneffekte generell (über alle Items einer Methode hinweg) oder itemspezifisch (Interaktion zwischen Item und Methode) auswirken. Anhand methodenspezifischer Traitparameter lässt sich bestimmen, ob anhand verschiedener Methoden dieselbe latente Traitvariable oder unterschiedliche latente Traitvariablen (Interaktion zwischen Person und Methode) erfasst werden. Rost & Walters Modelle stellen Erweiterungen von Rasch-Modellen dar.

Nussbeck et al. (2006) haben gezeigt, wie MTMM-Modelle für Variablen mit geordneten Antwortkategorien als faktorenanalytische Modelle für kategoriale Variablen definiert werden können. In Ergänzung zu den bei Rost & Walter (2006) behandelten Methodeneffekten zeigen sie, wie allgemeine latente Traitvariablen und merkmalsspezifische Methodeneffekte definiert werden können und wie die konvergente und diskriminante Validität im Rahmen dieser Modelle untersucht werden kann. Crayen et al. (2011) haben den Ansatz von Nussbeck et al. (2006) auf die Analyse von Longitudinaldaten zur Evaluation von Interventionseffekten übertragen.

Die genannten Modelle von Rost & Walter, Nussbeck et al. sowie Crayen et al. sind typischerweise für strukturell unterschiedliche Methoden geeignet. Strukturell unterschiedliche lassen sich von austauschbaren Methoden abgrenzen. Wie Eid et al. (2008) gezeigt haben, sind für für beide Arten von Methoden unterschiedliche Modelle geeignet. Wir wollen den Unterschied zwischen ihnen erläutern, indem wir verschiedene Beurteiler (Rater) als verschiedene Methoden betrachten (Kenny 1995). Austauschbare Methoden zeichnen sich dadurch aus, dass sie zufällig aus einer Menge zur Verfügung stehender und prinzipiell gleichwertiger Methoden gezogen werden. Wollen wir bspw. die Qualität der Lehre eines Professors bewerten, so können wir aus der Menge seiner Studierenden per Zufall eine gewisse Anzahl von Ratern auswählen. Die Studierenden wären untereinander austauschbar, da sie mehr oder weniger denselben Zugang zu den Merkmalen haben, die es zu beurteilen gilt (sie besuchen alle die zu evaluierende Lehrveranstaltung). Auch unterscheiden sie sich nicht hinsichtlich der Perspektive, die sie in Bezug auf die Beurteilung der Qualität der Lehre einnehmen (alle Studierenden haben die Perspektive der Lernenden). Ihre Beurteilungen sind gleichwertige Repräsentationen der Lehreinschätzung durch Studierende. Werden also per Zufall je drei Stu-

dierende als Rater der Qualität der Lehre unterschiedlicher Professoren aus deren Kursen gewählt, so liegen im Sinne der MTMM-Modelle drei austauschbare Methoden vor. Strukturell unterschiedliche Methoden werden im Gegensatz zu den austauschbaren nicht aus derselben Menge von Methoden gezogen. Wenn in unserem Beispiel zusätzlich noch ein Vertreter des hochschuldidaktischen Zentrums und der Studiendekan um ihre Einschätzungen gebeten würden, so lägen gänzlich unterschiedliche Perspektiven und auch Zugänge zu den zu beurteilenden Merkmalen vor. Die Hochschuldidaktiker und der Studiendekan gehören nicht zur Menge der Studierenden, sie sind im Sinne der MTMM-Modelle nicht austauschbar mit Studierenden, aber auch nicht untereinander. Es handelt sich bei den letzten beiden aufgenommenen Ratern um strukturell unterschiedliche Methoden.

Ein weiteres Beispiel für austauschbare Methoden sind Freunde, die einen gemeinsamen Freund beurteilen sollen. Der Freund selber ist allerdings nicht austauschbar, denn er beurteilt sich aus der Selbstperspektive, während seine Freunde ihn aus der Fremdperspektive beurteilen.

Typischerweise kommen in vielen Fragestellungen der Psychologie sowohl strukturell unterschiedliche als auch austauschbare Methoden (z. B. Rater) vor. Austauschbare Methoden sind durch eine Mehrebenenstruktur gekennzeichnet. Zunächst wird aus einer Menge zu beurteilender Personen (Merkmalsträger) eine Menge (Anzahl) gezogen. Dies entspricht der übergeordneten Einheit (Level-2). Für jede zu beurteilende Person wird dann eine Anzahl (Menge) von Ratern aus einer Menge von möglichen Ratern für diese Person gezogen. Dies entspricht der untergeordneten Einheit (Level-1). Die Rater sind somit in den zu beurteilenden Personen (Merkmalsträgern) geschachtelt.

Eid et al. (2008) haben dargelegt, wie Modelle für strukturell unterschiedliche und austauschbare Methoden als spezielle Mehrebenenmodelle der konfirmatorischen Faktorenanalyse definiert werden können. Nussbeck et al. (2009) zeigen, wie diese Modelle dann mit nur einer Ebene anhand spezifischer Restriktionen definiert werden können. Sie machen sich dabei den Umstand zunutze, dass spezifische Mehrebenenmodelle sich durch Restriktionen auch als klassische Modelle der Faktorenanalyse formulieren lassen (Curran 2003). Die von Eid et al. (2008) und Nussbeck et al. (2009) behandelten Modelle beziehen sich auf metrische beobachtete Variablen. Im Folgenden soll gezeigt und empirisch illustriert werden, wie MTMM-Modelle für strukturell unterschiedliche und austauschbare Rater für Variablen mit geordneten Antwortkategorien formuliert werden können. Wir beziehen uns dabei auf den Ansatz von Nussbeck et al. (2006), da selbst leistungsfähige Computer bei der Übertragung der Mehrebenenmodelle nach Eid et al. (2008) auf kategoriale Variablen mit dem Computerprogramm Mplus (Muthén & Muthén, 1998-2010) an Arbeitspeicherproblemen scheiterten. Wir formulieren zunächst die Modelle für ein Merkmal (Trait) und übertragen sie dann auf den Fall von zwei Merkmalen. Wir beziehen uns auf den einfachsten Fall von drei Methoden (Selbstbericht und zwei Fremdberichte). Aus didaktischen Gründen führen wir diese Modelle anhand eines empirischen Beispiels ein, der Erfassung des habituellen Stimmungserlebens durch eine Person (Selbstbericht) und durch zwei Freunde. Zunächst werden das habi-

tuelle Wohlbefinden (1. Merkmal) und in der Erweiterung die habituelle Wachheit (2. Merkmal) als Bestandteile des habituellen Stimmungserlebens modelliert. Die beiden Fremdberichte (Freunde der Person, die den Selbstbericht abgibt) werden als austauschbar angesehen, da angenommen wird, dass sie aus derselben Menge von Freunden stammen. Der Selbstbericht hingegen unterscheidet sich strukturell von den Fremdberichten (siehe oben). Es liegen 482 Selbstberichte und 964 Fremdberichte vor (zwei pro einzuschätzende Person).

3 Modelle zur Erfassung eines Merkmals mit mehreren Methoden

Wir stellen die Definition eines MTMM-Modells für Variablen mit geordneten Kategorien Schritt für Schritt dar und beginnen mit einem Modell für ein Merkmal (habituelles Wohlbefinden), das mit drei verschiedenen Methoden (Selbstbericht, Fremdbericht A, Fremdbericht B) erfasst wurde. Wir beziehen uns auf die drei Items zufrieden, wohl, glücklich, anhand derer die Zielperson einschätzen sollte, wie sie sich im Allgemeinen fühlt. Pro Item standen fünf Antwortkategorien zur Verfügung. Die unterste Kategorie war mit "überhaupt nicht", die höchste mit "sehr stark" bezeichnet. Die Kategorien dazwischen waren nicht verbal verankert.

3.1 Modell gemeinsamer Faktoren

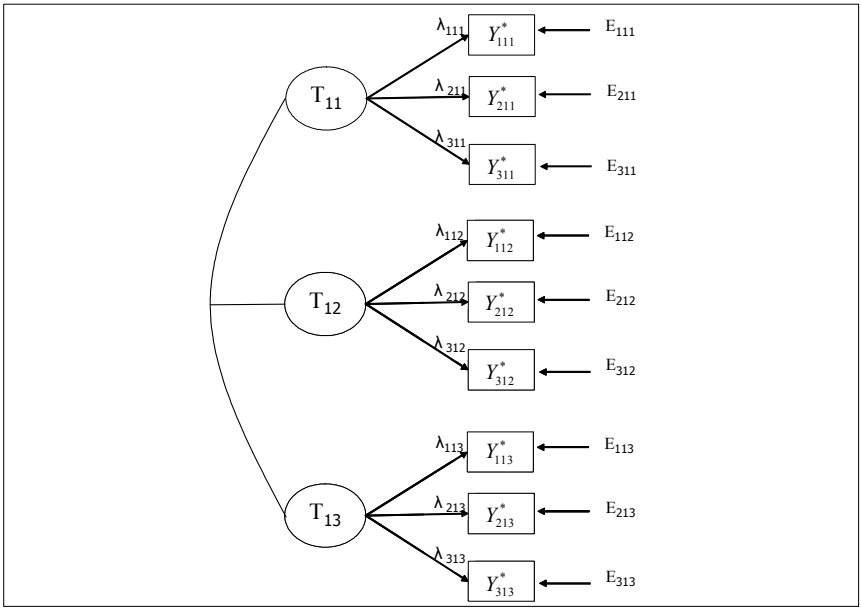


Abbildung 1: Modell gemeinsamer Faktoren für ein Merkmal (Trait) und drei Methoden. Y^*_{ijk} bezeichnet die latenten Antwortvariablen, die den beobachteten Variablen zugrunde liegen. T_{jk}: gemeinsamer Faktor. Der erste Index i (i = 1, …, m) kennzeichnet den Indikator einer Trait-Methoden-Einheit, der zweite Index j (j = 1, …, n) indiziert das betrachtete Merkmal (Trait), der letzte Index k gibt die Methode (k = 1, …, o) an.

In Abbildung 1 ist für diese Datensituation ein als faktorenanalytisches Modell für Variablen mit geordneten Antwortkategorien formuliertes Modell gemeinsamer Faktoren dargestellt. Es geht davon aus, dass jeder beobachteten kategorialen Antwortvariablen Y_{ijk} (z. B. zufrieden) eine kontinuierliche latente Antwortvariable Y^*_{ijk} zugrunde liegt. Der erste Index i (i = 1, …, m) kennzeichnet den Indikator einer Trait-Methoden-Einheit (zufrieden), der zweite Index j (j = 1, …, n) indiziert das betrachtete Merkmal (Trait: habituelles Wohlbefinden), der dritte Index k gibt die Methode (k = 1, …, o) an. Jedes Item umfasst eine beschränkte Anzahl von Antwortkategorien, denen die ganzen positiven Zahlen s = 0, …, c_{ijk}-1 zugeordnet werden. Mit c_{ijk} wird die Anzahl der Kategorien der Variablen Y_{ijk} bezeichnet. Die beiden Variablen Y^*_{ijk} und Y_{ijk} sind durch folgende Schwellenwertbeziehung miteinander verknüpft:

$$Y_{ijk} = \begin{cases} c_{ijk} - 1, & \text{für } \kappa_{(c_{ijk}-1)ijk} < Y^*_{ijk} \\ s, & \text{für } \kappa_{sijk} < Y^*_{ijk} \le \kappa_{(s+1)ijk}, \text{und } 0 < s < c_{ijk} - 1 \\ 0, & \text{für } Y^*_{ijk} \le \kappa_{1ijk} \end{cases} \quad (1)$$

Die Schwellenparameter κ_{sijk} zerlegen die kontinuierliche latente Antwortvariable Y^*_{ijk} in c_{ijk} Segmente. Jedes Segment entspricht einer Kategorie der beobachteten Variablen Y_{ijk}, die dem Modell zufolge durch Kategorisierung der kontinuierlichen Variablen Y^*_{ijk} entstanden ist. Für das Item zufrieden bedeutet dies beispielsweise, dass sich Versuchspersonen auf latenter Ebene hinsichtlich ihrer Zufriedenheit verorten lassen, Versuchspersonen wählen diejenige beobachtbare Antwortkategorie, die aufgrund der Schwellenbeziehung zu ihrer latenten Merkmalsausprägung passt. Da jeder Antwortkategorie ein Bereich der latenten Ausprägung zugeordnet wird, können Versuchspersonen, die die gleiche Antwortkategorie gewählt haben, Unterschiede in der latenten Ausprägung aufweisen.

Die latenten Antwortvariablen Y^*_{ijk} werden faktorenanalytisch zerlegt. Im Modell gemeinsamer Faktoren für ein Merkmal (Trait) und mehrere Methoden wird davon ausgegangen, dass alle beobachteten Variablen (z. B. zufrieden, wohl, glücklich), die zu derselben Trait-Methoden-Einheit (j, k) gehören (Zufriedenheit im Selbstbericht), einen gemeinsamen Faktor T_{jk} und eine Residualvariable E_{ijk} erfassen:

$$Y^*_{ijk} = \lambda_{ijk} T_{jk} + E_{ijk} \quad (2)$$

Da die Variablen Y^*_{ijk} nicht direkt beobachtbar sind, müssen bestimmte Annahmen zu ihrer Identifikation getroffen werden. Typischerweise geht man davon aus, dass sie multivariat standardnormalverteilt sind. Die Variablen T_{jk} und E_{ijk} seien ebenfalls normalverteilt mit einem Erwartungswert von 0. Die Residualvariablen E_{ijk} seien stochastisch unabhängig voneinander. Darüber hinaus muss pro Faktor eine Ladung auf einen Wert (üblicherweise den Wert 1) fixiert oder aber die Faktorvarianz normiert werden (üblicherweise auf den Wert 1). Die Korrelationen der Variablen Y^*_{ijk} sind die polychorischen Korrelationen, die faktorenanalytisch zerlegt werden.

Neben dieser Parametrisierung gibt es eine zweite, bei der nicht die Varianz der Variablen Y^*_{ijk} , sondern die Varianzen der Residualvariablen E_{ijk} auf den Wert Eins fixiert werden. Das Modell in dieser zweiten Parametrisierung ist äquivalent zu einem mehrdimensionalen Normalogiven-Graded-Response-Modell nach Samejima (1969). Diese Parametrisierung hat den Vorteil, dass die Parameter des faktorenanalytischen Modells denen des Item-Response-Modells entsprechen (vgl. hierzu ausführlich Eid 1995, Takane & deLeeuw 1987). Das Modell gemeinsamer Faktoren lässt sich wie folgt als Normalogiven-Graded-Response-Modell darstellen (Eid 1995). Zunächst wird für jede beobachtete Variable Y_{ijk} eine latente Antwortvariable π_{ijk} definiert, mit der sie durch folgende Wahrscheinlichkeitsfunktion verknüpft ist:

$$\text{prob}\left(Y_{ijk} \geq s \middle| \pi_{ijk}\right) = \Phi\left(\pi_{ijk} - \kappa_{sijk}\right) = \int_{-\infty}^{\pi_{ijk}-\kappa_{sijk}} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \qquad (3)$$

für alle s > 0. Mit κ_{sijk} werden die Schwellenparameter bezeichnet, die als Schwierigkeitsparameter interpretiert werden können. Je größer der Wert eines Parameters κ_{sijk} ist, desto geringer ist die Wahrscheinlichkeit, eine Antwort mindestens in der Kategorie s vorliegen zu haben. Mit Φ wird die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Für die bedingte Wahrscheinlichkeit einer Kategorie s der beobachteten Variablen Y_{ijk} ergibt sich hieraus:

$$\text{prob}\left(Y_{ijk} = s \middle| \pi_{ijk}\right) = \text{prob}\left(Y_{ijk} \geq s \middle| \pi_{ijk}\right) - \text{prob}\left(Y_{ijk} \geq s+1 \middle| \pi_{ijk}\right), \qquad (4)$$

Zusätzlich wird die Annahme der lokalen stochastischen Unabhängigkeit getroffen:

$$\text{prob}\left(Y_{111} = y_{111}, \ldots, Y_{mno} = y_{mno} \middle| \pi_{111}, \ldots, \pi_{mno}\right) = \qquad (5)$$

$$= \prod_{i=1}^{m} \prod_{j=1}^{n} \prod_{k=1}^{o} \text{prob}\left(Y_{ijk} = y_{ijk} \middle| \pi_{ijk}\right).$$

Diese Unabhängigkeitsannahme korrespondiert zu der Unabhängigkeit der Residuen in der faktorenanalytischen Darstellungsweise. Im Graded-Response-Modell wird die Variable π_{ijk} wird wie folgt zerlegt:

$$\pi_{ijk} = \lambda_{ijk} T_{jk}. \qquad (6)$$

Die Verbindung zwischen der faktorenanalytischen Darstellung und der in Form der Item-Response-Theorie lässt sich durch folgende Beziehung herstellen (Eid, 1995):

$$Y^*_{ijk} = \pi_{ijk} + \varepsilon^*_{ijk}. \qquad (7)$$

Von den latenten Variablen T_{jk} wird wie bei der ersten Parametrisierung angenommen, dass sie multivariat normalverteilt mit einem jeweiligen Erwartungswert von Null sind. Beide Parametrisierungen sind datenäquivalent, d. h., sie führen zur glei-

chen Modellanpassungsgüte und ihre Parameter können ineinander überführt werden (Eid 1995, Takane & deLeeuw 1987). Aufgrund der Äquivalenz zum Graded-Response-Modell greifen wir in diesem Beitrag auf die zweite Parametrisierung zurück, stellen das Modell allerdings faktorenanalytisch dar.

Die Parameter des faktorenanalytischen Modells können bspw. mit dem Computerprogramm Mplus (Muthén & Muthén 1998-2010) bestimmt werden, mit dem auch die Modellgültigkeit überprüft werden kann. Die erste Parametrisierung entspricht der Delta-Parametrisierung (derzeit Voreinstellung des Programms), die zweite erhält man, indem man die Theta-Parametrisierung wählt und alle Residualvarianzen auf Eins setzt. Dieses Vorgehen wurde zur Analyse der in diesem Beitrag berichteten Modelle gewählt. Die Parameter wurden mittels der Methode WLS-MV (Weighted Least Squares Mean and Variance adjusted) geschätzt, die für die vorliegende Stichprobengröße und die Anzahl der beobachteten Variablen angemessen ist (vgl. hierzu die Simulationsstudien von Nussbeck et al. 2006). Das WLS-MV-Verfahren bietet außerdem den Vorteil, dass es im Falle fehlender Werte für die Schätzung der polychorischen Korrelationen jeweils die paarweise verfügbare Information ausnutzt (Muthén & Muthén, 1998-2010), und selbst bei vergleichsweise vielen latenten Variablen handhabbar ist, wo klassische Full-Information-Maximum-Likelihood-Methoden an ihre Grenzen stoßen (vgl. Muraki & Carlson 1995).

Modell	Modellgüte				
	χ^2-Wert	df	p	CFI	RMSEA
M1: Keine Restriktionen	27.50	24	0.28	1.00	0.02
M2: Austauschbarkeit der Fremdberichte	40.41	40	0.45	1.00	0.01
M3: Austauschbarkeit der Selbst- und Fremdberichte	94.90	56	< 0.01	0.99	0.04
M4: Austauschbarkeit der Fremdberichte, Messäquivalenz mit Selbstbericht	56.71	53	0.34	1.00	0.01

Modell	Modellvergleiche			
	Vergleich mit Modell	χ^2-Differenzwert	df	p
M2: Austauschbarkeit der Fremdberichte	M1	15.00	16	0.53
M3: Austauschbarkeit der Selbst- und Fremdberichte	M2	46.53	16	< 0.01
M4: Austauschbarkeit der Fremdberichte, Messäquivalenz mit Selbstbericht	M2	17.53	13	0.18

Tabelle 1: Modellgütekoeffizienten für die Modelle gemeinsamer Faktoren

Das Modell gemeinsamer Faktoren (M1) erlaubt die Schätzung der konvergenten Validität auf latenter (messfehlerfreier) Ebene anhand der Interkorrelationen der Faktoren. Bezogen auf unser Beispiel weist das Modell eine sehr gute Modellanpassung auf (vgl. Tab. 1). Bei der Bewertung der Freiheitsgrade ist zu beachten,

dass die Freiheitsgrade beim WLSMV-Schätzer nicht in der üblichen Weise, sondern anhand einer komplexen Formel bestimmt werden (vgl. hierzu Muthén 1998-2004). Die geschätzten Parameter des Modells sind in Tabelle 2 angegeben. Die Korrelation zwischen dem Selbstberichtsfaktor und dem ersten Fremdberichtsfaktor beträgt r = 0.31, die zwischen dem Selbstberichtsfaktor und dem zweiten Fremdberichtsfaktor r = 0.29, die beiden Fremdberichtsfaktoren korrelieren zu r = 0.44 miteinander. Die Ergebnisse zeigen, dass die konvergente Validität bei den beiden Fremdberichten höher ist als die zwischen Selbst- und Fremdberichten. Die beiden Fremdrater sind einander ähnlicher als die Selbst- und Fremdrater. Die Höhe der konvergenten Validität liegt in einem Bereich, der für emotionale Konstrukte zu erwarten ist (Watson et al. 2000).

	Schwellenparameter				Ladungen
	$\hat{\kappa}_{1ijk}$	$\hat{\kappa}_{2ijk}$	$\hat{\kappa}_{3ijk}$	$\hat{\kappa}_{4ijk}$	$\hat{\lambda}_{ijk}$
Selbstbericht					
Zufrieden	-4.46	-2.36	-0.60	1.74	1.00
Wohl	-4.69	-3.13	-1.13	1.68	1.17
Glücklich	-4.71	-2.61	-0.71	1.76	1.18
Fremdbericht A					
Zufrieden	-3.74	-2.12	-0.42	2.53	1.00
Wohl	-4.58	-2.78	-0.95	2.28	1.13
Glücklich	-4.11	-2.00	-0.39	1.73	0.83
Fremdbericht B					
Zufrieden	-4.11	-2.00	-0.39	1.73	1.00
Wohl	-4.20	2.17	-0.53	2.37	0.88
Glücklich	-4.79	-2.41	-1.03	1.72	0.99

Varianzen, Kovarianzen und Korrelationen (kursiv) der Faktoren			
	T_{j1}	T_{j2}	T_{j3}
T_{j1}	1.85	*0.31*	*0.29*
T_{j2}	0.61	2.09	*0.44*
T_{j3}	0.58	0.96	2.29
Mittelwerte	0	0	0

Tabelle 2: Geschätzte Parameter der Modelle gemeinsamer Faktoren. (a) Modell M1: keine Restriktionen

Das Modell gemeinsamer Faktoren kann auch herangezogen werden, um zu überprüfen, ob es sich bei den beiden Fremdratern um austauschbare Methoden handelt. Ist dies der Fall, so sollten die Messmodelle der Fremdberichtsfaktoren identisch sein. Dies bedeutet, dass sich die Schwellenparameter der Items zwischen Rater A und B nicht unterscheiden dürfen. Darüber hinaus müssen die Faktorla-

	Schwellenparameter				Ladungen
	$\hat{\kappa}_{1ijk}$	$\hat{\kappa}_{2ijk}$	$\hat{\kappa}_{3ijk}$	$\hat{\kappa}_{4ijk}$	$\hat{\lambda}_{ijk}$
Selbstbericht					
Zufrieden	-4.45	-2.36	-0.60	1.74	1.00
Wohl	-4.70	-3.13	-1.13	1.68	1.17
Glücklich	-4.71	-2.61	-0.71	1.76	1.18
Fremdbericht A					
Zufrieden	-3.99	-2.17	-0.48	2.48	1.00
Wohl	-4.43	-2.55	-0.98	1.95	0.96
Glücklich	-4.43	-2.22	-0.45	1.87	0.90
Fremdbericht B					
Zufrieden	-3.99	-2.17	-0.48	2.48	1.00
Wohl	-4.43	-2.55	-0.98	1.95	0.96
Glücklich	-4.43	-2.22	-0.45	1.87	0.90

Varianzen, Kovarianzen und Korrelationen (kursiv) der Faktoren			
	T_{j1}	T_{j2}	T_{j3}
T_{j1}	1.84	*0.30*	*0.30*
T_{j2}	0.61	2.27	*0.44*
T_{j3}	0.61	0.99	2.27
Mittelwerte	0	0	0

Tabelle 2: Geschätzte Parameter der Modelle gemeinsamer Faktoren.
(b) Modell M2: Austauschbarkeit der Fremdberichte

dungen und die Erwartungswerte und Varianzen der beiden Faktoren gleich sein. Die Erwartungswerte waren in dem Modell ohne Restriktionen aus Identifikationsgründen für alle Faktoren auf den Wert Null gesetzt worden. Wenn Messäquivalenzrestriktionen auf zwei Faktoren gelegt werden, so ist es möglich, den Erwartungswert eines von ihnen zu schätzen. Aufgrund der Annahme der Austauschbarkeit werden aber die Erwartungswerte beider Faktoren weiterhin auf Null fixiert. Schließlich müssen die Kovarianzen der beiden Fremdberichtsfaktoren mit dem Selbstberichtsfaktor gleich sein. Durch die Annahme der Austauschbarkeit der Rater A und B können nun 16 Parameter weniger frei geschätzt werden (12 Schwellen, 2 Ladungen, 1 Varianz, 1 Kovarianz). Ein Modell mit diesen Restriktionen (M2) passt sehr gut auf die Daten (vgl. Tab. 1). Um zu überprüfen, ob dieses Modell zu einer signifikant schlechteren Modellanpassung führt, wurde es mit der Modellgüte des weniger restriktiven Modells verglichen. Bei geschachtelten Modellen kann ein χ^2-Differenztest zur Bestimmung des angemesseneren Modells herangezogen werden. Da die Differenz der beiden χ^2-Werte beim WLSMV-Schätzer jedoch nicht einer χ^2-Verteilung folgt, wurde mit dem Programm Mplus (Option DIFFTEST) ein

	Schwellenparameter				Ladungen
	$\hat{\kappa}_{1ijk}$	$\hat{\kappa}_{2ijk}$	$\hat{\kappa}_{3ijk}$	$\hat{\kappa}_{4ijk}$	$\hat{\lambda}_{ijk}$
Selbstbericht					
Zufrieden	-3.99	-2.20	-0.52	2.15	1.00
Wohl	-4.55	-2.73	-1.03	1.86	1.05
Glücklich	-4.52	-2.35	-0.53	1.83	1.00
Fremdbericht A					
Zufrieden	-3.99	-2.20	-0.52	2.15	1.00
Wohl	-4.55	-2.73	-1.03	1.86	1.05
Glücklich	-4.52	-2.35	-0.53	1.83	1.00
Fremdbericht B					
Zufrieden	-3.99	-2.20	-0.52	2.15	1.00
Wohl	-4.55	-2.73	-1.03	1.86	1.05
Glücklich	-4.52	-2.35	-0.53	1.83	1.00

Varianzen, Kovarianzen und Korrelationen (kursiv) der Faktoren			
	T_{j1}	T_{j2}	T_{j3}
T_{j1}	2.06	*0.34*	*0.34*
T_{j2}	0.71	2.06	*0.34*
T_{j3}	0.71	0.71	2.06
Mittelwerte	0	0	0

Tabelle 2: Geschätzte Parameter der Modelle gemeinsamer Faktoren.
(c) Modell M3: Austauschbarkeit der Selbst- und Fremdberichte

adäquater Differenztest durchgeführt. Dessen Ergebnis (vgl. Tab. 1) zeigt, dass das restriktivere Modell M2 keine signifikant schlechtere Modellanpassung als das nicht-restringierte Modell M1 aufweist. Die Annahme der Austauschbarkeit muss somit nicht verworfen werden. Die geschätzten Modellparameter sind in Tabelle 2 angegeben.

In einem nächsten Schritt haben wir überprüft, ob alle drei Methoden als austauschbar angesehen werden können, indem den Selbstberichtsvariablen dieselben Gleichheitsrestriktionen wie den Fremdberichten auferlegt werden (M3). Diese Annahme muss erwartungsgemäß auf einem α-Niveau von 0.05 verworfen werden (vgl. Tab. 1). Der adäquate Differenztest ergab, dass das Modell M3 signifikant schlechter auf die Daten passt als das Modell M2 (vgl. Tab. 1). Die theoretische Annahme, dass sich die Selbstberichtsmethode strukturell von den beiden austauschbaren Fremdberichtsmethoden unterscheidet, kann somit empirisch gestützt werden.

	Schwellenparameter				Ladungen
	$\hat{\kappa}_{1ijk}$	$\hat{\kappa}_{2ijk}$	$\hat{\kappa}_{3ijk}$	$\hat{\kappa}_{4ijk}$	$\hat{\lambda}_{ijk}$
Selbstbericht					
Zufrieden	-3.95	-2.14	-0.43	2.30	1.00
Wohl	-4.42	-2.63	-0.94	1.96	1.03
Glücklich	-4.41	-2.26	-0.44	1.92	0.98
Fremdbericht A					
Zufrieden	-3.95	-2.14	-0.43	2.30	1.00
Wohl	-4.42	-2.63	-0.94	1.96	1.03
Glücklich	-4.41	-2.26	-0.44	1.92	0.98
Fremdbericht B					
Zufrieden	-3.95	-2.14	-0.43	2.30	1.00
Wohl	-4.42	-2.63	-0.94	1.96	1.03
Glücklich	-4.41	-2.26	-0.44	1.92	0.98

Varianzen. Kovarianzen und Korrelationen (kursiv) der Faktoren			
	T_{j1}	T_{j2}	T_{j3}
T_{j1}	2.30	*0.30*	*0.30*
T_{j2}	0.64	2.04	*0.44*
T_{j3}	0.64	0.89	2.04
Mittelwerte	0.28	0	0

Tabelle 2: Geschätzte Parameter der Modelle gemeinsamer Faktoren.
(d) Modell M4: Austauschbarkeit der Fremdberichte, Messäquivalenz mit Selbstbericht

In einem weiteren Schritt haben wir überprüft, ob zwischen dem Selbstberichtsfaktor und den Fremdberichtsfaktoren Messäquivalenz besteht (M4). Der Unterschied zwischen dem Test auf Austauschbarkeit und dem auf Messäquivalenz besteht darin, dass bei der Messäquivalenz vorausgesetzt wird, dass die Schwellenwerte und die Faktorladungen sich nicht zwischen den Faktoren unterscheiden, der Erwartungswert und die Varianz des Selbstberichtsfaktors sich aber von den entsprechenden Parametern der Fremdberichtsfaktoren unterscheiden dürfen. Ein solchermaßen spezifiziertes Modell passt sehr gut auf die Daten und nicht signifikant schlechter als das Modell mit der Annahme der Austauschbarkeit der beiden Fremdrater (vgl. Tab. 1). Die geschätzten Parameter dieses Modells sind ebenfalls in Tabelle 2 zusammengestellt. Die Schwellenwerte und die Ladungen der drei verschiedenen Items sind sehr ähnlich, so dass nur geringe Unterschiede zwischen ihnen bestehen. Die Varianz des Selbstberichtsfaktors ist etwas größer als die Varianzen der Fremdberichtsfaktoren. Wie wir bereits oben gesehen haben, korrelieren die beiden Fremdberichtsfaktoren höher miteinander als die Fremdberichtsfaktoren mit dem Selbstberichtsfaktor. Der geschätzte Erwartungs-

wert des Selbstberichtsfaktors beträgt 0.28. Das selbstberichtete Wohlbefinden ist also im Mittel höher als das fremdberichtete. Die Erwartungswerte der Fremdberichtfaktoren sind aus Identifikationsgründen auf 0 fixiert.

Der Vorteil des Modells gemeinsamer Faktoren besteht darin, dass es eine Schätzung der latenten Korrelationen und eine Überprüfung der Austauschbarkeit der Methoden und der Messäquivalenz erlaubt. Es macht sich aber noch nicht die Austauschbarkeit der Fremdberichtsmethoden in vollem Umfang zunutze und enthält noch keine Variablen, die Methodeneffekte repräsentieren. Daher ist es zur Erfassung von Methodeneffekten und ihrer Erklärung nur begrenzt nutzbar.

3.2 Modell mit gemeinsamem Traitfaktor für die austauschbaren Methoden

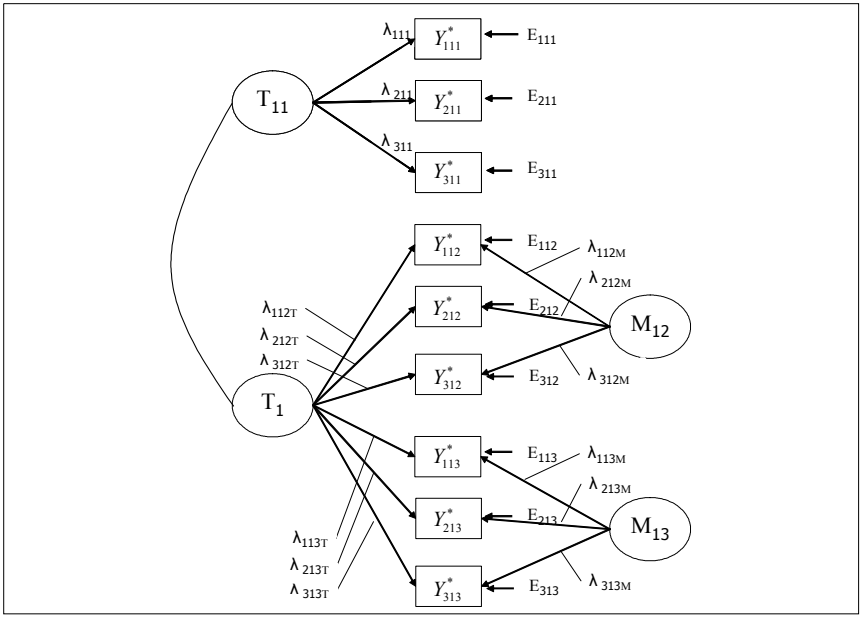


Abbildung 2: Modell mit einem gemeinsamen Traitfaktor für die austauschbaren Fremdrater. Y^*_{ijk} bezeichnet die latenten Antwortvariablen, die den beobachteten Variablen zugrunde liegen. T_{j1} = Traitfaktor der Selbstberichtsmethode; T_j = gemeinsamer Traitfaktor der Fremdberichtsmethode; M_{jk} = Methodenfaktoren. Der erste Index i (i = 1, …, m) kennzeichnet den Indikator einer Trait-Methoden-Einheit, der zweite Index j (j = 1, …, n) indiziert das betrachtete Merkmal (Trait), der letzte Index k gibt die Methode (k = 1, …, o) an.

Die Austauschbarkeit der beiden Fremdrater kann genutzt werden, um mehr Informationen über die Konvergenz der Methoden zu erhalten, insbesondere um einen gemeinsamen Traitfaktor für die Fremdberichte zu definieren. Machen wir uns dies anhand der Mehrebenenstruktur deutlich. Die beiden Fremdrater werden aus

der Population der potenziellen Fremdrater gezogen. Im Mehrebenenmodell lässt sich jede Zielperson (Level-2-Einheit) durch einen Parameter kennzeichnen, der den erwarteten Fremdberichtswert (das mittlere Urteil über alle potenziellen Rater hinweg) erfasst (group mean in der Begrifflichkeit der Mehrebenenanalyse). Die Werte der einzelnen Fremdurteile schwanken um diesen Parameter, der die Merkmalsausprägung der Person über alle Rater hinweg kennzeichnet. Im faktorenanalytischen Modell wird diese Level-2-Variable (erwartete Fremdberichtsurteile) durch einen gemeinsamen Faktor repräsentiert. Die Abweichungen der wahren Werte der einzelnen Fremdrater von den Werten dieses gemeinsamen Faktors sind Werte eines raterspezifischen Residualfaktors. Ein solches Modell ist in Abbildung 2 dargestellt. Es erlaubt nun die konvergente Validität auf der Ebene des gemeinsamen Faktors zu bestimmen, der die gemeinsame Sicht der beiden Fremdrater repräsentiert. Es ist zu erwarten, dass die konvergente Validität – gemessen über die Korrelation des Selbstberichtsfaktors und des gemeinsamen Fremdberichtsfaktors – größer ausfällt, da der gemeinsame Faktor frei von spezifischen Effekten der einzelnen Fremdrater ist. Eine höhere Korrelation spiegelt den Aggregationseffekt wider, den man aus der Persönlichkeitsforschung kennt (Epstein 1983): Durch die Ausmittelung der Werte verschiedener Rater bei der Aggregation werden spezifische Anteile reduziert, wodurch eine höhere Korrelation zu erwarten ist. In dem MTMM-Modell werden die Werte zwar nicht konkret ausgemittelt, der gemeinsame Faktor repräsentiert jedoch den gemeinsamen Anteil an den Fremdurteilen, der von spezifischen Effekten "befreit" ist.

Bezeichnen wir die Selbstberichtsmethode als erste Methode (k = 1), so lässt sich dieses Modell für ein Merkmal (Trait) j formal wie folgt schreiben:

$$Y^{*}_{ij1} = \lambda_{ij1} T_{j1} + E_{ij1} \tag{8}$$

und

$$Y^{*}_{ijk} = \lambda_{ijkT} T_{j} + \lambda_{ijkM} M_{jk} + E_{ijk} \tag{9}$$

für $k \neq 1$.

Die latente Variable T_{j1} ist der Traitfaktor der Selbstberichtsmethode. Die Variable T_j ist der gemeinsame Traitfaktor der Fremdberichtsmethode, er repräsentiert die Anteile der Fremdberichte, die über die Methoden generalisieren. Die latenten Variablen M_{jk} sind Methodenfaktoren. Sie repräsentieren die Abweichungen der einzelnen Fremdrater von ihrem gemeinsamen Traitwert. Sie stehen also für den Teil eines Fremdurteils, das nicht mit dem anderen Fremdurteil geteilt wird. Ist die Varianz der Methodenfaktoren groß, zeigt dies an, dass sich die Fremdrater stark unterscheiden. Die Traitvariablen können korreliert sein, während angenommen wird, dass die Methodenfaktoren, die raterspezifische Aspekte erfassen, mit den Traitfaktoren unkorreliert sind. Ebenfalls wird angenommen, dass die Methodenfaktoren untereinander unkorreliert sind, da sie die Anteile eines Fremdberichts repräsentieren, die spezifisch für einen Fremdrater sind, d. h. diejenigen Anteile wiederspiegeln, die nicht über die Fremdrater hinweg generalisieren. Die Kovari-

anzen zwischen den Traitfaktoren und den Methodenfaktoren wie auch die zwischen den Methodenfaktoren müssen daher auf Null fixiert werden. Darüber hinaus muss pro Faktor entweder eine Ladung auf einen Wert ungleich Null oder die Faktorvarianz auf einen Wert ungleich Null fixiert werden.

Es wird die Austauschbarkeitsbedingung angenommen, dass alle Parameter für die beiden Fremdrater gleich sind. Darüber hinaus kann auch in diesem Modell eine Messäquivalenzbedingung formuliert werden, die sich auf die Gleichheit der Schwellen und der Ladungen auf den Traitfaktoren zwischen Selbst- und Fremdbericht bezieht. Ist diese erfüllt, kann wiederum der Erwartungswert des Traitfaktors des Selbstberichts geschätzt werden. Ansonsten ist der Erwartungswert des Faktors T_j aus Identifikationsgründen auf Null fixiert. Die Erwartungswerte der Methodenfaktoren M_{jk} müssen aus theoretischen Gründen gleich Null sein (Eid et al. 2008).

In unserem Beispiel passt das Modell mit der Annahme der Austauschbarkeit der Fremdrater und der Annahme der Messäquivalenz zwischen allen Ratern sehr gut [χ^2 (df = 51, n = 482) = 55.07, p = .32, CFI = 1.00, RMSEA = 0.01]. Da sich die geschätzten Schwellenparameter nur unwesentlich von denen des Modells M4 unterscheiden, sind in Tabelle 3 nur die geschätzten Ladungskoeffizienten sowie die Varianzen, Kovarianzen und Korrelationen der latenten Variablen angegeben. Die Korrelation zwischen dem gemeinsamen Traitfaktor der Fremdberichte mit dem Traitfaktor des Selbstberichts ist mit r = 0.45 – wie erwartet – größer als die des Selbstberichtsfaktors mit den einzelnen Fremdberichtsfaktoren in Modell M4. Der geschätzte Erwartungswert des Selbstberichtsfaktors beträgt 0.28 und ist somit etwas größer als der auf Null fixierte Mittelwert des gemeinsamen Traitfaktors der Fremdberichte.

	Ladungen			Kon[1]	MS[1]
	$\hat{\lambda}_{ij1}$	$\hat{\lambda}_{ijkT}$	$\hat{\lambda}_{ijM}$		
Zufrieden	1.00	1.00	1.00	0.45	0.55
Wohl	1.01	1.01	1.12	0.40	0.60
Glücklich	0.98	0.98	0.96	0.46	0.54

Varianzen, Kovarianzen und Korrelationen (kursiv) der Faktoren				
	T_{j1}	T_{j2}	M_{j2}	M_{j3}
T_{j1}	2.32	*0.45*	*0.00*	*0.00*
T_j	0.65	0.90	*0.00*	*0.00*
M_{j2}	0.00	0.00	1.10	*0.00*
M_{j3}	0.00	0.00	0.00	1.10
Mittelwerte	0.28	0	0	0

[1] Konsistenz- und Methodenspezifität beziehen sich auf die Items der Fremdberichte

Tabelle 3: Geschätzte Parameter des Modells mit gemeinsamem Traitfaktor für die austauschbaren Methoden. Kon = Konsistenz, MS = Methodenspezifität

Die Ladungen und die Varianzen der latenten Variablen können herangezogen werden, um das Ausmaß der Methodenspezifität der einzelnen Items bei den Fremdratern zu bestimmen. Die wahre (nicht messfehlerbedingte) Varianz einer Fremdberichtsvariablen ist gleich der Summe der mit den quadrierten Ladungen gewichteten Faktorvarianzen, also: $\lambda^2_{ijkT}\mathrm{Var}(T_j)+\lambda^2_{ijkM}(M_{jk})$. Teilt man die Varianz, die durch den gemeinsamen Traitfaktor bestimmt wird, durch die wahre Gesamtvarianz, erhält man einen Varianzanteil, der die Konsistenz (Kon) über die verschiedenen Fremdberichte hinweg kennzeichnet:

$$\mathrm{Kon}(\mathrm{Fremdbericht})_{ijk} = \frac{\lambda^2_{ijkT}\mathrm{Var}(T_j)}{\lambda^2_{ijkT}\mathrm{Var}(T_j)+\lambda^2_{ijkM}\mathrm{Var}(M_{jk})}. \quad (10)$$

Das Gegenteil ist der Methodenspezifitätskoeffizient (MS) eines Fremdberichts, der den Anteil an der wahren Gesamtvarianz widerspiegelt, der spezifisch für einen Fremdbericht ist:

$$\mathrm{MS}(\mathrm{Fremdbericht})_{ijk} = \frac{\lambda^2_{ijkM}\mathrm{Var}(M_{jk})}{\lambda^2_{ijkT}\mathrm{Var}(T_j)+\lambda^2_{ijkM}\mathrm{Var}(M_{jk})}. \quad (11)$$

Beide Koeffizienten summieren sich zu Eins auf. Je größer der Konsistenzkoeffizient im Vergleich zum Methodenspezifitätskoeffizienten ist, umso größer ist die Konvergenz der Fremdberichte. Für die verschiedenen Items sind die Konsistenz- und Methodenspezifitätskoeffizienten in Tabelle 3 angegeben. Sie zeigen, dass die Methodenspezifitätskoeffizienten etwas größer sind als die Konsistenzkoeffizienten. Messfehlerfreie Unterschiede in Fremdberichten spiegeln etwas stärker raterspezifische als gemeinsame Einflüsse, wobei der gemeinsame Anteil nicht zu vernachlässigen ist.

3.3 CTC(M-1)-Modell für strukturell unterschiedliche und austauschbare Methoden für ein Merkmal

Das in Kapitel 3.2 vorgestellte Modell mit einem gemeinsamen Traitfaktor für austauschbare Methoden erlaubt die Messung und Quantifizierung von Methodeneinflüssen auf der Ebene der einzelnen Fremdrater. Es enthält jedoch noch keine latente Variable, die den gemeinsamen Methodeneinfluss der austauschbaren Rater im Vergleich zum Selbstbericht repräsentiert. Dies ermöglicht das CTC(M-1)-Modell für strukturell unterschiedliche und austauschbare Methoden (Nussbeck et al. 2009), das ursprünglich zur Analyse von MTMM-Daten mit strukturell unterschiedlichen Methoden entwickelt wurde (Eid 2000, Eid et al. 2003, Eid et al. 2008). Dessen Grundidee besteht darin, Methoden miteinander zu kontrastieren (vgl. hierzu ausführlich Eid et al. 2008, Geiser et al. 2008). Dabei ist es notwendig, eine der strukturell unterschiedlichen Methoden als Referenzmethode auszuwählen. Die latente Traitvariable, die mit dieser erfasst wird, dient als unabhängige Variable in einer latenten Regressionsanalyse zur Vorhersage der Y*-Werte der beobachteten Variablen, die zu den anderen Methoden gehören (vgl. Abb. 3). In unserem Bei-

spiel wurde der Selbstbericht als Referenzmethode ausgewählt. Der darauf basierende Traitfaktor T_{j1} misst das mit dem Selbstbericht erhobene "wahre" (messfehlerfreie) Wohlbefinden. Die Bedeutung dieses Faktors hat sich somit im Vergleich zu dem Modell mit gemeinsamem Traitfaktor für austauschbare Methoden (vgl. Abb. 2) nicht geändert. Dieser Traitfaktor wird nun herangezogen, um die Fremdurteile vorherzusagen. Diese Vorhersage ist typischerweise nicht perfekt, da sich Selbst- und Fremdurteile unterscheiden. Dem CTC(M-1)-Modell für strukturell unterschiedliche und austauschbare Methoden zufolge gibt es zwei Quellen dafür: Zum einen gibt es Abweichungen, die auf die gemeinsame Sicht der Fremdrater zurückgeführt werden können, die mit den Selbstratern nicht geteilt wird. Dies ist ein gemeinsamer Methodeneinfluss, der über die Fremdrater generalisiert. In dem in Abbildung 3 dargestellten Modell wird dieser Einfluss durch den gemeinsamen Methodenfaktor CM_j repräsentiert. Zum anderen unterscheiden sich die Fremdrater auch untereinander. Hierfür stehen die spezifischen Methodenfaktoren M_{jk}, die den Teil eines Fremdberichts verkörpern, der weder mit der Zielperson noch mit den anderen Fremdratern geteilt wird und nicht auf den Messfehler zurückgeführt werden kann.

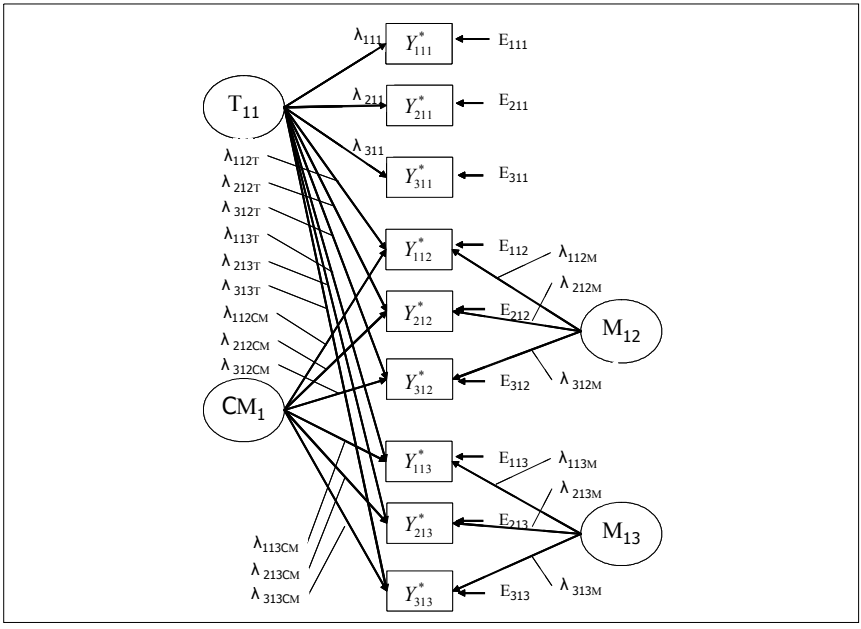


Abbildung 3: CTC(M-1)-Modell für strukturell unterschiedliche und austauschbare Rater. Y^*_{ijk} bezeichnet die latenten Antwortvariablen, die den beobachteten Variablen zugrunde liegen. T_{j1} = Traitfaktor der Selbstberichtsmethode, die als Referenzmethode gewählt wurde; CM_j = gemeinsamer Methodenfaktor der Fremdberichtsmethode; M_{jk} = spezifische Methodenfaktoren. Der erste Index i (i = 1, …, m) kennzeichnet den Indikator einer Trait-Methoden-Einheit, der zweite Index j (j = 1, …, n) indiziert das betrachtete Merkmal (Trait), der letzte Index k gibt die Methode (k = 1, …, o) an.

Formal lässt sich dieses Modell wie folgt darstellen. Bezeichnen wir wieder die Selbstberichtsmethode als erste Methode (k = 1), so ist das Modell für ein Merkmal (Trait) j durch zwei Gleichungen beschrieben:

$$Y^{*}_{ij1} = \lambda_{ij1} T_{j1} + E_{ij1} \tag{12}$$

und

$$Y^{*}_{ijk} = \lambda_{ijkT} T_{j1} + \lambda_{ijkCM} CM_{j} + \lambda_{ijkM} M_{jk} + E_{ijk} \tag{13}$$

für $k \neq 1$.

Die latente Variable T_{j1} ist der Traitfaktor des Selbstberichts, der als Referenzmethode gewählt wurde. Die Variable CM_j ist der gemeinsame Methodenfaktor der Fremdberichte, die latenten Variablen M_{jk} sind deren spezifischen Methodenfaktoren. Die λ-Koeffizienten sind die entsprechenden Ladungen auf den latenten Faktoren. Für die austauschbaren Methoden werden wieder der Austauschbarkeitsannahme zufolge, alle Parameter über über die Methoden hinweg gleichgesetzt. Da sich die Faktorenstruktur der strukturell unterschiedlichen Methoden unterscheidet, wird keine weitere Messäquivalenzannahme zwischen ihnen getroffen. Aus Identifikationsgründen muss wiederum für jeden Faktor eine Ladung fixiert werden, typischerweise auf den Wert Eins, und es wird angenommen, dass die Erwartungswerte aller Faktoren gleich Null sind. Darüber hinaus sind alle Faktoren miteinander unkorreliert. Die Fremdberichtsvariablen werden somit in einen Satz unabhängiger Einflussgrößen zerlegt, wodurch sich ihre Varianz folgendermaßen additiv zusammensetzt:

$$Var\left(Y^{*}_{ijk}\right) = \lambda^{2}_{ijkT} Var\left(T_{j1}\right) + \lambda^{2}_{ijkCM} Var\left(CM_{j}\right) + \lambda^{2}_{ijkM} Var\left(M_{jk}\right) + Var\left(E_{ijk}\right) \tag{14}$$

Aufgrund der gewählten Parametrisierung sind die Fehlervarianzen auf Eins gesetzt. Die wahre Varianz lässt sich somit in drei Quellen zerlegen: 1. Einflüsse des Traits (hier: des Traits der Selbstberichtsmethode), 2. Einflüsse des gemeinsamen Methodenfaktors (hier: der gemeinsame Anteil der Fremdberichte, den diese nicht mit dem Selbstbericht teilen), und 3. Einflüsse des spezifischen Methodenfaktors (hier: die Anteile der Fremdberichte, die weder mit dem Selbstbericht noch mit dem anderen Fremdbericht geteilt werden). Hierauf aufbauend lassen sich drei Varianzkomponenten definieren:

Der Konsistenzkoeffizient (Konsistenz mit dem Selbstbericht):

$$Kon\left(Selbstbericht\right)_{ijk} = \frac{\lambda^{2}_{ijkT} Var\left(T_{j1}\right)}{\lambda^{2}_{ijkT} Var\left(T_{j1}\right) + \lambda^{2}_{ijkCM} Var\left(CM_{j}\right) + \lambda^{2}_{ijkM} Var\left(M_{jk}\right)} \tag{15}$$

gibt den Anteil an der wahren Varianz einer Fremdberichtsvariablen an, der durch den Selbstberichts-Traitfaktor determiniert wird. Der gemeinsame Methodenspezifitätskoeffizient (GMS)

$$\text{GMS}(\text{Fremdbericht})_{ijk} = \frac{\lambda^2_{ijkCM}\text{Var}(CM_j)}{\lambda^2_{ijkT}\text{Var}(T_{j1}) + \lambda^2_{ijkCM}\text{Var}(CM_j) + \lambda^2_{ijkM}\text{Var}(M_{jk})} \quad (16)$$

spiegelt den Anteil an der wahren Varianz einer Fremdberichtsvariablen wider, der auf gemeinsame Methodeneinflüsse zurückgeht, also von den verschiedenen Fremdratern, nicht aber mit dem Selbstbericht geteilt wird. Der spezifische Methodenspezifitätskoeffizient (SMS)

$$\text{SMS}(\text{Fremdbericht})_{ijk} = \frac{\lambda^2_{ijkM}\text{Var}(M_{jk})}{\lambda^2_{ijkT}\text{Var}(T_{j1}) + \lambda^2_{ijkCM}\text{Var}(CM_j) + \lambda^2_{ijkM}\text{Var}(M_{jk})} \quad (17)$$

kennzeichnet den Anteil an der wahren Varianz einer Fremdberichtsvariablen, der darauf zurückgeführt werden kann, dass jeder Fremdrater eine eigene Sicht auf die Zielperson hat.

In unserer Anwendung weist das in Abbildung 3 dargestellte Modell eine sehr gute Modellanpassung auf [χ^2 (df = 36, n = 482) = 37.98, p = .38, CFI = 1.00, RMSEA = 0,01]. Die geschätzten Modellparameter sind in Tabelle 4 zusammengestellt. Die Konsistenz- und Methodenspezifitätskoeffizienten zeigen, dass die Hauptquelle der wahren Varianz einer Fremdberichtsvariablen der spezifische Methodenanteil ist. Fremdrater unterscheiden sich somit vor allem dadurch, dass sie eine eigene Sicht auf die Zielperson haben. Die zweitstärkste Varianzquelle ist der gemeinsame Methodenanteil. Fremdbeurteilungen stimmen bei der Beurteilung einer Zielperson in einem gewissen Maße überein, diese gemeinsame Einschätzung unterscheidet sich über die Zielpersonen hinweg. Die kleinste, aber immer noch bedeutsame Varianzquelle ist die Selbstbeurteilung. Fremdurteile (über Zielpersonen hinweg) unterscheiden sich daher auch dadurch, dass sie eine Quelle der Unterschiedlichkeit erfassen, die mit der Zielperson geteilt wird.

	Ladungen				Varianzkomponenten		
	GT		GM	SM	Kon	GMS	SMS
	SB	FB					
Zufrieden	1.00	0.33	1.00	1.00	0.09	0.38	0.52
Wohl	1.17	0.32	0.91	1.17	0.08	0.28	0.64
Glücklich	1.18	0.29	0.91	.93	0.09	0.38	0.54

	GT	GM	SM
Varianzen	1.84	0.82	1.11
Mittelwerte	0	0	0

Tabelle 4: Geschätzte Parameter des CTC(M-1)-Modells für strukturell unterschiedliche und austauschbare Methoden. GT = gemeinsamer Trait; SB = Selbstberichte; FM = Fremdberichte; GM = gemeinsamer Methodenfaktor; SM = spezifischer Methodenfaktor; Kon = Konsistenzkoeffizient; GMS = gemeinsame Methodenspezifität; SMS = spezifische Methodenspezifität

4 Modelle zur Erfassung mehrerer Merkmale mit mehreren Methoden

Die bisher beschriebenen Modelle eignen sich zur Analyse eines Merkmals (Traits) mit mehreren Methoden. Sie lassen sich leicht zur Analyse mehrerer Merkmale und somit von MTMM-Daten erweitern. Hierzu müssen sie für einzelne Merkmale zu

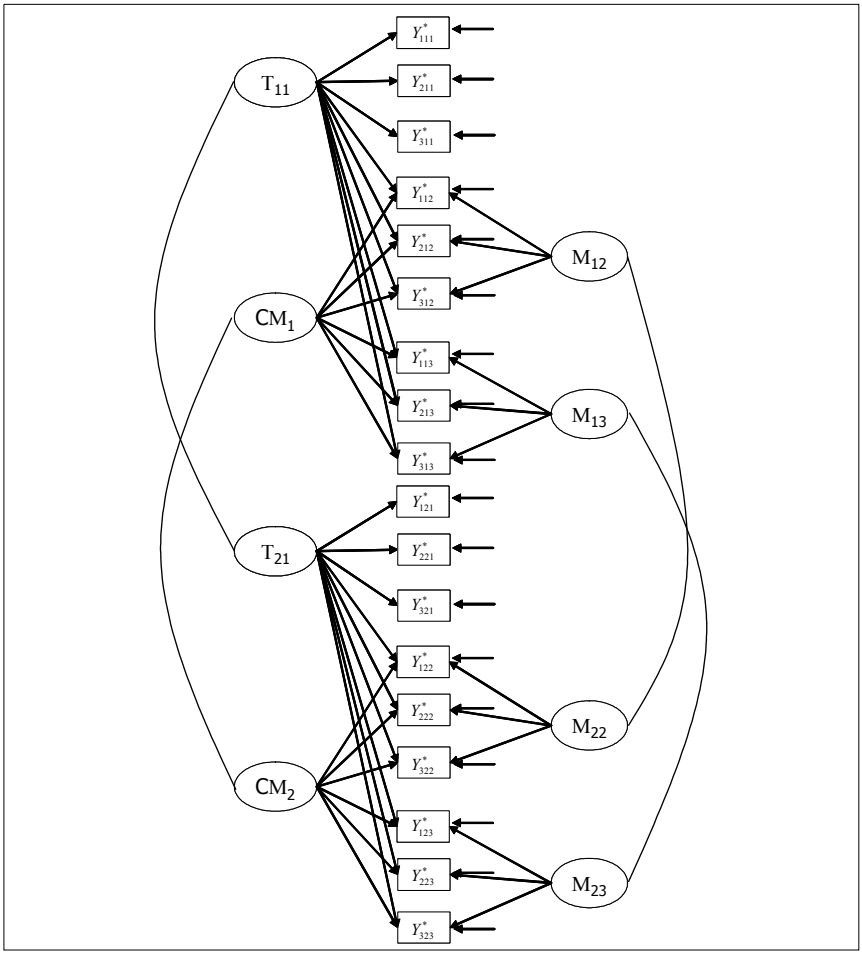


Abbildung 4: CTC(M-1)-Modell für strukturell unterschiedliche und austauschbare Rater für zwei Merkmale. Y^*_{ijk} bezeichnet die latenten Antwortvariablen, die den beobachteten Variablen zugrunde liegen. T_{j1} = Traitfaktor der Selbstberichtsmethode, die als Referenzmethode gewählt wurde; CM_j = gemeinsamer Methodenfaktor der Fremdberichtsmethode; M_{jk} = spezifische Methodenfaktoren. Der erste Index i (i = 1, …, m) kennzeichnet den Indikator einer Trait-Methoden-Einheit, der zweite Index j (j = 1, …, n) indiziert das betrachtete Merkmal (Trait), der letzte Index k gibt die Methode (k = 1, …, o) an. Aus Gründen der Lesbarkeit sind die Ladungskoeffizienten und die Residuen nicht bezeichnet.

einem Gesamtmodell verknüpft werden. Wir wollen dies nur am CTC(M-1)-Modell für strukturell unterschiedliche und austauschbare Methoden illustrieren.

In Abbildung 4 ist ein CTC(M-1)-Modell für strukturell unterschiedliche und austauschbare Methoden für zwei Merkmale (Traits) dargestellt, in dem es für jedes Merkmal ein Messmodell nach Abbildung 3 (Gleichungen 12 und 13) gibt. Die zulässigen Korrelationen haben folgende Bedeutung (siehe auch Nussbeck et al. 2009):

1. Die Korrelationen der Traitfaktoren T_{j1} zeigen die diskriminante Validität in Bezug auf die Selbstberichtsmethode an. Geringe Korrelationen weisen auf hohe diskriminante Validität hin.
2. Die Korrelationen der gemeinsamen Methodenfaktoren können zur Beurteilung der Generalisierung von gemeinsamen Methodeneffekten über die verschiedenen Merkmale herangezogen werden. Sind diese Korrelationen z. B. positiv, so bedeutet dies, dass Zielpersonen, die von ihren Fremdratern im Vergleich zu anderen Zielpersonen in Bezug auf ein Merkmal höher (bzw. niedriger) als aufgrund des Selbstberichts zu erwarten ist, eingeschätzt werden, auch bezüglich eines anderen Merkmals tendenziell höher (bzw. niedriger) eingeschätzt werden.
3. Die Korrelationen der spezifischen Methodenfaktoren über Merkmale hinweg zeigen die Generalisierung spezifischer Methodeneffekte an. Positive Korrelationen können bspw. dahingehend interpretiert werden, dass Fremdrater, die ihre Zielperson in Bezug auf ein Merkmal höher bzw. niedriger einschätzen als andere Fremdrater, dies tendenziell auch in Bezug auf ein anderes Merkmal tun. Einzelne Fremdrater könnten z. B. die Tendenz haben, die zu beurteilende Person grundsätzlich positiver oder negativer einzuschätzen.

Zur Illustrierung haben wir in unser Beispiel zusätzlich das Merkmal Wachheit aufgenommen, das über die drei Items ausgeruht, munter und frisch sowohl von den Zielpersonen als auch von den Fremdratern eingeschätzt wurde. Dieses Modell muss zwar gemäß χ^2-Test auf dem α = 0.05 Niveau verworfen werden [χ^2 (df = 147, n = 482) = 192.34, p < .01], es weist aber einen sehr gute approximative Modellanpassung auf (CFI = 0.99, RMSEA = 0.03), so dass die Ergebnisse im Folgenden interpretiert werden sollen. Für die Items zur Erfassung der habituellen Wachheit liegen die Konsistenzkoeffizienten zwischen 0.06 und 0.10, die gemeinsame Methodenspezifität zwischen 0.15 und 0.30 und die spezifischen Methodenspezifitätskoeffizienten zwischen 0.64 und 0.79. Im Vergleich zu den Wohlbefindensitems ist der Anteil der spezifischen Methodeneinflüsse größer, was daran liegen mag, dass es schwieriger ist, die Wachheit als das Wohlbefinden zu beurteilen.

Die beiden Traitvariablen korrelieren zu r = 0.77 und zeigen auf der Selbstberichtsebene eine geringe diskriminante Validität an. Die gemeinsamen Methodenfaktoren korrelieren zu r = 0.75, die generellen Methodeneffekte generalisieren somit stark über die Merkmale hinweg. Zielpersonen, deren Wohlbefinden von ihren

Fremdratern höher eingeschätzt wird als aufgrund ihrer eigenen Einschätzung zu erwarten ist, werden tendenziell auch in Bezug auf ihre Wachheit so beurteilt werden. Ähnliches gilt für die spezifischen Methodeneinflüsse. Die Korrelationen der spezifischen Methodenfaktoren betragen r = 0.70. Fremdrater, die z. B. das Wohlbefinden ihrer Zielperson höher einschätzen als andere Fremdrater, tun dies tendenziell auch bezüglich der Wachheit.

5 Diskussion

Im vorliegenden Beitrag wurde dargestellt, wie MTMM-Modelle für strukturell unterschiedliche und austauschbare Methoden auf der Grundlage der konfirmatorischen Faktorenanalyse für Variablen mit geordneten Antwortkategorien definiert werden können. Das empirische Beispiel zeigte, dass sich diese Modelle erfolgreich auf die Daten anwenden ließen und interessante Informationen über die Quellen interindividueller Unterschiede bei Fremdberichten lieferten. In allen Fällen war der raterspezifische Einfluss größer als der gemeinsame Methodeneinfluss, der wiederum größer als der des Selbstberichts war.

Den Ergebnissen zufolge muss die Annahme der Austauschbarkeit der beiden Fremdrater nicht verworfen werden, während die Annahme, dass der Selbstbericht mit den Fremdberichten austauschbar sei, nicht haltbar ist. Dies unterstreicht die Unterschiede zwischen strukturell unterschiedlichen und austauschbaren Methoden und weist darauf hin, dass die Definition eines gemeinsamen Traitfaktors, der den Fremdurteilen zugrunde liegt, sinnvoll ist. Er repräsentiert die Merkmalsausprägungen der Zielpersonen auf der Basis der Fremdurteile, die aber von raterspezifischen Einflüssen "bereinigt" sind. Da die Austauschbarkeitsannahme für den Selbstbericht verworfen werden musste, wäre es bedenklich, diesen gemeinsamen Faktor den Selbstberichten in gleicher Weise zugrunde zu legen. Dies wäre auch aus inhaltlichen Gründen problematisch: Was würde ein solcher gemeinsamer Faktor bedeuten, wenn sich die Methoden strukturell unterscheiden? Diese Problematik lässt sich anhand der Aggregation von Werten, die häufig zur Ausmittelung spezifischer Einflüsse herangezogen wird, verdeutlichen. Während die Aggregation der Fremdberichte sinnvoll ist, um einen allgemeinen Fremdbeurteilungswert zu erhalten, der die Person kennzeichnet, ist die Aggregation über Selbst- und Fremdberichte hinweg problematisch, da nicht klar ist, welcher spezifische Anteil ausgemittelt wird. An einem Beispiel illustriert: Angenommen, die Zielperson fühlt sich äußerst unglücklich, wird aber von beiden Fremdratern als sehr glücklich wahrgenommen. Wäre dann der angemessene Traitwert der Mittelwert dieser drei Urteile oder – übertragen auf ein faktorenanalytisches Modell – der Wert aus dem gemeinsamen Faktor, der ihnen zugrunde gelegt wird? Dies wäre psychologisch betrachtet sicherlich nicht sinnvoll, da in der Diskrepanz zwischen Selbst- und Fremdbericht strukturelle Unterschiede in der Sichtweise hervortreten. Das CTC(M-1)-Modell hat eine andere Perspektive darauf. Es kontrastiert die Traitfaktoren, die unterschiedlichen Methoden zugrunde liegen. Hierdurch wird es möglich, verschiedene Quellen für Methodeneinflüsse zu identifizieren. Die Methodenfaktoren re-

präsentieren die Einflüsse, die für eine nicht perfekte Konvergenz verantwortlich sind. Diese können in weitergehenden Analysen durch Kovariaten erklärt werden. In Bezug auf strukturell unterschiedliche Methoden nimmt das vorgestellte Modell eine methodenorientierte statt einer traitorientierten Perspektive ein. Wenn Methoden strukturell verschieden sind, dann sollte es darum gehen, die Methodeneffekte zu identifizieren, zu erklären und zu verstehen. Bei austauschbaren Methoden ist eine traitorientierte Sicht jedoch durchaus sinnvoll, was die Erklärung von Methodeneffekten nicht ausschließt.

In den vorgestellten Modellen wurde angenommen, dass die beobachteten Variablen zur Erfassung einer Trait-Methodeneinheit unidimensional sind. Dies muss nicht immer so sein. Gerade wenn die Selbst- und die Fremdbeurteilung anhand derselben Skalen erfolgen, ist es möglich, indikatorspezifische Anteile einer beobachteten Variablen zu identifizieren. Dies sind die Anteile, die diese mit allen anderen beobachteten Variablen gemeinsam hat, die sich auf dieselbe Skala beziehen, die aber nicht mit den anderen beobachteten Variablen, die zu den anderen Skalen gehören, geteilt wird. Liegen solche indikatorspezifischen Anteile vor, dann passen die in diesem Kapitel vorgestellten Modelle nicht auf die Daten. Sie lassen sich aber leicht auf den Fall indikatorspezifischer Einflüsse verallgemeinern. Wie dies geschehen kann, haben bspw. Eid et al. (2006a) und Eid et al. (2008) gezeigt.

Ein symmetrisches Rating-Scale-Modell für Fragebogen mit invertierten Items

Thorsten Meiser, David Hoeffler & Fabian Jasper

Das Rasch-Modell ist ein messtheoretisch fundiertes Modell der probabilistischen Testtheorie, das die Abbildung der Personenausprägung und der Itemschwierigkeit hinsichtlich einer vorgegebenen Merkmalsdimension auf einer Differenzenskala ermöglicht (Rost 2004, Kapitel 3.1). Als Merkmalsdimension können sowohl Facetten der kognitiven Leistungsfähigkeit als auch unterschiedliche Konstrukte der Persönlichkeit untersucht werden. Für die Persönlichkeitsdiagnostik ist insbesondere die Erweiterung des Rasch-Modells von dichotomen Items auf die Analyse polytomer Items mit ordinalen Antwortkategorien von Bedeutung, da Persönlichkeitsfragebogen oftmals Selbsteinschätzungen in Form von abgestuften Ratings zu vorgegebenen Iteminhalten erfordern. Rasch-Modelle für polytome Items bieten hierbei eine geeignete formale Grundlage, um Annahmen über die Prozesse zu spezifizieren und zu testen, die den beobachtbaren Antworten auf derartigen Ratingskalen zugrunde liegen (vgl. Rost 1999).

In diesem Beitrag wird die Anwendung polytomer Rasch-Modelle auf den häufig anzutreffenden Fall diskutiert, in dem ein Persönlichkeitsfragebogen sowohl positiv formulierte als auch invertierte Items enthält. Während die Iteminhalte positiv formulierter Items dem zu erfassenden Konstrukt entsprechen, repräsentieren die Inhalte invertierter Items das Gegenteil des zu erfassenden Konstrukts. Beispielsweise stimmt die Aussage "Ich bin gerne mit vielen Leuten zusammen" mit der Ausrichtung einer Skala zur Erfassung von Extraversion überein und bildet daher ein positiv formuliertes Item, wohingegen die Aussage "Ich bevorzuge es, die meisten Dinge allein zu tun" eher introvertiertes Verhalten beinhaltet und damit auf einer Extraversionsskala ein invertiertes Item darstellt. Invertierte Items werden in zahlreichen Fragebogen eingesetzt, um allgemeinen Antworttendenzen entgegenzuwirken und die Aufmerksamkeit für die spezifischen Iteminhalte aufrechtzuerhalten. Bei der Auswertung werden die Antworten auf invertierte Items dann rekodiert oder "umgepolt", so dass eine hohe Antwortkategorie bei invertierten Items als äquivalent zu einer niedrigen Antwortkategorie bei positiv formulierten Items betrachtet wird und umgekehrt.

Die Rekodierung von Antworten auf invertierte Items mag unproblematisch erscheinen, so lange die Auswertung und Interpretation auf der Basis der klassischen Testtheorie erfolgt und primär auf den Summenwert der Itemantworten in einem

Fragebogen zielt. Werden hingegen Annahmen über das Zustandekommen von beobachtbaren Antworten in einzelnen Items als Funktion der latenten Personeneigenschaft explizit gemacht, wie es bei einer Formalisierung durch polytome Rasch-Modelle geschieht, dann erfordert die Rekodierung von Items eine kritischere Betrachtung. Dazu werden im nächsten Abschnitt zunächst die Grundlagen von Rasch-Modellen für polytome Items mit ordinalen Antwortkategorien skizziert. Darauf aufbauend werden mögliche Probleme, die durch die Rekodierung invertierter Items verursacht werden können, diskutiert. Neben einigen alternativen Modellen wird anschließend ein symmetrisches Rating Scale-Modell als Lösungsansatz vorgestellt. Zur Illustration werden Daten zu einem deutschsprachigen Fragebogen zur Erfassung des Persönlichkeitskonstrukts Personal Need for Structure (Machunsky & Meiser 2006) reanalysiert.

1 Rasch-Modelle für polytome Items

1.1 Definition und Parametrisierung der Schwellenwahrscheinlichkeit

Einen zentralen Stellenwert bei der Formulierung von Rasch-Modellen für polytome Items nimmt das Konzept der Schwellenwahrscheinlichkeit ein. Die Schwellenwahrscheinlichkeit γ_{vix} ist definiert als die bedingte Wahrscheinlichkeit, mit der Person v, v = 1,…,n, in Item i mit m_i+1 Antwortkategorien von 0 bis m_i, i = 1,…,k, die Antwortkategorie x wählt; gegeben, dass die Person v eine der beiden Kategorien x-1 oder x wählt, wobei x = 1,…,m_i (vgl. Rost 1988b, Kapitel 3.2.2). Die Schwellenwahrscheinlichkeit wird analog zum Rasch-Modell für dichotome Items durch eine logistische Funktion beschrieben, die als Argument die Differenz aus der Personeneigenschaft und der Schwierigkeit der Schwelle enthält:

$$\gamma_{vix} = \frac{\text{prob}(X_{vi} = x)}{\text{prob}(X_{vi} = x-1) + \text{prob}(X_{vi} = x)} \qquad (1)$$

$$= \frac{\exp(\theta_v - \sigma_{ix})}{1 + \exp(\theta_v - \sigma_{ix})}$$

X_{vi} stellt hierbei die Zufallsvariable der Antwort von Person v auf Item i dar und kann ganzzahlige Werte zwischen Null und m_i annehmen. Der Schwellenparameter σ_{ix} entspricht der Schwierigkeit des Übergangs von Kategorie x-1 zu Kategorie x in dem vorliegenden Item i. Der Personenparameter θ_v repräsentiert die Ausprägung von Person v auf der zu erfassenden Merkmalsdimension.

Wie aus Gleichung (1) hervorgeht, ist die Schwellenwahrscheinlichkeit kleiner als .5, wenn der Schwellenparameter σ_{ix} größer ist als der Personenparameter θ_v. In diesem Fall ist also die Wahrscheinlichkeit, dass Person v in Antwortkategorie x-1 verbleibt, größer als die Wahrscheinlichkeit, dass Person v die höhere Kategorie x wählt. Bei abnehmenden Werten für die Personenausprägung bzw. bei steigender Schwierigkeit der Schwelle konvergiert die Schwellenwahrscheinlichkeit gegen die

untere Asymptote Null. Hingegen ist die Schwellenwahrscheinlichkeit größer als .5, wenn der Personenparameter θ_v den itemspezifischen Schwellenparameter σ_{ix} übersteigt. In diesem Fall ist also die Wahrscheinlichkeit, dass Person v die höhere Kategorie x wählt, größer als die Wahrscheinlichkeit, dass Person v in Kategorie x-1 verbleibt. Mit steigender Personenausprägung bzw. fallender Schwierigkeit der Schwelle konvergiert die Schwellenwahrscheinlichkeit gegen die obere Asymptote Eins. Nimmt der Personenparameter θ_v den gleichen Wert an wie der Schwellenparameter σ_{ix}, so beträgt die Schwellenwahrscheinlichkeit exakt .5. Für Personen mit einer Merkmalsausprägung von $\theta_v = \sigma_{ix}$ sind die bedingten Wahrscheinlichkeiten für die benachbarten Antwortkategorien x-1 und x also identisch.

Für ein Item mit m_i+1 Antwortkategorien 0,…, m_i lassen sich m_i Schwellenwahrscheinlichkeiten spezifizieren. Analog zur Itemcharakteristik-Kurve im dichotomen Rasch-Modell können die Schwellenwahrscheinlichkeiten in polytomen Rasch-Modellen für gegebene Schwellenparameter σ_{ix} als Funktion des Personenparameters θ abgebildet werden. Der obere Teil von Abbildung 1 zeigt die resultierenden Schwellen-Charakteristik-Kurven für die fünf Schwellenwahrscheinlichkeiten γ_{vi1}, …, γ_{vi5} eines fiktiven Items mit sechs Antwortkategorien 0,…,5.

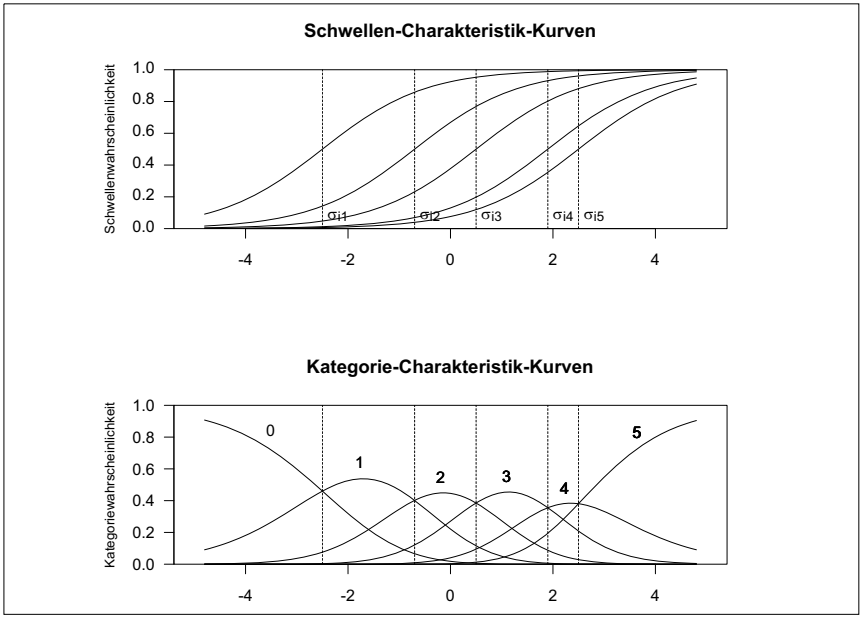


Abbildung 1: Schwellen-Charakteristik-Kurven und Kategorie-Charakteristik-Kurven für ein fiktives Item mit sechs Antwortkategorien und Schwellenparametern σ_{i1} = -2.5, σ_{i2} = -0.7, σ_{i3} = 0.5, σ_{i4} = 1.9 und σ_{i5}=2.5

Wie oben ausgeführt, sind die Schwellenwahrscheinlichkeiten als bedingte Wahrscheinlichkeiten einer Antwortkategorie x bei gegebener Beobachtung von Kategorie x-1 oder x definiert. Aus ihnen lassen sich jedoch die unbedingten Katego-

rienwahrscheinlichkeiten berechnen, d. h. die Wahrscheinlichkeiten der m_i+1 Antwortkategorien eines Items i mit einem Vektor von m_i Schwellenparametern. Die Wahrscheinlichkeit, mit der Person v in Item i die Antwortkategorie x, $x = 0,\ldots,m_i$, zeigt, resultiert dabei als

$$\text{prob}(X_{vi} = x) = \frac{\exp\left(x\theta_v - \sum_{j=1}^{x}\sigma_{ij}\right)}{1 + \sum_{s=1}^{m_i}\exp\left(s\theta_v - \sum_{j=1}^{s}\sigma_{ij}\right)} \tag{2}$$

(zur Herleitung vgl. z. B. Rost, 2004, Kapitel 3.3). Mit der Definition $\sigma_{i0} = 0$ kann die Darstellung der Kategorienwahrscheinlichkeiten zu

$$\text{prob}(X_{vi} = x) = \frac{\exp\left(x\theta_v - \sum_{j=1}^{x}\sigma_{ij}\right)}{\sum_{s=0}^{m_i}\exp\left(s\theta_v - \sum_{j=0}^{s}\sigma_{ij}\right)} \tag{3}$$

vereinfacht werden. Der untere Teil von Abbildung 1 zeigt deren Verlauf als Funktion des Personenparameters θ für das fiktive Beispielitem mit sechs Antwortkategorien.

In polytomen Rasch-Modellen kann ein streng monotoner Anstieg der Schwellenparameter, $\sigma_{i1} < \sigma_{i2} < \ldots < \sigma_{im_i}$, als empirisches Kriterium für die theoretische Annahme betrachtet werden, dass die Antwortkategorien 0, 1,..., m_i ordinale Information enthalten, so dass die Wahl einer höheren Antwortkategorie eine stärkere Ausprägung der latenten Eigenschaft widerspiegelt als die Wahl einer niedrigeren Antwortkategorie (für eine ausführliche Begründung vgl. Andrich 1985, 2010; Andrich et al. 1997). Die strenge Monotonie der Schwellenparameter innerhalb eines Items stellt zusätzlich sicher, dass es für jede Antwortkategorie ein Intervall des Personenparameters gibt, in dem die Kategorie die modale Auftretenswahrscheinlichkeit aufweist (vgl. Abb. 1 unten).

1.2 Spezielle Rasch-Modelle für ordinale Antwortkategorien

Die Spezifikationen der Schwellen- und Kategorienwahrscheinlichkeiten in den Gleichungen (1) bis (3) entsprechen dem allgemeinen eindimensionalen Rasch-Modell für polytome Items mit ordinalen Antwortkategorien. Für die folgenden Spezialfälle dieses allgemeinen Modells bietet sich eine Reparametrisierung der Schwellenparameter σ_{ix} durch eine kategorieunabhängige Itemschwierigkeit β_i und einen itemspezifischen Kategorienparameter τ_{ix} an:

$$\sigma_{ix} = \beta_i + \tau_{ix}\,. \tag{4}$$

Dabei repräsentiert β_i die allgemeine Schwierigkeit des Items i und ist formal als Mittelwert der Schwellenparameter $\sigma_{i1},\ldots,\sigma_{im_i}$ definiert. Die itemspezifischen Ka-

tegorienparameter τ_{ix} stellen die Abweichungen der einzelnen Schwellenparameter σ_{ix} von ihrem Mittelwert β_i dar und summieren sich daher für jedes Item zu Null auf:

$$\beta_i = \frac{\sum_{j=1}^{m_i} \sigma_{ij}}{m_i} \quad \text{und} \quad \sum_{j=1}^{m_i} \tau_{ix} = 0\,. \tag{5}$$

Sofern keine Verteilungsannahme über den Personenparameter θ getroffen wird, muss eine zusätzliche Restriktion eingeführt werden, um Identifizierbarkeit der Modellparameter herzustellen. Hierzu wird im Allgemeinen die Summennormierung der Itemparameter, $\sum_{i=1}^{k} \beta_i = 0$, verwendet.

Für die Analyse von Items mit ordinalen Antwortkategorien wurden spezielle Rasch-Modelle vorgeschlagen, die durch unterschiedliche Restriktionen der itemspezifischen Kategorienparameter τ_{ix} in Gleichung (4) hervorgehen. Werden keine Restriktionen eingeführt, so dass die Kategorienparameter innerhalb jedes Items und zwischen den Items beliebig variieren können, so resultiert das Partial-Credit-Modell (PCM; Masters 1982, Masters & Wright 1997). Das PCM wurde zunächst für Leistungsitems vorgeschlagen, die entweder eine Sequenz von Lösungsschritten erfordern oder mehrere Antwortalternativen vorgeben, die hinsichtlich ihrer Nähe zur korrekten Lösung abgestuft sind. Gleichzeitig gilt es als hierarchisch übergeordnetes Rahmenmodell für die nachfolgenden Spezialfälle zur Analyse von Fragebogenitems mit ordinalem Antwortformat.

Werden in einem Fragebogen k Items mit unterschiedlichen Iteminhalten, aber identischen Antwortkategorien 0,…, m vorgegeben, so ist die Annahme naheliegend, dass sich die Items hinsichtlich ihrer Itemschwierigkeiten β_i voneinander unterscheiden, während die relative Bedeutung der einzelnen Antwortkategorien über die Items konstant bleibt. D. h., durch die unterschiedlichen Iteminhalte variieren die Items hinsichtlich ihrer generellen Lokation auf dem latenten Kontinuum, aber die Abweichungen der einzelnen Schwellenparameter von ihrem Mittelwert sowie voneinander bleiben aufgrund des identischen Antwortformats über die Iteminhalte hinweg unverändert. Diese Annahme wird durch das Rating-Scale-Modell (RSM; Andrich 1978) ausgedrückt, das durch die Restriktion

$$\tau_{ix} = \tau_x \tag{6}$$

aus dem PCM hervorgeht. Der Schwellenparameter σ_{ix} setzt sich im RSM daher additiv aus dem Schwierigkeitsparameter des Items β_i und einem itemunabhängigen Kategorienparameter τ_x zusammen. Als Konsequenz sind die Abstände zwischen benachbarten Schwellenparametern σ_{ix} und $\sigma_{i(x+1)}$ für alle Items gleich, $\sigma_{i(x+1)} - \sigma_{ix} = \tau_{(x+1)} - \tau_x$.

Eine alternative Modellvorstellung besteht darin, dass die Schwellen innerhalb jedes Items äquidistant sind, während sich die Schwellenabstände zwischen den Items aufgrund einer itemspezifischen Dispersion unterscheiden können. Diese Annahme beinhaltet das Dispersions-Modell (DM; Andrich 1982), das mit der Restriktion

$$\tau_{ix} = \left(x - \frac{m+1}{2} \right) \delta_i \tag{7}$$

als Spezialfall des PCM dargestellt werden kann (vgl. Rost 1988b, S. 178f.). Der Parameter δ_i repräsentiert den Abstand benachbarter Schwellen in Item i. Setzt man den Dispersionsparameter δ_i über die Items hinweg konstant, so resultiert ein Spezialfall des RSM, bei dem die Schwellenabstände sowohl zwischen als auch innerhalb der Items identisch sind.

Eine alternative Kombination aus dem RSM und dem DM ergibt sich durch die Hinzunahme eines itemspezifischen Dispersionsparameters im RSM (Rost 1988a):

$$\tau_{ix} = \tau_x + \left(x - \frac{m+1}{2} \right) \delta_i \tag{8}$$

Dieses Rating-Scale-Dispersion-Modell (RSDM) drückt die Annahmen aus, dass das einheitliche Antwortformat einerseits zu itemunabhängigen Kategorienparametern τ_x führt und dass die einzelnen Items andererseits neben ihrer Schwierigkeit auch durch eine spezifische Dispersion δ_i charakterisiert sind. Dieses Modell stellt damit eine Erweiterung des RSM und des DM und gleichzeitig einen Spezialfall des PCM dar.

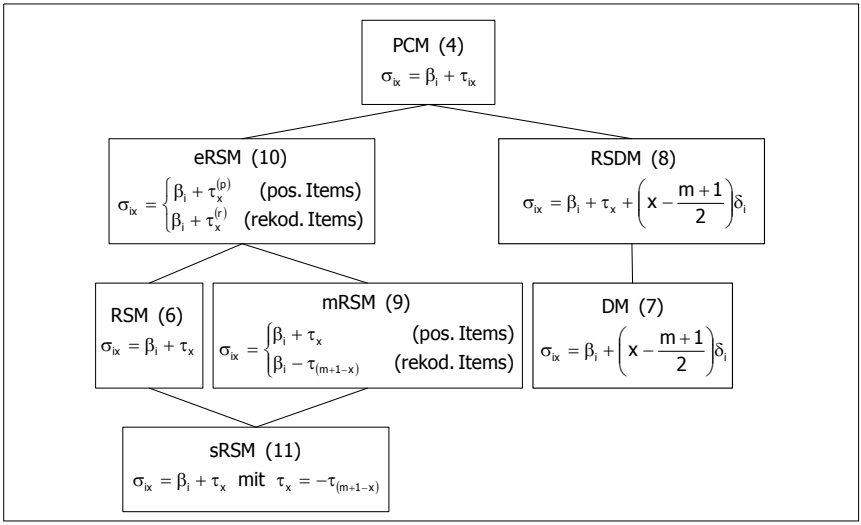


Abbildung 2: Spezifikation der Schwellenparameter σ_{ix} in den verschiedenen Rasch-Modellen für polytome Items mit ordinalem Antwortformat. Linien zwischen den Modellen zeigen Hierarchien an, in denen das untere Modell durch Restriktionen aus dem oberen hervorgeht. PCM = Partial-Credit-Modell; eRSM = erweitertes Rating-Scale-Modell mit separaten Kategorienparametern für positiv formulierte und rekodierte invertierte Items; RSDM = Rating-Scale-Dispersion-Modell; mRSM = modifiziertes Rating-Scale-Modell mit gespiegelten Kategorienparametern für positiv formulierte und rekodierte invertierte Items; RSM = ursprüngliches Rating-Scale-Modell; DM = Dispersion-Modell; sRSM = symmetrisches Rating-Scale-Modell mit punktsymmetrischen Kategorienparametern relativ zum Mittelpunkt der Ratingskala

Die bislang skizzierten Modelle für polytome Items mit ordinalem Antwortformat verdeutlichen, dass sich der formale Rahmen des Rasch-Modells dazu eignet, theoretische Annahmen über die Interpretation und Nutzung von Ratingskalen in Fragebogen zu formulieren und zu testen. So impliziert das RSM, dass das Antwortformat unabhängig vom Iteminhalt interpretiert und über die Items hinweg konstant genutzt wird, so dass die Schwellenabstände auf dem latenten Kontinuum über die Items hinweg invariant bleiben. Hingegen impliziert das DM, dass die Antwortkategorien innerhalb jedes Items als äquidistant wahrgenommen und verwendet werden, während sich die Schwellenabstände über die Items hinweg unterscheiden können. Abbildung 2 bietet eine Übersicht über die Festlegung der Schwellenparameter σ_{ix} in den Rasch-Modellen, die in diesem Abschnitt und in den folgenden Abschnitten dargestellt sind. Verbindungslinien zwischen den einzelnen Modellen symbolisieren Modellhierarchien, bei denen weiter oben abgebildete Modelle die weiter unten abgebildeten Modelle als Spezialfälle enthalten.

1.3 Erweiterungen polytomer Rasch-Modelle

Die unterschiedlichen Annahmen hinsichtlich der Antwortkategorien und Schwellenabstände wurden auch auf andere Modellfamilien übertragen. Während sich das PCM, das RSM, das DM sowie deren Integration zum RSDM zunächst auf das Rasch-Modell mit kontinuierlicher Personenvariable θ beziehen, wurden analoge Parametrisierungen der Schwellenabstände für ordinale Antwortkategorien in Latent-Class-Modellen vorgeschlagen, bei denen die kategoriale Klassenzugehörigkeit die latente Personenvariable bildet (Rost 1988b, 1988c). Darüber hinaus wurden die oben beschriebenen Rasch-Modelle für polytome Items als Komponenten in diskreten Mischverteilungsmodellen formuliert, so dass sich die Item- und Kategorienparameter zwischen latenten Subpopulationen unterscheiden können (Rost 1991, von Davier & Rost 1995). Weitere Möglichkeiten der Spezifikation und Prüfung von Annahmen bezüglich der Item- oder Schwellenparameter ergeben sich aus dem linearen Rating-Scale-Modell (Fischer & Parzer 1991) bzw. dem linearen Partial-Credit-Modell (Fischer & Ponocny 1994, 1995). Hierbei werden die Itemparameter des RSM bzw. die itemspezifischen Schwellenparameter des PCM als Linearkombinationen elementarer Basisparameter spezifiziert, so dass theoretisch motivierte Annahmen flexibel als Restriktionen in das Modell integriert und empirisch geprüft werden können. So können etwa das RSM und die im folgenden Abschnitt dargestellten Varianten dieses Modells durch geeignete Festlegung der Linearkombinationen von Basisparametern im Rahmen des linearen Partial-Credit-Modells formuliert werden.

2 Probleme bei der Analyse invertierter Items mit Rekodierung

Wenn positiv formulierte Items und invertierte Items mit entsprechender Rekodierung gemeinsam analysiert werden, können einige der zuvor skizzierten Rasch-Modelle für polytome Items zu schwerwiegenden Fehlspezifikationen führen (Rost

2004, S. 223). Diese Gefahr besteht insbesondere dann, wenn die Interpretation und Nutzung des Antwortformats vonseiten der Testpersonen unabhängig vom Iteminhalt ist, wenn unterschiedliche Schwellenabstände innerhalb der Items vorherrschen und wenn gleichzeitig das festgelegte Rasch-Modell itemunabhängige Komponenten der Schwellenabstände postuliert, wie dies etwa beim RSM von Andrich (1978) und bei der Erweiterung zum RSDM von Rost (1988a) der Fall ist.

2.1 Nutzung des Antwortformats und Antwortstile

Die Nutzung des Antwortformats, das in Persönlichkeitsfragebogen meist aus einer mehrstufigen Ratingskala besteht, hängt von der subjektiven Interpretation der Antwortkategorien sowie von der generellen Bevorzugung bestimmter Kategorien ab. Die Interpretation der Antwortkategorien bezieht sich auf die wahrgenommene Unterschiedlichkeit der einzelnen Antwortalternativen im Hinblick auf die zu erfassende Eigenschaftsdimension, die bei Gültigkeit des eindimensionalen Rasch-Modells durch die Schwellenabstände repräsentiert wird. Je größer der Abstand zwischen zwei benachbarten Schwellen ist, desto größer ist das Intervall auf der latenten Eigenschaftsdimension, in dem die dazwischen liegende Kategorie die modale Antwortwahrscheinlichkeit aufweist (vgl. Abb. 1). Werden die Kategorien als sehr ähnlich wahrgenommen, sollten die Schwellen daher dicht beieinander liegen, so dass bereits kleine Unterschiede der Personeneigenschaft den Wechsel in eine höhere Kategorie wahrscheinlich machen. Werden die Kategorien hingegen als sehr unterschiedlich interpretiert, sollten die Abstände zwischen den Schwellen groß sein, so dass ein deutlicher Zuwachs der zugrunde liegenden Personeneigenschaft erforderlich ist, bis die Wahrscheinlichkeit der nächsthöheren Kategorie dominiert. Aufgrund der Gleichheit der vorgegebenen Antwortalternativen über die Items eines Fragebogens bietet das RSM ein naheliegendes Analysemodell, da die Annahme itemunabhängiger Kategorienparameter τ_x impliziert, dass die Schwellenabstände über unterschiedliche Items hinweg konstant sind. Selbstverständlich stellt diese Invarianz der Schwellenabstände zwischen Items jedoch lediglich eine empirisch testbare Hypothese dar, die etwa verletzt ist, wenn die Bedeutung der Antwortkategorien mit dem Iteminhalt variiert.

Neben der wahrgenommenen Semantik der vorgegebenen Antwortalternativen beeinflussen persönliche Antwortstile die Nutzung des Antwortformats, wie etwa eine generelle Zustimmungs- bzw. Ablehnungstendenz oder die generelle Bevorzugung extremer bzw. moderater Antwortkategorien. Derartige Antwortstile bilden interindividuelle Variablen, die sich neben der zu erfassenden Persönlichkeitsdimension im Antwortverhalten widerspiegeln und relativ stabil über die Zeit sowie konsistent über unterschiedliche Inhaltsdomänen sind (etwa Billiet & Davidov 2008, Weijters et al. 2010). In mehreren Analysen wurde die Struktur des Antwortverhaltens in Fragebogen zu unterschiedlichen Persönlichkeitsdimensionen mit Hilfe von Mischverteilungs-Rasch-Modellen für polytome Items (Rost 1991, von Davier & Rost 1995) untersucht. Dabei zeigten sich Antwortstile durch subpopulationsspezifische Schwellenabstände, die insbesondere für eine Variation hinsichtlich der Bevorzugung extremer gegenüber moderaten Antwortkategorien bei den

üblicherweise verwendeten Likert-Items sprechen (Eid & Rauber 2000, Gollwitzer et al. 2005, Hinz et al. 2003, Meiser & Machunsky 2008, Rost et al. 1999). Vergleichbare Mischverteilungs-Modelle, die eine Untersuchung subpopulationsspezifischer Antwortstile auch im Falle von Thurstone-Items erlauben, wurden von Roberts et al. (2010) vorgeschlagen.

Von besonderer Relevanz für diesen Beitrag sind spezifische Effekte auf die Nutzung des Antwortformats, die von invertierten Items gegenüber positiv formulierten Items ausgehen. Hierzu wurde einerseits berichtet, dass Reaktionen auf invertierte Items zu interindividuellen Unterschieden im Antwortverhalten führen, die substanzielle Konsistenzen über Inhaltsbereiche und Stabilitäten über Zeitintervalle aufweisen sowie mit anderen Persönlichkeitseigenschaften korrelieren (DiStefano & Motl 2006, Horan et al. 2003). Somit haben invertierte Items offenbar das Potenzial, individuelle Antwortstile zu induzieren. Andererseits zeigte ein Vergleich der Schwellenabstände für positiv formulierte und invertierte Items in einem Mischverteilungs-Rasch-Modell neben den allgemeinen Antwortstilen zur Bevorzugung extremer versus mittlerer Antwortkategorien keinen spezifischen Effekt der Itempolung auf das Antwortverhalten der Subpopulationen (Hinz et al. 2003).

Ungeachtet dieser empirischen Befunde für einzelne Datensätze werden im Folgenden die theoretischen Implikationen diskutiert, die eine Anwendung polytomer Rasch-Modelle auf positiv formulierte und invertierte Items mit Rekodierung mit sich bringt.

2.2 Konsequenzen der Rekodierung invertierter Items für die Auswertung durch Rasch-Modelle

Das RSM enthält die nahe liegende Annahme, dass für Fragebogenitems mit identischer Anzahl und Art von Antwortkategorien eine konstante Nutzung des Antwortformats vorausgesetzt werden kann, die unabhängig vom spezifischen Inhalt und der Schwierigkeit der einzelnen Items ist. Was bedeutet diese Modellvorstellung jedoch beim Auftreten invertierter Items, und ist das RSM bei Gültigkeit der Annahme einer konstanten Nutzung des Antwortformats auch zur Analyse positiv formulierter und invertierter Items mit Rekodierung angemessen?

Im oberen Teil von Abbildung 3 sind die Kategorie-Charakteristik-Kurven eines positiv formulierten Items mit sechs Antwortkategorien 0,…,5 als durchgezogene Linien dargestellt. Zusätzlich sind die Verläufe der fünf Schwellenwahrscheinlichkeiten (gestrichelte Funktionskurven) sowie die Lokationen der Schwellenparameter (vertikale gestrichelte Linien) eingezeichnet. Wie aus der Abbildung hervorgeht, unterscheiden sich die Abstände zwischen den einzelnen Schwellen deutlich. So ist etwa das Intervall, in dem die Kategorie 1 die modale Antwortwahrscheinlichkeit aufweist, erheblich breiter als das entsprechende Intervall für Kategorie 4.

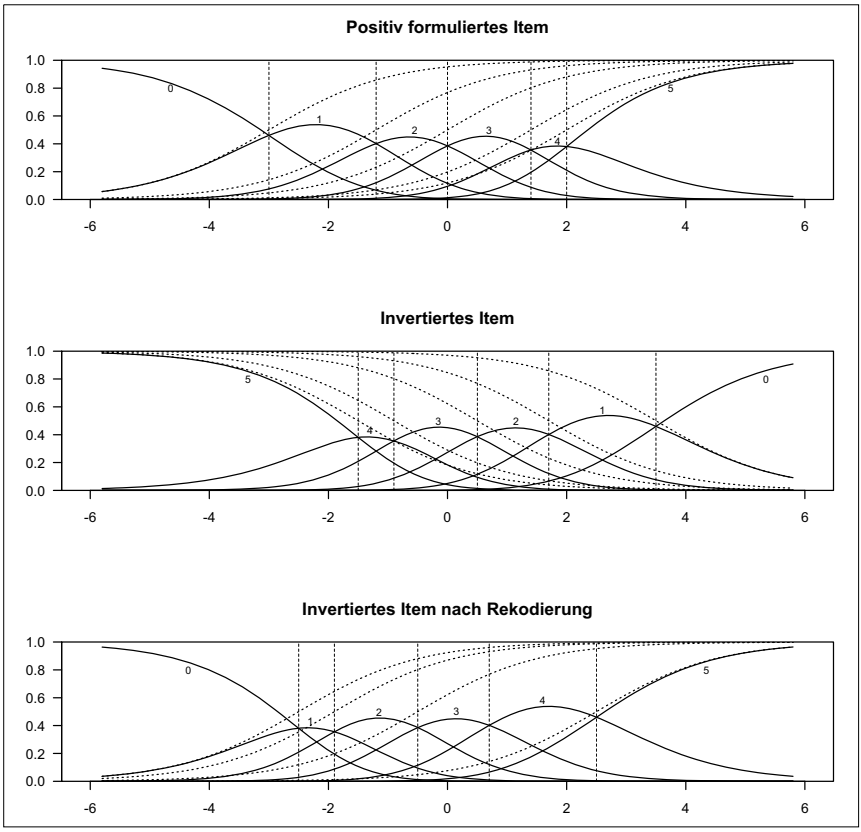


Abbildung 3: Kategorienwahrscheinlichkeiten (durchgezogene Funktionsverläufe) und Schwellen-Charakteristik-Kurven (gestrichelte Funktionsverläufe) für positiv formulierte, invertierte und rekodierte Items mit sechs Antwortkategorien. Vertikale gestrichelte Linien zeigen die Lokation der Schwellenparameter an.

Der mittlere Teil von Abbildung 3 zeigt die Kategorienwahrscheinlichkeiten für ein invertiertes Item vor der Rekodierung. Im Gegensatz zu positiv formulierten Items ist bei invertierten Items hier zu beachten, dass im Falle einer schwachen Personenausprägung des latenten Merkmals die Wahrscheinlichkeiten für hohe Antwortkategorien größer sind als die Wahrscheinlichkeiten für niedrige Antwortkategorien, während im Falle einer starken Personenausprägung die Wahrscheinlichkeiten für niedrige Antwortkategorien größer sind als für hohe. So sollte das Beispielitem "Ich bevorzuge es, die meisten Dinge allein zu tun" in einem Fragebogen zur Erfassung von Extraversion mit großer Wahrscheinlichkeit eine hohe Zustimmung bei solchen Personen auslösen, die eine schwache Ausprägung des Merkmals Extraversion aufweisen, während Personen mit einer starken Ausprägung von Extraversion zu geringer Zustimmung tendieren. Daher sind die Schwellen-Charakteristik-Kurven im mittleren Teil der Abbildung monoton fallend, d. h. die bedingte Wahr-

scheinlichkeit der jeweils höheren Antwortkategorie eines Paares benachbarter Kategorien sinkt mit steigendem Personenparameter. Entsprechend weist die höchste Antwortkategorie die modale Wahrscheinlichkeit bei sehr niedrigen Ausprägungen der latenten Eigenschaft auf, und mit steigendem Personenparameter nehmen die übrigen Antwortkategorien in abfallender Reihenfolge die modale Antwortwahrscheinlichkeit an. Dabei wird zusätzlich vorausgesetzt, dass die Antworten zu positiv formulierten und invertierten Items dieselbe latente Persönlichkeitsdimension widerspiegeln, jedoch mit umgekehrtem Vorzeichen.

Der obere und mittlere Teil von Abbildung 3 drückt die spezifische Annahme aus, dass die Schwellenabstände für beide Itemarten identisch sind. Beispielsweise ist der Abstand zwischen der Schwelle für die Kategorien 3 und 4 und der Schwelle für die Kategorien 4 und 5 des positiv formulierten Items (rechte Schwellen im oberen Teil von Abb. 3) genauso groß wie der Abstand zwischen der Schwelle für die Kategorien 5 und 4 und der Schwelle für die Kategorien 4 und 3 des invertierten Items (linke Schwellen im mittleren Teil von Abb. 3). Dadurch ist das Intervall, in dem Kategorie 4 die Antwortwahrscheinlichkeiten dominiert, in beiden Items gleich breit. Analoges gilt für die anderen Schwellenabstände und Antwortkategorien. Diese Invarianz der Schwellenabstände entspricht der Modellvorstellung, dass die Interpretation der Antwortkategorien und die Verwendung des Antwortformats nicht vom Inhalt der Items abhängen und sich damit auch nicht zwischen positiv formulierten und invertierten Items unterscheiden.

Im unteren Teil von Abbildung 3 sind die Konsequenzen dieser Modellvorstellung für ein invertiertes Item nach Rekodierung skizziert. Durch die Rekodierung werden die Antwortkategorien so umbenannt, dass eine stärkere Ausprägung des latenten Merkmals analog zu positiv formulierten Items die Wahrscheinlichkeit für eine hohe Antwortkategorie steigert. Daher sind die Schwellen-Charakteristik-Kurven für rekodierte Items wiederum monoton steigend, und die Kategorien nehmen in aufsteigender Reihenfolge die modale Antwortwahrscheinlichkeit an. Da die Rekodierung jedoch lediglich die Bezeichnung der Kategorien umdreht, so dass hohen Antwortkategorien niedrige Werte zugeordnet werden und umgekehrt, verändern sich die aus dem Antwortverhalten der Testpersonen resultierenden Schwellenabstände der ursprünglichen Antwortkategorien durch diese nachträgliche Umbenennung nicht. Somit entspricht der Abstand zwischen der Schwelle für die Kategorien 5 und 4 und der Schwelle für die Kategorien 4 und 3 vor der Rekodierung (linke Schwellen im mittleren Teil von Abb. 3) dem Abstand der Schwelle für die Kategorien 0 und 1 und der Schwelle für die Kategorien 1 und 2 nach der Rekodierung (linke Schwellen im unteren Teil von Abb. 3). Entsprechendes gilt für alle übrigen Schwellenabstände.

Wie auch ein direkter Vergleich zwischen dem oberen und dem unteren Teil von Abbildung 3 demonstriert, werden durch die Rekodierung invertierter Items also Schwellenabstände zwischen unterschiedlichen Kategorien "vertauscht". Während für positiv formulierte und für nicht-rekodierte invertierte Items beispielsweise das Intervall, in dem die Kategorie 1 die modale Antwortwahrscheinlichkeit aufweist, deutlich breiter ist als das entsprechende Intervall für Kategorie 4, gilt für reko-

dierte Items das Gegenteil. Obwohl den fiktiven Items in Abbildung 3 die Modellvorstellung zugrunde liegt, dass Testpersonen das Antwortformat unabhängig vom Iteminhalt in konstanter Weise interpretieren und nutzen, und obwohl damit die zentrale Annahme des RSM als gültig vorausgesetzt wird, führt das RSM nach der Rekodierung invertierter Items zu einer Fehlspezifikation, da die Schwellenabstände für positiv formulierte und invertierte Items mit Rekodierung keineswegs gleich sind. Eine derartige Problematik stellt sich für das PCM und das DM nicht, da diese Modellvarianten keine Annahmen über die Konstanz von Schwellenabständen zwischen den Items eines Fragebogens enthalten (vgl. Rost 2004, S. 223). Dafür verzichten diese Modelle jedoch auf eine Nutzung der wesentlichen Information, dass die Items identische Antwortalternativen beinhalten, und können daher als weniger theoretisch stringent und sparsam betrachtet werden. Im Folgenden werden Modifikationen und Erweiterungen des RSM betrachtet, die sich unter speziellen Voraussetzungen für die gemeinsame Analyse von positiv formulierten und invertierten Items mit Rekodierung eignen.

In Abbildung 3 wurde die Annahme getroffen, dass die Testpersonen das Antwortformat für positiv formulierte und invertierte Items in identischer Weise interpretieren und nutzen. Statt des ursprünglichen RSM, das unter dieser Voraussetzung zu einer Fehlspezifikation für rekodierte Items führt, ist in diesem Fall eine Parametrisierung der Kategorienparameter angezeigt, durch die die Reihenfolge der Schwellenabstände von invertierten Items mit Rekodierung gegenüber positiv formulierten Items umgekehrt wird. Dies kann durch eine Spezifikation der Schwellenparameter als

$$\sigma_{ix} = \beta_i + \tau_x \qquad \text{für positiv formulierte Items und} \tag{9}$$

$$\sigma_{ix} = \beta_i - \tau_{(m+1-x)} \qquad \text{für rekodierte invertierte Items}$$

gewährleistet werden. Durch die Parametrisierung der Schwellenparameter in diesem modifizierten Rating-Scale-Modell (mRSM) bleibt die Anzahl der Kategorienparameter im Vergleich zu dem ursprünglichen RSM unverändert. Das mRSM berücksichtigt jedoch die Vertauschung der Schwellenabstände infolge der Rekodierung (vgl. Abb. 3).

Alternativ zu der restriktiven Annahme, dass das Antwortformat unabhängig von der Itempolung genutzt wird, können die Testpersonen dessen Nutzung in Reaktion auf invertierte Items gegenüber positiv formulierten Items auch ändern. Falls die Testpersonen das Antwortformat konstant über alle positiv formulierten Items einerseits und konstant über alle invertierten Items andererseits nutzen, jedoch einen Unterschied in der Nutzung des Antwortformats zwischen positiv formulierten und invertierten Items machen, dann ist ein erweitertes Rating-Scale-Modell (eRSM) angemessen, in dem getrennte Kategorienparameter für positiv formulierte und invertierte Items spezifiziert werden. Die Schwellenparameter ergeben sich demnach als

$$\sigma_{ix} = \beta_i + \tau_x^{(p)} \quad \text{für positiv formulierte Items und} \tag{10}$$

$$\sigma_{ix} = \beta_i + \tau_x^{(r)} \quad \text{für rekodierte invertierte Items.}$$

Damit enthält das eRSM doppelt so viele Kategorienparameter wie das ursprüngliche RSM bzw. das mRSM. Gleichzeitig umfasst das eRSM sowohl das RSM als auch das mRSM als Spezialfälle, da sich die Schwellenparameter der Gleichungen (6) und (9) durch geeignete Restriktionen in (10) ergeben (vgl. Abb. 2).

Als Extremfall des eRSM für positiv formulierte und rekodierte inverse Items ist ein Antwortprozess denkbar, bei dem die Testpersonen das Antwortformat zwischen den beiden Itemtypen in entgegengesetzter Weise verwenden. Das bedeutet, dass die Testpersonen hohe Antwortkategorien für invertierte Items spontan wie niedrige Antwortkategorien für positiv formulierte Items interpretieren und umgekehrt. Als Konsequenz würden die Schwellenabstände für invertierte Items vor der Rekodierung einer Spiegelung der Schwellenabstände für positiv formulierte Items entsprechen. Bezogen auf Abbildung 3 würde dies etwa beinhalten, dass das breite Intervall, in dem Kategorie 1 im Falle positiv formulierter Items dominant ist, zu einem ebenso breiten Intervall für Kategorie 4 im Falle invertierter Items vor der Rekodierung führte, während das schmale Intervall für Kategorie 4 bei positiv formulierten Items zu einem ebenso schmalen Intervall für Kategorie 1 bei invertierten Items vor der Rekodierung führte. Die nachfolgende Rekodierung hätte bei diesem Antwortprozess zur Folge, dass die Schwellenabstände rekodierter invertierter Items identisch zu den Schwellenabständen positiv formulierter Items sind, so dass das ursprüngliche RSM für die gemeinsame Analyse positiv formulierter und rekodierter invertierter Items angemessen wäre.

Anders ausgedrückt erfordert die Anwendung des urspünglichen RSM auf Itemmengen, die sowohl positiv formulierte als auch invertierte Items mit Rekodierung umfassen, die starke Annahme, dass die Testpersonen das Antwortformat für positiv formulierte und invertierte Items in exakt entgegengesetzter Weise verwenden. Diese Annahme kann durch einen Modellvergleich zwischen dem RSM und dem hierarchisch übergeordneten eRSM spezifisch geprüft werden.

3 Ein symmetrisches Rating-Scale-Modell

Im vorangehenden Abschnitt wurden Rasch-Modelle für ordinale Items diskutiert, die in Abhängigkeit von den spezifischen Annahmen hinsichtlich möglicher Unterschiede in der Nutzung des Antwortformats für positiv formulierte versus invertierte Items angemessen sind oder zu Fehlspezifikationen führen. Das eRSM, das separate Kategorienparameter für die beiden Itemtypen vorsieht, ist sowohl dem mRSM mit gespiegelten Kategorienparametern für rekodierte Items als auch dem ursprüngliche RSM hierarchisch übergeordnet. Durch diese Modellhierarchie werden die speziellen Annahmen des mRSM bzw. des RSM empirisch prüfbar. Gleich-

zeitig enthält das eRSM jedoch mehr Parameter als die beiden anderen und ist somit weniger sparsam.

Daher soll hier ein alternativer Weg beschritten werden, indem statt einer Modellerweiterung durch zusätzliche Parameter ein sparsameres Ausgangsmodell formuliert wird. Dieses sparsame Ausgangsmodell soll so beschaffen sein, dass auch eine Vertauschung der Schwellenabstände invertierter Items durch Rekodierung (vgl. Abb. 3) zu äquivalenten Kategorienparametern für positiv formulierte und rekodierte invertierte Items führt. Damit soll sich das Ausgangsmodell ohne weitere Modifikationen für die gemeinsame Analyse beider Itemarten eignen, sofern die Testpersonen das Antwortformat entweder in identischer oder in entgegengesetzter Weise für positiv formulierte und invertierte Items verwenden. Ein Modell, das diese Kriterien erfüllt, kann durch die Spezifikation symmetrischer Schwellenparameter

$$\sigma_{ix} = \beta_i + \tau_x \quad \text{mit} \quad \tau_x = -\tau_{(m+1-x)} \tag{11}$$

formuliert werden. Bei dem resultierenden symmetrischen Rating-Scale-Modell (sRSM) handelt es sich um eine restringierte Version des RSM bzw. des mRSM, in der postuliert wird, dass sich die Kategorienparameter und Schwellenabstände innerhalb der Items symmetrisch zum Mittelpunkt der Ratingskala verhalten. Abbildung 4 illustriert das sRSM anhand eines fiktiven Items mit sechs Antwortkategorien.

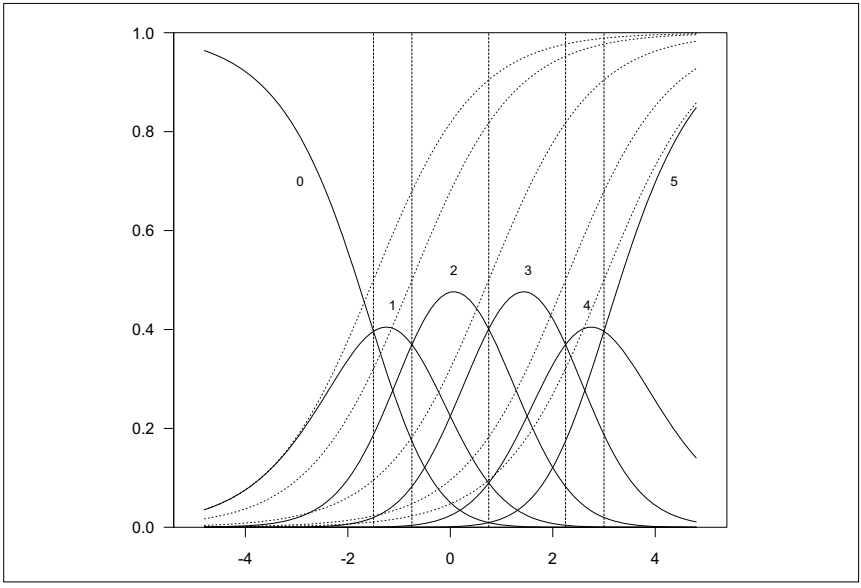


Abbildung 4: Kategorienwahrscheinlichkeiten (durchgezogene Funktionsverläufe), Schwellen-Charakteristik-Kurven (gestrichelte Funktionsverläufe) und Lokation der Schwellenparameter (vertikale gestrichelte Linien) eines fiktiven Items nach dem symmetrischen Rating Scale-Modell

Im Gegensatz zum mRSM wird die Annahme gespiegelter Schwellenabstände beim sRSM nicht zwischen positiv formulierten und rekodierten invertierten Items eingeführt, sondern die Symmetrieannahme richtet sich hier auf die Schwellenabstände innerhalb der Items jedes Itemtyps. Wie in Abbildung 2 dargestellt, kann das sRSM gleichzeitig sowohl als Spezialfall des mRSM als auch des ursprünglichen RSM betrachtet werden. Dies hat zur Folge, dass es im Falle der Gültigkeit des sRSM gleichgültig ist, ob die Testpersonen das Antwortformat für positiv formulierte und invertierte Items identisch oder entgegengesetzt verwenden, da eine Umkehrung der symmetrischen Schwellenabstände ohne Auswirkung bleibt. Damit bietet das sRSM eine Ausgangsbasis für die gemeinsame Analyse positiv formulierter und rekodierter Fragebogenitems, die unter verschiedenen Annahmen im Hinblick auf den zugrundeliegenden Antwortprozess sinnvoll ist. Im Gegensatz zum PCM trägt das sRSM durch die itemunabhängigen Kategorienparameter der Tatsache Rechnung, dass die Items mit einem einheitlichen Antwortformat vorgegeben werden. Im Unterschied zum DM wird nicht gefordert, dass die Schwellen innerhalb der Items äquidistant sind, und es werden keine itemspezifischen Dispersionen vorgesehen.

Das sRSM erscheint nicht nur aufgrund seiner formalen Eignung für die simultane Analyse positiv formulierter und rekodierter invertierter Items attraktiv. Mehrere Gründe sprechen auch für die theoretische und empirische Plausibilität des Modells zur Analyse von Fragebogenitems mit konstantem Antwortformat. So ist es theoretisch naheliegend, dass Testpersonen die Entfernung der Antwortkategorien vom Mittelpunkt einer Ratingskala in beide Richtungen als äquivalent wahrnehmen, während die Abstufungen der Kategorien innerhalb jeder Richtung keineswegs als gleichwertig interpretiert werden müssen. Für das Beispielitem in Abbildung 4 bedeutet dies etwa, dass die Antwortkategorie 1 als gleich weit entfernt von einem imaginären Skalenmittelpunkt der Gleichgültigkeit oder Unentschiedenheit wahrgenommen wird wie die Antwortkategorie 4, während die Entfernung des Übergangs von Kategorie 1 zu Kategorie 2 (bzw. von Kategorie 3 zu Kategorie 4) vom mittleren Übergang zwischen den Kategorien 2 und 3 als deutlich gravierender empfunden wird als von dem äußeren Übergang von Kategorie 0 zu Kategorie 1 (bzw. von Kategorie 4 zu Kategorie 5). Eine derartige Interpretation und Nutzung des Antwortformats führt zu Schwellenabständen, die zum Mittelpunkt der Ratingskala symmetrisch, jedoch keineswegs äquidistant sind, was exakt dem sRSM entspricht. Zusätzlich korrespondiert die für das Modell zentrale Annahme punktsymmetrischer Kategorienparameter mit vergleichbaren Konzepten zur Analyse ordinaler Zustimmungsitems anhand von probabilistischen Unfolding-Modellen (vgl. etwa Roberts et al. 2010, S. 178).

Deskriptiv scheint das sRSM auch eine gute empirische Übereinstimmung mit den Ergebnissen einiger früherer Studien aufzuweisen. So ergaben sich relativ symmetrisch wirkende Schwellenparameter in Analysen unterschiedlicher Fragebogen mit ordinalen Ratingitems durch das PCM (Gollwitzer et al. 2005, Hinz et al. 2003) und durch das RSM (Meiser & Machunsky 2008). Da in diesen vorangegangenen Analysen jedoch keine expliziten Symmetrieannahmen bezüglich der Schwellen-

bzw. Kategorienparameter implementiert und getestet wurden, steht eine direkte empirische Bewertung des sRSM bislang aus.

4 Anwendung des symmetrischen Rating-Scale-Modells auf Items zum Konstrukt Personal Need for Structure

Zur Illustration der in den Abschnitten 2 und 3 vorgestellten polytomen Rasch-Modelle für positiv formulierte und rekodierte invertierte Items wurden Daten von n = 1789 Testpersonen aus der Studie von Meiser & Machunsky (2008) reanalysiert, die die deutschsprachige Version des Fragebogens zur Erfassung von Personal Need for Structure (Machunsky & Meiser 2006) bearbeitet haben. Das Konstrukt Personal Need for Structure beinhaltet das Aufsuchen von Struktur im Sinne von Ordnung und Vorhersagbarkeit sowie negative Reaktionen und Vermeidungstendenzen in Bezug auf unstrukturierte und unvorhersehbare Situationen (Neuberg & Newsom 1993, Thompson et al. 2001). Der Fragebogen zur Erfassung des Konstrukts umfasst 12 Items mit einem einheitlichen sechsstufigen Antwortformat mit den Endpunkten "trifft gar nicht zu" und "trifft vollkommen zu".

Von den insgesamt 12 Items beziehen sich vier auf das Aufsuchen von Struktur und acht auf negative Reaktionen auf unstrukturierte Situationen. Für die Reanalyse wurden sechs Items ausgewählt, die Reaktionen auf unstrukturierte Situationen zum Gegenstand haben. Die ersten beiden Items dieses Inhaltsbereiches wurden ausgeschlossen, um zu gewährleisten, dass die Testpersonen bis zur Bearbeitung der analysierten Items eine stabile Nutzung des Antwortformats entwickeln konnten. Die verbleibenden sechs Analyseitems lauteten:

1. Ich genieße es, spontan zu sein. (invertiert)
2. Ich mag unklare Situationen nicht.
3. Ich hasse es, meine Pläne in der letzten Minute zu ändern.
4. Ich bin ungern mit Leuten zusammen, deren Verhalten nicht vorhersehbar ist.
5. Ich genieße die Herausforderung, mich in unvorhersehbaren Situationen zu befinden. (invertiert)
6. Ich fühle mich unwohl, wenn die Regeln in einer Situation unklar sind.

Die beiden invertierten Items wurden vor der statistischen Analyse rekodiert.

Auf die ausgewählten Analyseitems wurden mehrere Rasch-Modelle mit unterschiedlichen Restriktionen hinsichtlich der Schwellenabstände für positiv formulierte und invertierte Items mit Rekodierung angewendet. Die Modellparameter wurden anhand der Conditional-Maximum-Likelihood-Methode mit dem R-Paket eRm (Mair et al. 2010; vgl. auch Mair & Hatzinger 2007) geschätzt, wobei die unterschiedlichen Modelle als Spezialfälle des linearen Partial-Credit-Modells spezifiziert wurden. Für die Modellvergleiche wurde das Bayesian Information Criterion (BIC; Schwarz 1978; vgl. auch Burnham & Anderson 2002) als deskriptives Maß für die Modellpassung in Relation zur Anzahl der Modellparameter herangezogen. Auf inferenzstatistische Vergleiche durch bedingte Log-Likelihood-Quotiententests

wurde verzichtet, da nicht alle der hier verwendeten Modelle in einem hierarchischen Verhältnis zueinander stehen (vgl. Abb. 2).

Modell	ln(L)	n(P)	BIC
RSM	-11494.44	9	23056.28
mRSM	-11495.71	9	23058.82
sRSM	-11498.12	7	23048.67
eRSM	-11464.63	13	23026.62
PCM	-11377.95	29	22973.09

Tabelle 1: Goodness-of-Fit Statistiken für unterschiedliche Modellvarianten. ln(L) = Log-Likelihood, n(P) = Anzahl der Parameter, RSM = Rating-Scale-Modell, mRSM = modifiziertes Rating-Scale-Modell, sRSM = symmetrisches Rating-Scale-Modell, eRSM = erweitertes Rating-Scale-Modell, PCM = Partial-Credit-Modell

Modell	Kategorieparameter τ				
	Schwelle 1	Schwelle 2	Schwelle 3	Schwelle 4	Schwelle 5
RSM	-1.584	-0.415	-0.071	0.520	1.550
mRSM[a]	-1.631	-0.427	0.027	0.498	1.534
mRSM[b]	-1.534	-0.498	-0.027	0.427	1.631
sRSM	-1.569	-0.466	0.000	0.466	1.569

[a] Kategorieparameter für positiv formulierte Items
[b] Kategorieparameter für rekodierte invertierte Items

Tabelle 2: Geschätzte Kategorienparameter.[1] RSM = Rating-Scale-Modell, mRSM = modifiziertes Rating-Scale-Modell, sRSM = symmetrisches Rating-Scale-Modell

Zunächst wurden die Daten mit dem RSM, dem mRSM sowie dem sRSM ausgewertet. Die ermittelte Log-Likelihood, die Anzahl der Modellparameter sowie die resultierenden BIC-Werte dieser und einiger weiterer Modelle sind in Tabelle 1 dargestellt. Tabelle 2 enthält die geschätzten Kategorienparameter des RSM, mRSM und sRSM. Anhand der BIC-Werte weisen das RSM und das mRSM eine nahezu identische Modellgüte auf, wobei das RSM mit einem geringfügig kleineren Wert eine etwas bessere Passung zeigt. Die geschätzten Kategorienparameter für die fünf Schwellen des sechsstufigen Antwortformats zeigen in beiden Modellen, dass die Abstände zwischen den Schwellen für die extremeren Antwortkategorien größer sind als die Abstände zwischen den Schwellen für die mittleren Antwortkategorien. Da dies gleichermaßen für die Schwellenabstände zwischen niedrigen und hohen Antwortkategorien gilt, lassen die geschätzten Kategorienparameter deskriptiv ein symmetrisches Muster erkennen. Für das mRSM ist weiterhin zu beachten, dass die Kategorienparameter für rekodierte invertierte Items gegenüber

1 Parameterschätzung anhand der Conditional Maximum Likelihood-Methode mit dem R-Paket eRm (Mair et al. 2010).

den positiv formulierten Items gespiegelt sind, wie in Tabelle 2 illustriert wird. Die symmetrische Anordnung der Kategorienparameter und Schwellenabstände wird durch den niedrigeren BIC-Wert des sRSM sowohl im Vergleich zum RSM als auch zum mRSM bestätigt. Der reduzierte BIC-Wert spricht dafür, dass die größere Sparsamkeit des sRSM nicht mit einem gravierenden Verlust an Modellpassung einhergeht. Damit hat sich das sRSM für die hier reanalysierten Daten gegenüber den anderen beiden Modellen empirisch bewährt. In Übereinstimmung mit den zuvor diskutierten theoretischen Gründen, die für das sRSM zur gemeinsamen Analyse positiv formulierter und rekodierter invertierter Items sprechen, kann dieses Modell somit als sparsame und geeignete Modellvariante angesehen werden.

Die in Tabelle 2 aufgeführten Modelle enthalten die Annahme, dass die Kategorienparameter und Schwellenabstände homogen über die analysierten Items sind, auch wenn im mRSM die Reihenfolge der Schwellenabstände für rekodierte invertierte Items umgekehrt wird. Diese Annahme wurde durch den Vergleich mit alternativen Modellen geprüft. Das eRSM, das separate Kategorienparameter für positiv formulierte und rekodierte invertierte Items zulässt, ist den zuvor diskutierten Modellen mit homogenen Kategorienparametern aufgrund eines niedrigeren BIC-Wertes überlegen, wie Tabelle 1 zeigt. Dies scheint zunächst für eine unterschiedliche Nutzung des Antwortformats zwischen den beiden Itemtypen zu sprechen, d. h. für unterschiedliche Schwellenabstände bei positiv formulierten Items gegenüber invertierten Items mit Rekodierung. Allerdings wird diese Interpretation durch weitergehende Analysen in Frage gestellt, in denen andere Aufteilungen der Items vorgenommen wurden als die nach Itempolung. So führt etwa eine Spezifikation separater Kategorienparameter für die ersten drei und die letzten drei Items zu einem BIC-Wert von 23052.30, der die Werte für das RSM und das mRSM ebenfalls unterschreitet, obwohl sich in beiden Teilmengen jeweils zwei positiv formulierte und ein rekodiertes invertiertes Item befinden. Werden für die Items 3 und 4 separate Kategorienparameter festgelegt, die sich von den gemeinsamen Kategorienparametern für die Items 1, 2, 5 und 6 unterscheiden können, so resultiert ein BIC-Wert von nur 22898.60. Dieser Wert ist sogar niedriger als der Wert des eRSM, obwohl die beiden Items 3 und 4 positiv formuliert sind und das alternative Modell somit Besonderheiten in der Nutzung des Antwortformats für zwei positiv formulierte Items gegenüber den anderen beiden positiv formulierten und den zwei rekodierten invertierten Items beinhaltet.

Diese Ergebnisse sprechen dafür, dass die Heterogenität der Schwellenparameter keineswegs auf spezifische Unterschiede des Antwortverhaltens in Abhängigkeit von der Itempolung zurückgeht, sondern vielmehr Variationen der Modellparameter einzelner Items unabhängig von deren Polung reflektiert. Diese Interpretation wird durch einen abschließenden Vergleich mit dem PCM unterstützt, das als Extremfall heterogener Kategorienparameter und Schwellenabstände angesehen werden kann, da es separate Kategorienparameter für jedes Item vorsieht und damit auf Homogenitätsannahmen in Teilmengen der Items verzichtet. Wie in Tabelle 1 zu sehen ist, zeigt das PCM einen niedrigeren BIC-Wert als das eRSM und seine Spezialfälle RSM, mRSM und sRSM und lässt damit die Annahme homogener

Kategorienparameter innerhalb der Teilmengen von Items mit gleicher Polung fragwürdig erscheinen. Gleichzeitig ist der BIC-Wert des PCM höher als derjenige des zuvor analysierten Modells mit gemeinsamen Kategorieparametern für die beiden positiv formulierten Items 3 und 4 einerseits und für die gemischt positiv formulierten und invertierten Items 1, 2, 5 und 6 andererseits. Damit scheint die Annahme homogener Kategorienparameter zumindest für die gemischte Teilmenge der Items 1, 2, 5 und 6 gerechtfertigt.

Zusammengefasst sprechen die Ergebnisse dafür, dass die Symmetrieannahme des sRSM im Vergleich zum übergeordneten RSM für die hier analysierten Daten empirisch haltbar ist und dass Unterschiede der Schwellenabstände zwischen Items keine spezifischen Effekte der Itempolung reflektieren.

5 Schlussfolgerungen und Ausblick

Rasch-Modelle für polytome Items mit ordinalem Antwortformat erlauben eine präzise Spezifikation und Prüfung von Annahmen über die latenten Prozesse, die zu den beobachteten Wahlen von Antwortkategorien führen. Damit unterscheiden sich Rasch-Modelle wesentlich von der klassischen Testtheorie, die die einzelnen bzw. aggregierten Itemantworten naiv als Messwerte interpretiert und keine tiefergehende Analyse des Antwortverhaltens vorsieht. Die formalen Implikationen unterschiedlicher polytomer Rasch-Modelle wurden hier in Bezug auf die Verwendung invertierter Items mit Rekodierung diskutiert. Dabei zeigte sich, dass das übliche RSM nur unter der äußerst restriktiven Annahme für die gemeinsame Analyse positiv formulierter und rekodierter invertierter Items geeignet ist, dass die Testpersonen das Antwortformat für invertierte Items gegenüber positiv formulierten Items entgegengesetzt verwenden, d. h. Schwellenabstände für hohe und niedrige Antwortkategorien spontan vertauschen. Benutzen die Testpersonen das Antwortformat hingegen in der gleichen Weise für beide Itemarten, so kann das übliche RSM zu gravierenden Fehlspezifikationen führen, während das mRSM mit gespiegelten Schwellenabständen für rekodierte invertierte Items dem Antwortverhalten gerecht wird. Das ebenfalls diskutierte sRSM erlaubt im Falle seiner Gültigkeit eine problemlose Kombination positiv formulierter und invertierter Items, da die Symmetrie der Kategorienparameter äquivalente Schwellenabstände für beide Itemtypen garantiert — und zwar unabhängig von der Beibehaltung oder Vertauschung der Schwellenabstände für invertierte Items durch die Testpersonen. Diese Eigenschaft macht das sRSM zu einem theoretisch und pragmatisch attraktiven Ausgangsmodell zur Analyse von Fragebogenitems, die positiv formulierte und invertierte Items mit konstantem Antwortformat umfassen.

Mit der Verwendung invertierter Items wird im Allgemeinen das Ziel verfolgt, den Einfluss genereller Antworttendenzen auf den Summenwert eines Fragebogens durch Aggregation über positiv formulierte und rekodierte invertierte Items zu reduzieren. Die genauere testtheoretische Betrachtung mit Rasch-Modellen für ordinale Antwortkategorien hat dabei jedoch mögliche Probleme hinsichtlich der im-

pliziten oder expliziten Annahme aufgezeigt, dass die Antwortkategorien über die unterschiedlichen Items hinweg gleich interpretiert und verwendet werden. Denn genau diese Annahme führt durch die nachträgliche Rekodierung invertierter Items zur Ungleichheit von Schwellenabständen für Kategorien, die numerisch den gleichen Wert haben (Rost 2004, S. 223). Vorliegende Studien deuten jedoch darauf hin, dass diese theoretische Problematik empirisch nicht allzu gravierend ist. So zeigten sich in einer Analyse von Fragebogendaten zu den Big-Five-Persönlichkeitsfaktoren neben generellen Antwortstilen zur Präferenz moderater versus extremer Antwortkategorien keine systematischen Effekte der Invertierung und Rekodierung von Items auf die Schwellenabstände im PCM (Hinz et al. 2003). Ähnlich ergab eine umfassende Reanalyse zahlreicher Datensätze aus der Persönlichkeitsdiagnostik anhand einer Serie von Rasch-Modellen für ordinale Items durch Hoeffler (2010), dass eine Heterogenität der Schwellenabstände über die Items besteht, die sich nicht auf spezielle Unterschiede zwischen positiv formulierten und invertierten Items mit Rekodierung zurückführen lässt. Die in diesem Beitrag vorgestellte Reanalyse von Daten aus der Studie von Meiser & Machunsky (2008) bestätigt diese Einschätzung.

In der Übersicht von Hoeffler (2010) erwies sich das PCM dem üblichen RSM sowie den in Abschnitt 2.2 skizzierten modifizierten und erweiterten Modellvarianten weitgehend überlegen. Gleichzeitig war das eRSM dem RSM und dem mRSM überlegen, was zunächst für eine prinzipiell unterschiedliche Interpretation und Nutzung des Antwortformats bei positiv formulierten versus rekodierten invertierten Items spricht. Um den spezifischen Effekt invertierter Items mit Rekodierung auf die Schwellenabstände zu prüfen, führte Hoeffler Zusatzanalysen durch, in denen analog zum eRSM getrennte Kategorienparameter für Teilmengen von Items spezifiziert wurden. Diese Teilmengen wurden jedoch zufällig gebildet und nicht anhand der Itempolung definiert. Die Ergebnisse zeigten, dass das Modell mit getrennten Kategorienparametern für Teilmengen von Items auch bei zufälliger Aufteilung der Items in der Mehrheit der Datensätze eine empirische Überlegenheit gegenüber dem üblichen RSM aufweist. Somit erlaubt die Untersuchung von Hoeffler die vorläufige Schlussfolgerung, dass Unterschiede der Schwellenabstände für positiv formulierte und rekodierte invertierte Items eher ein Bestandteil der allgemeinen Heterogenität in der Nutzung des Antwortformats zwischen Items darstellt und weniger als spezifische Effekte der Itempolung angesehen werden sollten.

Auch wenn diese empirischen Befunde darauf hindeuten, dass testtheoretische Probleme aufgrund der Analyse invertierter Items mit Rekodierung in vielen Fällen praktisch nicht gravierend sind, ist es eine Aufgabe der testtheoretischen Modellbildung, die theoretisch angenommenen Antwortprozesse formal abzubilden und dadurch prüfbar zu machen. Die hier vorgestellten Varianten des Rasch-Modells für polytome Items zur Berücksichtigung invertierter Items können hierzu einen Beitrag leisten. Eine äußerst flexible Möglichkeit zur Spezifikation und Prüfung unterschiedlicher Annahmen über die Nutzung des Antwortformats erwächst darüber hinaus aus aktuellen Ansätzen zur Formulierung von Modellen der Item-Response-Theorie als nicht-lineare Mehrebenenmodelle (Raudenbusch & Bryk 2002, Rijmen

et al. 2003). Neben der Formulierung von Modellannahmen über die Nutzung des Antwortformats in Abhängigkeit von Itemeigenschaften wie der Itempolung erlaubt dieser formale Rahmen zusätzlich die Berücksichtigung von Kovariaten der Testpersonen im Hinblick auf die Personenausprägung des zu erfassenden Konstrukts sowie der Interpretation und Nutzung des Antwortformats. Somit können etwa Einflüsse von Personenvariablen auf die Auswirkungen der Itempolung (vgl. DiStefano & Motl 2006) direkt in ein theoretisch motiviertes Rasch-Modell integriert werden.

Latent-Class- und Mixed-Rasch-Modelle zur Identifizierung skalierbarer und unskalierbarer Personengruppen in der Allgemeinen Depressionsskala

Ferdinand Keller

1 Einführung

1.1 Die Allgemeine Depressionsskala und ihre faktorielle Struktur

Die Allgemeine Depressionsskala (ADS) ist die deutsche Version der Center of Epidemiological Studies Depression Scale (CES-D) von Radloff (1977), die speziell für die Untersuchung großer Stichproben in epidemiologischen Studien entwickelt wurde. Die Skala soll ökonomisch und zuverlässig depressive Symptome erfassen und so möglicherweise depressive Personen erkennen helfen. Mittlerweile ist die CES-D sehr weit verbreitet, sie wurde in mehr als 20 Sprachen übersetzt und wird in zahlreichen epidemiologischen wie auch klinischen Studien zur Abschätzung von Depressivität eingesetzt (Hautzinger et al. 2012).

Erste psychometrische Analysen von Radloff (1977) erbrachten eine gute interne Konsistenz der aus 20 Items bestehenden Skala (zur genaueren Beschreibung vgl. Kap. 2.2). In faktorenanalytischen Auswertungen konnten vier Faktoren unterschieden werden, die sie mit Depressivität, Wohlbefinden, somatische Symptome und interpersonelle Probleme bezeichnete. Diese vier-faktorielle Struktur wurde seither vielfach bestätigt. Shafer (2006) kommt in einer Metaanalyse von 28 Studien, in denen die faktorielle Struktur untersucht wird, zu dem Ergebnis, dass eine Lösung mit vier Faktoren am angemessensten ist und es im Vergleich zu den anderen untersuchten Instrumenten zur Messung von Depressivität wenig Variabilität über die Studien hinweg gibt. In konfirmatorischen Faktorenanalysen wurde weitgehend konsistent die bessere Anpassung einer vier-faktoriellen Struktur gefunden (vgl. Hautzinger et al. 2012), und dieses Ergebnis ließ sich auch für Jugendliche bestätigen (Barkmann et al. 2008, Crockett et al. 2005, Keller 2010). Es wurden jedoch auch andere Lösungen beschrieben, insbesondere eine Unterteilung in zwei Faktoren, bei der sich die positiv und die negativ formulierten Items aufteilten. Auch eine eindimensionale Lösung wurde vorgeschlagen, und in der Praxis hat sich die Angabe des Gesamtwertes der ADS mit der Argumentation durchgesetzt, dass alle Items konzeptuell mit Depression verbunden sind und deshalb unabhängig von den Faktorladungen eingeschlossen werden sollten (Edwards

et al. 2010, Hautzinger et al. 2012). Generell besteht aber weitgehende Einigkeit darüber, dass die positiv formulierten Items zusätzliche Varianz "erzeugen"; sei es inhaltlich, indem sie positiven Affekt ansprechen, oder sei es als Methodenfaktor.

1.2 Effekte negativer und positiver Itemformulierungen

Der Faktor Wohlbefinden wird durch vier positiv formulierte Items abgefragt, die in regelmäßigen Abständen (Itemnummern 4, 8, 12, 16) eingestreut sind und entsprechend umgepolt werden, bevor sie in den Summenwert eingehen. Diese Items wurden eingefügt, um Antwortmuster, Verzerrungen und unaufmerksames Beantworten bzw. unlautere Angaben zu erkennen bzw. zu verhindern. "Four items were worded in the positive direction to break tendencies toward response set as well as to assess positive affect (or its absence)" (Radloff 1977, S. 386). Hautzinger & Bailer (1993) haben deshalb ein so genanntes Lügenkriterium vorgeschlagen, das insbesondere bei anonymen Befragungen und der Untersuchung großer Gruppen, wenn die persönliche Interaktion gering ist, helfen soll, fraglich valide Antworten herauszufiltern. Ausgangspunkt war die Beobachtung, dass in großen Stichproben Personen immer wieder dadurch auffallen, dass sie alle Fragen mit 0 (selten oder überhaupt nicht) beantworten. Dieses stereotype Ankreuzen führt dazu, dass die vier positiv formulierten Items nach der Umkodierung mit 3 in die Auswertung eingehen, obgleich alle 16 anderen Items 0 Punkte erbringen und die 12 Punkte aus den positiv formulierten Items überhaupt nicht ins Bild passen (Hautzinger & Bailer 1993). Sie empfehlen daher zur Artefaktkontrolle, Fragebögen mit einem Missverhältnis von negativen zu positiven Summenwerten aus der Analyse herauszunehmen (zur Berechnung vgl. Kap. 2.2). Personen, die immer dieselbe Kategorie oder offensichtlich Muster ankreuzen, können nach visueller Kontrolle ebenfalls eliminiert werden.

Polungsartefakte sind aber auch dann zu erwarten, wenn eine bestimmte Personengruppe die unterschiedliche Polung nicht richtig erkennen kann. Matschinger et al. (2000) haben mit Hilfe probabilistischer Testmodelle in einer epidemiologischen Studie zu demenziellen Erkrankungen bei einer Stichprobe älterer Menschen deutliche Hinweise darauf gefunden, dass eine Gruppe die wechselnde Polung übersehen hat. Bei großen Schuluntersuchungen besteht das Problem, dass Jugendliche teilweise weniger motiviert sind, die Fragebögen auszufüllen, als dies bei Studenten oder erwachsenen Probanden erwartet werden kann. Bei Jugendlichen besteht daher nicht so sehr das Problem verringerter kognitiver Fähigkeiten als vielmehr die Gefahr, dass sie die Fragebögen nachlässig oder desinteressiert ausfüllen, wie es von den Durchführenden hin und wieder vermutet wird. Daher besteht in solchen Studien ein großes Interesse daran, die nicht instruktionsgemäß ausgefüllten Fragebögen herauszuziehen, so dass sich z. B. in Interventionsstudien die Interventionseffekte insgesamt besser und valider einschätzen lassen und eine Erhöhung der Teststärke für die Signifikanzbeurteilung erzielt werden kann.

Neben solchen, eher praktischen und in der Durchführung begründeten Problemen wurden aber immer wieder auch theoretische Einwände vorgebracht, nach denen sich der positive Faktor als eigenständig erweisen sollte. Theoretischer Hintergrund dessen ist die Diskussion um die Unkorreliertheit von positivem und negativem Affekt, wie sie von Watson & Tellegen (1985) propagiert wurde, bzw., ob Affektdimensionen uni- oder bipolar zu konzeptualisieren sind (Russell & Carroll 1999). Stansbury et al. (2006) vermuten bei ihrer Stichprobe von älteren Menschen, dass die positiven Items ein eigenes Konstrukt darstellen, das gering mit den anderen Subskalen korreliert. Unaufmerksamkeit beim Ausfüllen könne am Zustandekommen dieser geringen Korrelation beteiligt sein, aber nur als ergänzende Erklärung. Kulturelle Unterschiede wurden ebenfalls als mögliche Ursachen vermutet (Iwata et al. 1998), aber auch diese Autoren verweisen auf die Theorie von Watson & Tellegen.

In der teststatistischen Literatur wurde wiederholt versucht, Personen mit anscheinend inkonsistenten Antwortmustern aufgrund von speziell konstruierten Abweichungsindizes zu identifizieren (z. B. Zickar & Drasgow 1996). Auf diese Ansätze wird hier nicht weiter eingegangen, weil sie häufig davon ausgehen, dass diese Personen bewusst etwas vortäuschen und sogar betrügen wollen.

In der deutschen Version (ADS) besteht zusätzlich das Problem, dass das Item 8 eine schlechte Trennschärfe besitzt, insbesondere bei Jugendlichen, und auch Item 4 diesbezüglich abfällt, so dass aus messtheoretischer Perspektive weitere Probleme entstehen und Fragen zur Dimensionalität des Instruments mit Effekten der Itemformulierung konfundiert sein können.

1.3 Ergebnisse einer eigenen Studie mit Mixed-Rasch-Ansätzen

Unser erster Ansatz (Keller et al. 2008) ging daher der Frage nach, ob die z. T. schlechteren psychometrischen Eigenschaften auch (teilweise) durch die Herausarbeitung von Antwortmustern erklärbar sind, die auf ein "falsches" bzw. nicht instruktionsgemäßes Ausfüllen der ADS zurückzuführen sein könnten. Zu erwarten war eine kleine Gruppe "unskalierbarer" Personen, die (im Idealfall) als eigenständige latente Klasse typisiert werden könnte und die insbesondere durch Übersehen der Polung der positiven Items gekennzeichnet ist. Zur Überprüfung dessen wurden klassifizierende Testmodelle verwendet, mit denen – in der einfachsten Form als Latent-Class-Analyse (LCA) – Eigenschaften des Messinstruments und mögliche abweichende Antwortmuster simultan betrachtet werden können (Formann 2002). Als exploratorische Analyse ermöglicht es die LCA auch, die Geordnetheit der Antwortprofile über die Subgruppen hinweg zu überprüfen (Rost 2002).

Eine darauf aufbauende und ebenfalls zu prüfende Modellvariante bestand in der Annahme, dass die Mehrzahl der Antworten dem Rasch-Modell folgt, aber eine (oder auch mehrere) Klasse(n) unskalierbarer Personen existieren. Dieser Ansatz wird als Hybrid-Modell bezeichnet (vgl. Rost & Langeheine 1997). Als weitere Hypothese sollte geprüft werden, ob es zwei (oder mehr) Gruppen gibt, in denen das

Rasch-Modell zwar passt, aber mit unterschiedlichen Parametern, z. B. infolge unterschiedlicher response sets (Mixed-Rasch-Modell, MRM). Rost et al. (1999) zeigen die Nützlichkeit dieses Ansatzes bei der Auswertung von Subskalen des NEO-FFI und Gollwitzer et al. (2005) anhand der Daten zum Ärger-Inventar STAXI.

Damit wurde das Ziel verfolgt, dass eine Verbesserung der psychometrischen Eigenschaften der ADS auch durch die Identifizierung einer (latenten) Gruppe von Personen gelingen kann, die aus ganz unterschiedlichen Gründen "unskalierbar" sind. Daher wurden Mischverteilungsmodelle eingesetzt, mit denen man sich gewissermaßen im Bereich von "Testoptimierung durch Personenselektion" (Rost 1996, Kap. 6.3.2) bewegt. Das Ziel bestand nicht darin, eine Veränderung der Skala oder sogar Neukonstruktion von Items vorzunehmen, sondern bestehende, gut eingeführte Instrumente in der praktischen Anwendung in großen Studien optimal einzusetzen, z. B. in epidemiologischen Untersuchungen und Präventionsprojekten.

Dazu wurden insgesamt drei Stichproben von 13-15-jährigen Realschülern und Gymnasiasten untersucht (n = 1077; 547; 1113). Nach dem Vergleich der verschiedenen MR-, LCA- und Hybrid-Modelle ergab sich eine 5-Klassen-Lösung der LCA als konsistentes Ergebnis über alle drei Stichproben hinweg. Dabei zeigten sich jeweils vier Klassen, deren Profile sich nicht überschnitten und die im Wesentlichen unterschiedliche Schweregrade der Depressivität abbildeten. Eine weitere Klasse (im Folgenden als Cluster 4 bezeichnet), die etwa 12-15% der Personen bzw. Bögen umfasste, war dadurch gekennzeichnet, dass sie bei negativen Items geringe Werte aufwies, bei den positiven aber vergleichsweise hohe. Dieses Profil ist vereinbar mit der Beobachtung, dass einige Jugendliche die Polung der Items übersehen haben bzw. wenig aufmerksam waren.

Weitere Anhaltspunkte dafür ergaben sich im Mittelwertsvergleich der latenten Klassen bei der Lügenskala des Eysenck Personality Inventory und der Infrequency-Skala von Chapman und Mitarbeitern, die unübliches Antwortverhalten erfasst (z. B. Ankreuzen, ohne den Inhalt der Fragen gelesen zu haben). Die Mittelwerte von Cluster 4 waren jedoch nur leicht erhöht und nicht signifikant verschieden von allen anderen Klassen. Dagegen zeigte sich, wie zu erwarten, ein deutlich erhöhter Mittelwert im Lügenindex nach Hautzinger & Bailer (1993). Ebenso wies es in den Personen-Fit-Maßen den höchsten Mittelwert auf, was sich mit dem Befund von Matschinger et al. (2000) deckt. Cluster 4 enthielt zudem überproportional Personen männlichen Geschlechts und etwas jüngere Schüler (vermehrt 13-jährige). Auch wenn dies zunächst für die Unaufmerksamkeitshypothese spricht, wäre eine alternative Erklärung, dass sich eine leichte oder beginnende Depression nur im verminderten positiven Erleben abzeichnet, aber noch nicht im Auftreten negativer Symptome.

Eine Erweiterung des Rasch-Modells um eine latente Klasse von unskalierbaren Personen (Hybrid-Modell) in der Hoffnung, eine große Klasse zu finden, für die das Rasch-Modell gilt, und eine kleine Klasse, die eine unskalierbare Restkategorie darstellt, erwies sich als nicht weiterführend.

Unser Resümee war, dass die Reliabilität und Validität der ADS-Daten – auch für auf die Auswertung in Form eines globalen Summenwerts – weitgehend gegeben ist (Keller et al. 2008). Die positiv formulierten Items sollten genutzt werden, um eine Gruppe von Unskalierbaren zu identifizieren und dadurch bessere Testeigenschaften zu erhalten. Die Ergebnisse der Analysen legten freilich nahe, dass es, entgegen der ursprünglichen Vermutung, wohl nur einige wenige unskalierbare Personen gibt, die als Falschankreuzer bezeichnet werden könnten. Es gab eher Hinweise darauf, dass ein Problem dahinter steht, was mit der erwähnten Diskussion um Uni- bzw. Bipolarität von positivem und negativem Affekt zusammenhängen könnte.

1.4 Fragestellung/Replikationsvorhaben

In diesem Beitrag soll eine Replikation dieser Befunde für die Jugendlichen mittels einer Stichprobe von 665 Schülern vorgenommen werden. Weiterhin stehen zwei große Erwachsenenstichproben zur Verfügung, um die Existenz dieses hypothetischen Clusters 4 auch altersübergreifend zu analysieren. Dabei variierte auch die Vorgabe der ADS, die in der einen Untersuchung wie üblich als Fragebogen vorgelegt wurde, in der anderen aber Teil des mündlichen Interviews war.

Da Hautzinger & Bailer (1993) mit ihrem Lügenkriterium einen manifesten Indikator zur Feststellung fraglich valider Antworten vorgeschlagen haben, kann auch der Frage nachgegangen werden, inwieweit sich die probabilistische Zuordnung der Personen eines vermuteten Clusters 4 durch das Lügenkriterium ersetzen lässt, d. h. inwieweit beide deckungsgleich sind.

In den letzten Jahren sind weitere Analysen zum Problem des wording (= positive und negative Itemformulierungen in eindimensionalen Skalen) publiziert worden. Ansätze, die auf probabilistischen Testmodellen basieren, sind eher selten geblieben. Beispielsweise verwendeten Rauch et al. (2008) das graded response model von Samejima (vgl. Rost 1996) zur psychometrischen Analyse einer Skala zur Erfassung von Optimismus, die positiv und negativ formulierte Items enthält. Im Unterschied dazu wurden in einer Reihe von Studien Strukturgleichungsmodelle (SEM) benutzt, z. B. Edwards et al. (2010) für die CES-D oder Höfling et al. (2011) für die Modifizierung einer Achtsamkeitsskala mit positiv und negativ formulierten Items. Marsh et al. (2010) haben am Beispiel der Rosenberg-Self-Esteem-Skala die unterschiedlichen Sichtweisen und die möglichen Auswertungsmodelle zusammengefasst. In umfassenden SEM-Analysen eines großen Längsschnittdatensatzes über vier Messzeitpunkte fanden sie, dass neben dem inhaltlichen Faktor (self-esteem) noch zwei latente Methodenfaktoren (einer für positive und einer für negative Items) anzunehmen sind, die über die Zeit hinweg stabil bleiben. Vautier & Pohl (2009) bestätigten in einer Test-Retest-Studie mit dem Angstfragebogen STAI, der positiv und negativ formulierte Items enthält, ebenfalls die Bipolarität des Angstkonstrukts, zu dem zeitlich stabile Methodeneffekte dazukamen.

Im vorliegenden Beitrag wird weiterhin, auch zur besseren Replikation, der MRM- bzw. LCA-Ansatz beibehalten. Methodisch wird also versucht, Personengruppen zu identifizieren, während Faktorenanalysen bzw. SEM auf der Seite der Itemklassifikation ansetzen.

2 Methodik

2.1 Stichproben

a) Jugendlichenstichprobe: In der Studie von Plener et al. (2009) wurden Schüler der 9. Klasse in Ulm (n = 665) untersucht. 57.1% von ihnen waren weiblich. Sie waren zwischen 14 und 17 Jahren alt, das mittlere Alter lag bei 14.8 Jahren (SD = 0.7). Vollständig ausgefüllt hatten die ADS 656 Schüler.

b) Normstichprobe: Hierbei handelt es sich um die Leipziger Repräsentativerhebung der deutschen Bevölkerung von Brähler und Kollegen, deren ADS-Daten bereits in das Manual von Hautzinger et al. (2012) eingingen und die hier für die Thematik dieses Beitrags nachanalysiert werden. Nach der Reduzierung auf die deutschen Teilnehmer umfasst die Stichprobe 2428 Personen, von 2343 liegt ein vollständiger Datensatz in der ADS vor. Eine auf die Personen reduzierte Stichprobe, die oberhalb des Lügenkriteriums von -28 liegt, enthält noch 2005 Probanden (85.6%). Das Geschlechterverhältnis ist leicht zugunsten der weiblichen Teilnehmer verschoben (45.1% männlich, 54.9% weiblich). Die Teilnehmer waren zwischen 14 und 91 Jahren alt, das mittlere Alter liegt bei 48.0 Jahren (SD = 18.3).

c) Älterenstichprobe: Dies ist eine Stichprobe aus dem Deutschen Alterssurvey (DEAS) von Tesch-Römer und Kollegen. Die ADS-Daten sind ebenfalls bereits im Manual verarbeitet und werden hier nachanalysiert. Verwendet wurde die Panel-Stichprobe, von der die Daten aus 2002 vorliegen. Sie umfasst 1524 Personen mit auswertbaren mündlichen Interviews (Schiel et al. 2003), von 1425 liegt eine vollständige ADS vor. Die ADS wurde in umfangreichen mündlichen Interviews erhoben. Das Alter lag im Mittel bei 63 Jahren; der jüngste war 46, der älteste Teilnehmer 91 Jahre alt. Das Geschlechterverhältnis ist leicht zugunsten der männlichen Teilnehmer verschoben (53.3% männlich, 46.7% weiblich).

2.2 Verwendete Instrumente

Allgemeine Depressionsskala (Hautzinger & Bailer 1993): Die ADS besteht aus 20 Items, die depressive Symptome erfassen (vgl. Tab. 1), die anhand einer Skala von 0 (= selten oder überhaupt nicht), 1 (= manchmal), 2 (= öfters) und 3 (= meistens, die ganze Zeit) beurteilt werden. Für diese vier Abstufungen steht jeweils eine Spezifizierung zur Verfügung, z. B. bedeutet "manchmal", dass das Symptom an einem bis zwei Tagen der letzten Woche aufgetreten ist. Alle Aussagen werden nach der Häufigkeit ihres Auftretens in der letzten Woche beurteilt. Beispielhaft sind im Folgenden ein negativ und ein positiv formuliertes Item aufgeführt:

Während der letzten Woche …

6. war ich deprimiert/niedergeschlagen.
12. war ich fröhlich gestimmt.

Für die Berechnung des Lügenkriteriums wird die Summe der 16 negativ formulierten Items gebildet, von denen die Summe der vier positiv formulierten Items abgezogen wird, die (zum Ausgleich) mit dem Faktor 4 gewichtet wird. Kreuzt also eine Person immer die 0 an, erhält sie 0 in der Summe der negativen Items und in den positiven Items nach Umkodierung jeweils 3 (= 12 als Summe der vier positiven Items). Der Lügenwert errechnet sich zu -48 (= 0 – 4*12). Wird ein kritischer Wert von kleiner -28 erreicht, dann sollte der Fragebogen gemäß der Empfehlung von Hautzinger & Bailer (1993) nicht länger als glaubwürdig angesehen werden.

Zur Verbesserung der psychometrischen Eigenschaften und der Screening-Qualität wurden immer wieder verkürzte Versionen der ADS vorgeschlagen (z. B. Santor & Coyne 1997; Cole et al. 2004, die das Rasch-Modell verwendeten; Übersicht in Hautzinger et al. 2012). Auch Hautzinger & Bailer (1993) schlagen eine Kurzversion vor (ADS-K), in der Items mit schlechter Trennschärfe herausgenommen wurden. Die ADS-K ist Grundlage mehrerer Studien in Deutschland, u.a. der Stichprobe zum Alterssurvey in diesem Beitrag. Die entfallenden Items sind aus Tabelle 1 ersichtlich.

2.3 Benutzte Modelle und Beurteilung ihrer Güte

Die Analysen zum Rasch-Modell (RM) und dem Mixed-Rasch-Modell wurden mit dem Programm WINMIRA 2001 v1.45 von v. Davier (2001) berechnet. Dabei wurde jeweils ein Partial-Credit-Modell angenommen, d. h. die Schwellenparameter sind für jedes Item unterschiedlich und klassenabhängig verschieden (Rost 1996). WINMIRA bietet auch die Möglichkeit, mit einer geglätteten Scoreverteilung (smoothed) zu arbeiten und dadurch Parameter einzusparen.

Latent-Class-Analysen wurden ursprünglich ebenfalls mit WINMIRA unter der Partial-Credit-Annahme berechnet. Alternativ wurde der Latent-Cluster-Ansatz zur Berechnung latenter Klassen genutzt, der in Latent GOLD (Vermunt & Magidson 2003) zur Verfügung gestellt wird. Darin wird nicht das vollständig unrestringierte Partial-Credit-Modell geschätzt, sondern ein Modell, in dem die Klassen durch eine Verschiebung der Itemschwierigkeit, die für alle Kategorien eines Items gleich ist, unterschieden sind (Vermunt & Magidson 2006). Wegen der daraus resultierenden Parameterreduzierung, die sich insbesondere bei Lösungen mit vielen Klassen in stabileren Schätzungen und besserer Vergleichbarkeit mit dem RM in den Informationskriterien BIC und AIC (Rost 1996) auswirkt, werden im Folgenden die Ergebnisse mit Latent GOLD in der Version 4.5 berichtet. Zur Bestimmung der optimalen Klassenzahl wird der BIC herangezogen, jedoch angesichts sehr unterschiedlicher Stichprobengrößen und -homogenität kombiniert mit inhaltlichen Überlegungen zur Herausarbeitung der hauptsächlich interessierenden und zu replizierenden Klassen.

3 Ergebnisse

3.1 Testgütekriterien nach der klassischen Testtheorie

Die Reliabilität der ADS bei der Ulmer Schulstichprobe, berechnet über Cronbachs α, liegt bei .88. Für die beiden Erwachsenenstichproben beträgt Cronbachs α .88 bzw. .90. Zieht man für die Normstichprobe nur die Personen heran, die oberhalb des Lügenkriteriums liegen (n = 2005), steigt α auf .92. Die Trennschärfen der einzelnen Items finden sich in Tabelle 1. Sie sind durchgängig akzeptabel bis sehr gut, eine Ausnahme stellt jedoch das Item 8 dar, das in der Jugendlichenstichprobe sogar eine leicht negative Trennschärfe aufweist (analog zu bisherigen Befunden bei Jugendlichen). Bei Erwachsenen trennt es besser, liegt aber immer noch unterhalb des gängigen Kriteriums von .30.

	Älterenstichprobe	Normstichprobe		Jugendliche
	n = 1425	n = 2005	n = 2343	n = 656
1. Grübeln	.57	.63	.61	.42
2. Appetit	–	.62	.60	.40
3. Fehlende Reagibilität	.67	.69	.66	.59
4. Selbstwert	–	.39	.33	.40
5. Konzentration	.61	.58	.54	.39
6. Niedergeschlagen	.75	.75	.72	.72
7. Alles anstrengend	.59	.59	.55	.47
8. Hoffnung	–	.28	.26	-.10
9. Leben ein Fehler	.45	.66	.63	.63
10. Angst	.59	.70	.67	.52
11. Schlafprobleme	.46	.57	.54	.40
12. Fröhliche Stimmung	.52	.54	.48	.60
13. Reduzierte Sprache	.46	.51	.47	.46
14. Einsam	.49	.69	.66	.65
15. Andere unfreundlich	–	.53	.50	.28
16. Leben genießen	.48	.42	.39	.55
17. Weinen	–	.58	.56	.61
18. Traurig	.69	.71	.68	.74
19. Andere ablehnend	.35	.64	.61	.51
20. Antriebslos	.59	.71	.68	.43

Tabelle 1: Trennschärfekoeffizienten für die verschiedenen Stichproben

3.2 Jugendlichenstichprobe

Der Modellvergleich in der Jugendlichenstichprobe zeigt (vgl. Tab. 2), dass das einfache Rasch-Modell (1-RM) gemäß BIC zugunsten der 2-Klassen-Lösung des Mixed-Rasch-Modells (2-RM) zu verwerfen ist. Die Berechnungen mit smoothed

weisen jeweils noch deutlich niedrigere Informationskriterien auf, doch ist die Anpassung der Scoreverteilung nicht gut bei 1-RM (RMSEA = 0.478) und bei 2-RM nur für eine Klasse (RMSEA = 0.822 und 0.0780), weshalb diese Lösungen nicht weiter betrachtet werden.

Unter den Latent-Class-Modellen wird gemäß BIC eine 4-Klassen-Lösung (4-LC) favorisiert. Bei den Modellanalysen der verschiedenen Stichproben Jugendlicher in Keller et al. (2008) ergab sich, dass bei einer Stichprobe ähnlicher Größe ebenfalls ein 3- oder 4-Klassen-Modell ein Minimum im BIC hatte, dass aber bei den anderen beiden Stichproben mit ca. 1100 Jugendlichen eine Lösung mit 5 Klassen das Optimum war. Aus Gründen der inhaltlichen Vergleichbarkeit wird daher wieder die 5- Klassen-Lösung betrachtet.

	ln(L)	n(P)	BIC	AIC	Reliabilität	Klassengröße
Partial-Credit-Modelle						
1-RM	-13404.8	119	27581.4	27047.6	.85	100
2-RM	-13009.0	237	27555.1	26491.9	.84/.74	57/43%
1-RM smoothed	-13472.0	61	27339.6	27065.9	.86	100
2-RM smoothed	-13084.9	123	26967.6	26415.8	.85/.76	50/50%
LC-Modelle						
3-LC	-13092.9	102	26847.4	26389.8	–	
4-LC	-13003.2	123	**26804.3**	26252.5	–	
5-LC	-12941.0	144	26816.1	26170.1	–	
6-LC	-12892.9	165	26856.0	26115.7	–	

Tabelle 2: Vergleich der Modellanpassung für verschiedene Rasch- und LC-Modelle in der Jugendlichenstichprobe (n = 656), Ln(L) = Log-Likelihood; n(P) = Anzahl Parameter

Die Profile der Erwartungswerte (vgl. Abb. 1) zeigen, dass Klasse 2 die geringsten Depressionswerte aufweist, und auch Klasse 1 hat eher niedrige. Die beiden Klassen umfassen über die Hälfte der Jugendlichen. Klasse 3 ist eine Gruppe mit "mittlerer" Depressivität und Klasse 5, in die etwa 9% der Jugendlichen fallen, hat mit einem Mittelwert von 37.0 hohe Werte (vgl. Tab. 3). Von diesen vier Klassen, die weitgehend geordnete Profile aufweisen, unterscheidet sich Klasse 4. Jugendliche in diesem Cluster 4 weisen in den meisten Items ähnlich niedrige Werte wie die Klassen 1 und 2 auf, haben aber bei den positiv formulierten Items deutlich höhere Werte und kommen deshalb auf einen Gesamtmittelwert von 15.6. Ersichtlich wird auch, dass das Item 8 ("dachte voller Hoffnung an die Zukunft") nicht diskriminiert.

Wegen der Diskrepanz zwischen positiv und negativ formulierten Items weist Cluster 4 auch den höchsten Mittelwert im Lügenscore auf (vgl. Tab. 3) und unterscheidet sich damit von allen anderen Gruppen (Scheffé-Test). Eine Auszählung nach dem Lügenkriterium < -28 ergibt, dass nur 34 Jugendliche (5.2%) einen klei-

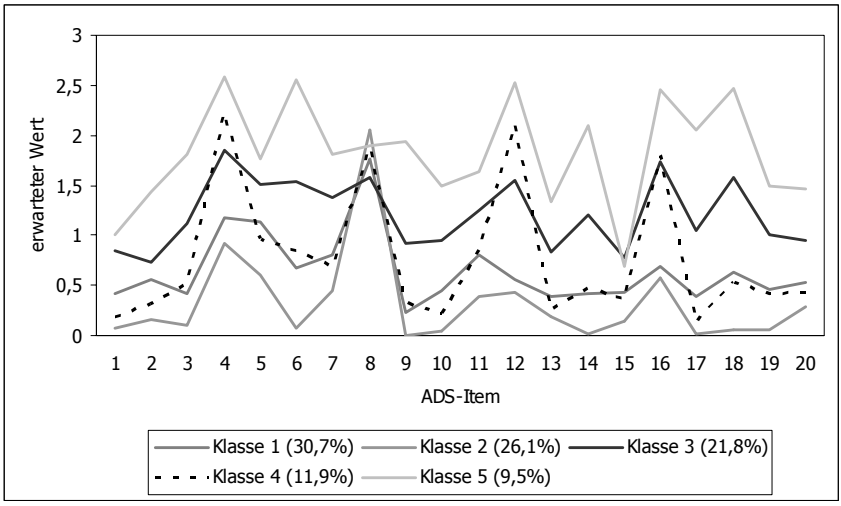


Abbildung 1: Profil der Erwartungswerte für die fünf Klassen in der Jugendlichenstichprobe (n = 656)

neren Wert aufweisen (dies steht im Gegensatz zu [vorwiegend] Erwachsenenstichproben, z. B. der Normstichprobe, in der ca. 15% das Lügenkriterium unterschreiten). Die Altersunterschiede zwischen den Klassen sind signifikant, dürften aber praktisch nicht bedeutsam sein. Beim Geschlechterverhältnis findet sich, wie erwartet, ein höherer Anteil von Mädchen in den Klassen mit mittlerer bzw. höherer Depressivität.

	Klasse 1 (n = 195)	Klasse 2 (n = 179)	Klasse 3 (n = 150)	Klasse 4 (n = 73)	Klasse 5 (n = 59)	Signifikanz
Geschlecht (% w.)	60.5	28.5	76.0	57.5	83.1	χ^2 = 98.8, df = 4, p<.001
Alter	14.8	14.7	14.8	14.9	15.0	F(4,651) = 2.62, p = .034
Lügen-kriterium	-7.7	-13.4	-9.3	-25.5	-10.8	F(4,651) = 59.00, p < .001
ADS-Gesamtwert	13.0 (3.2)	6.6 (2.8)	24.3 (4.9)	15,6 (3.4)	37.0 (4.9)	F(4,651) = 946.2, p < .001

Tabelle 3: Vergleich der fünf Klassen in der Jugendlichenstichprobe von 656 Schülern (Mittelwerte und Standardabweichungen)

Ein Vergleich dieser Jugendlichenstichprobe mit der Stichprobe von 13-15-jährigen Realschülern aus Keller et al. (2008) zeigt, dass die ADS-Mittelwerte der ersten drei Klassen sehr ähnlich sind (13.6; 6.3; 23.5). Für das "auffällige" Cluster 4 ergab sich 10.7 und für die Klasse mit ausgeprägter Depressivität 33.5. Beide Klassen weisen

also in der jetzigen Stichprobe noch etwas höhere Werte auf, was eventuell mit dem höheren Alter erklärbar ist, denn die Probanden der ersten Stichprobe waren etwa 0.4–0.8 Jahre jünger. Insgesamt ergibt sich eine gute Replikation der Muster.

3.3 Normstichprobe

Ein Vergleich der BIC legt nahe (vgl. Tab. 4), dass das einfache RM nicht die optimale Beschreibung der ADS-Antworten darstellt. Die Unterteilung in zwei Raschhomogene Gruppen erweist sich als besser geeignet, die Daten zu klassifizieren. Bei einem 3-RM ist es schwierig, stabile Lösungen zu finden. Als beste Lösung ergab sich mit BIC = 72820.1 jedoch wieder eine deutlich weniger sparsame Anpassung. Die Berechnungen mit smoothed weisen jeweils noch deutlich niedrigere Informationskriterien auf, doch ist die Anpassung der Scoreverteilung nicht gut bei 1-RM (RMSEA = 0.215) und bei 2-RM nur für eine Klasse (RMSEA = 0.054 und 0.446), weshalb diese Lösungen – wie schon bei den Jugendlichen – nicht weiter betrachtet werden.

Die LC-Modelle sind insgesamt sparsamer gemäß den Informationskriterien, doch ergibt sich bis zur Lösung mit acht Klassen noch kein Minimum. Dieses 8-LC-Modell ist jedoch bereits schwierig zu schätzen, denn es treten zahlreiche lokale Maxima auf und die kleinste Klasse enthält nur noch ca. 1.7% der Personen. Aus inhaltlichen Überlegungen heraus wird daher wieder eine 5-Klassen-Lösung betrachtet.

	ln(L)	n(P)	BIC	AIC	Reliabilität	Klassengröße
Partial-Credit-Modelle						
1-RM	-36895.0	119	74713.4	74028.0	.86	100
2-RM	-35223.1	237	72285.2	70920.3	.77/.88	60/40%
1-RM smoothed	-37022.6	61	74518.6	74167.3	.87	100
2-RM smoothed	-35332.7	123	71619.8	70911.5	.79/.90	59/41%
LC-Modelle						
4-LC	-35656.0	123	72266.3	71557.9	–	
5-LC	-35366.0	144	71849.4	71020.0	–	
6-LC	-35229.9	165	71740.1	70789.8	–	
7-LC	-35111.1	186	71665.4	70594.2	–	
8-LC	-34975.1	207	71556.4	70364.3	–	

Tabelle 4: Vergleich der Modellanpassung für verschiedene Rasch- und LC-Modelle in der Normstichprobe (n = 2343), Ln(L) = Log-Likelihood; (nP) = Anzahl Parameter

Abbildung 2 zeigt die Erwartungswerte für die einzelnen Items in den Klassen. Klasse 1 und 2 weisen niedrige bzw. sehr geringe Depressionswerte auf und umfassen zusammen knapp über die Hälfte der Personen. Klasse 3 ist, wie schon bei der Jugendlichenstichprobe, eine Gruppe mit "mittlerer" Depressivität und Klasse 5,

in die 9.5% der Personen fallen, hat mit einem Mittelwert von 30.9 bereits hohe Werte (vgl. Tab. 5). Cluster 4 weist in den meisten Items ähnlich niedrige Werte wie die Klassen 1 und 2 auf, bei den positiv formulierten aber deutlich höhere. Die Profilverläufe und ADS-Mittelwerte sind damit denen der Jugendlichenstichprobe sehr ähnlich, die insgesamt etwas höhere Mittelwerte aufweist. Item 8 diskriminiert etwas besser als bei der Stichprobe Jugendlicher, wie schon anhand der in Tabelle 1 dokumentierten Trennschärfen zu vermuten war.

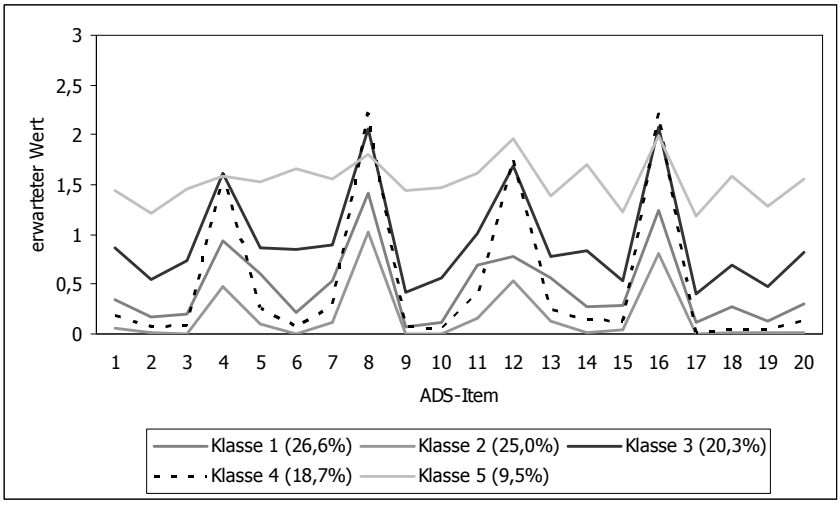


Abbildung 2: Profil der Erwartungswerte für die fünf Klassen in der Normstichprobe (n = 2343)

Ein demografischer Vergleich ergibt (vgl. Tab. 5), dass zwar mehr weibliche Personen in den beiden depressiven Klassen vertreten sind, aber weniger ausgeprägt als bei den Jugendlichen. Die Altersunterschiede sind signifikant und die Altersmittelwerte nehmen mit der Depressionsschwere zu. Im Lügenscore weist Cluster 4 wieder den höchsten Mittelwert auf.

Aus der Gesamtstichprobe wurde ein Anteil von 14.4% nach dem Lügenkriterium ausgeschlossen (vgl. Kap. 2.1). Zur Feststellung der Überlappung dieser Gruppe mit Cluster 4 interessiert daher auch die Verteilung der Lügenscores in den einzelnen Klassen, die Abbildung 3 zeigt (zur besseren Visualisierung wurden die Daten durch Addieren einer Zufallszahl "verrauscht"). In Cluster 4 lässt sich einerseits eine klare Häufung der Personen mit dem Kriterium < -28 erkennen, andererseits werden ihm aber auch viele Personen ohne erhöhten Lügenwert zugeordnet. Zudem fallen viele Personen aus Klasse 3, die eine mittlere Depressivität gemäß dem ADS-Summenwert aufweist, unter die Grenze des Lügenkriteriums. Auch in den anderen drei Klassen gibt es einige wenige Personen, die das Lügenkriterium erfüllen würden. Beispielhaft dargestellt werden soll eine Person in Klasse 5 (hohe Depression). Es handelt sich um eine 64-jährige Frau, die in den negativ formulierten Items eine Summe von 19 hat und in den positiv formulierten überall 0 an-

kreuzt, d. h. nach Umkodierung jeweils 3 und in der Summe 12 erhält. Daraus ergibt sich ein Lügenwert von 19 – 4*12 = -29. Folglich müsste sie herausgenommen werden, obwohl der ADS-Summenwert von 31 und das Antwortmuster nicht weiter auffällig erscheinen, denn es werden differenziert die Kategorien 0, 1 und 2 angekreuzt und die negativ formulierten Symptome ergeben mit 19 bereits einen über dem Durchschnitt liegenden Wert. Die Zuordnungswahrscheinlichkeit zu Klasse 5 ist relativ eindeutig mit p(Klasse = 5) = 0.88.

	Klasse 1 (n = 619)	Klasse 2 (n = 601)	Klasse 3 (n = 481)	Klasse 4 (n = 426)	Klasse 5 (n = 216)	Signifikanz
Geschlecht (% w.)	52.3	50.7	64.2	50.2	58.3	χ^2 = 27.4, df = 4, p<.001
Alter	48.4 (17.8)	44.9 (18.0)	49.7 (18.7)	49.9 (16.8)	51.8 (19.0)	F(4,2338) = 9.21, p<.001
Lügen-kriterium	-12.2	-10.7	-18.5	-30.0	-5.6	F(4,2338) = 525.8, p < .001
ADS-Gesamtwert	9.3 (2.6)	3.5 (1.9)	18.9 (3.7)	10.2 (2.5)	30.9 (5.4)	F(4,2338) = 3958.0, p < .001

Tabelle 5: Vergleich der fünf Klassen in der Normstichprobe von 2343 Personen (Mittelwerte und Standardabweichungen)

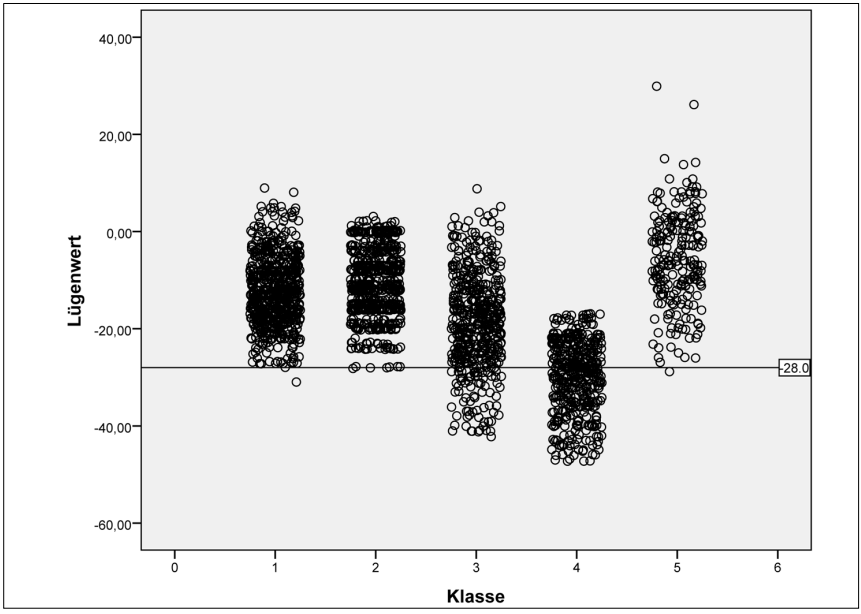


Abbildung 3: Verteilung des Lügenscores in den fünf Klassen der Normstichprobe (n = 2343)

3.4 Älterenstichprobe

Der Modellvergleich in der Älterenstichprobe (vgl. Tab. 6), in der nur die Kurzversion der ADS zum Einsatz kam, ergibt ganz analog zur Jugendlichenstichprobe, dass das einfache RM gemäß BIC zugunsten des 2-RM zu verwerfen ist. Die Berechnungen mit smoothed weisen deutlich niedrigere BIC auf, doch ist die Anpassung der Scoreverteilung weder bei 1-RM (RMSEA = 0.253) noch bei 2-RM (RMSEA = 0.180 und 0.152) gut, weshalb diese Lösungen nicht weiter betrachtet werden.

	ln(L)	n(P)	BIC	AIC	Reliabilität	Klassengröße
Partial-Credit-Modelle						
1-RM	-15311.9	89	31270.2	30801.9	.81	100
2-RM	-14982.4	177	31250.2	30318.9	.69/.85	59/41%
1-RM smoothed	-15368.6	46	31071.2	30829.1	.81	100
2-RM smoothed	-15072.1	93	30819.6	30330.3	.74/.86	55/45%
LC-Modelle						
4-LC	-15124.0	93	30923.3	30434.0	–	
5-LC	-15050.6	109	30892.8	30319.2	–	
6-LC	-14987.4	125	**30882.6**	30224.9	–	
7-LC	-14950.6	141	30925.1	30183.2	–	

Tabelle 6: Vergleich der Modellanpassung für verschiedene Rasch- und LC-Modelle in der Älterenstichprobe (n = 1425) für ADS-K, Ln(L) = Log-Likelihood; n(P) = Anzahl Parameter

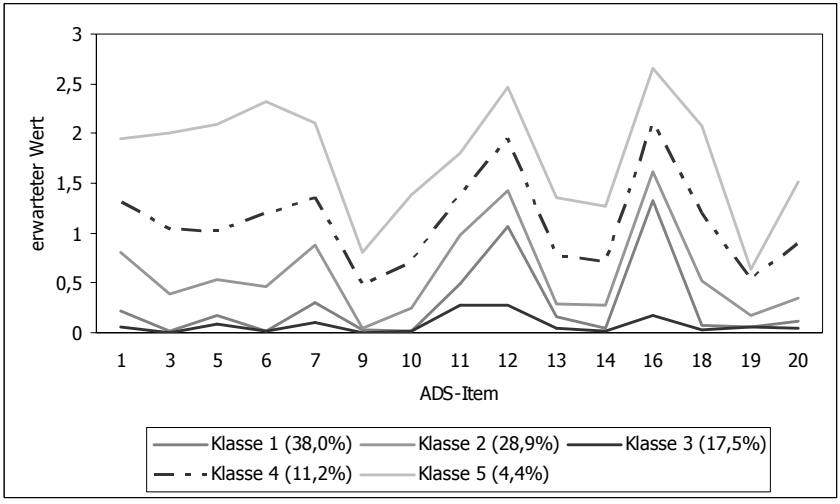


Abbildung 4: Profil der Erwartungswerte für die fünf Klassen in der Älterenstichprobe (n = 1425)

Unter den LC-Modellen wird gemäß BIC eine 6-LC favorisiert. Bei der 5-LC, die zur besseren Vergleichbarkeit dargestellt wird (vgl. Abb. 4), ergibt sich im Unterschied zu den beiden anderen Stichproben, dass die Itemprofile aller 5 Klassen geordnet sind, d. h. es gibt keine "überschneidende" Klasse. Bei der 6-LC tendiert eine Klasse (25%) in die Richtung des bisherigen Clusters 4, da bei den positiv formulierten Items 9 und 12 höhere Werte auftreten als bei den anderen. Die Erwartungswerte liegen aber bei 0.42 bzw. 0.49 und damit deutlich unter den Profilen in den Abbildungen 1 und 2; auch liegen diese Werte unterhalb von denen der drei Klassen mit höheren Depressionswerten, so dass die 6 Profile weitgehend geordnet sind und lediglich eine Tendenz in Richtung eines Clusters 4 vermutet werden könnte.

4 Diskussion

Eine Bewertung der psychometrischen Eigenschaften der ADS mit Kennwerten der klassischen Testtheorie (KTT) zeigt, dass von einer hohen Reliabilität der Skala in allen drei Stichproben ausgegangen werden kann, die für die Normstichprobe nach der Bereinigung durch das Lügenkriterium erwartungsgemäß noch gesteigert wird. Für Jugendliche sind die Trennschärfen zumeist etwas schlechter, teilweise aber auch besser, z. B. in den positiv formulierten Items 12 und 16. Item 8, das bei den Erwachsenen noch akzeptabel ist, diskriminiert bei Jugendlichen nicht, wie schon die Untersuchung anderer Stichproben von Jugendlichen gezeigt hat (vgl. Hautzinger et al. 2012, wo eine Umformulierung dieses Items zu finden ist, das danach vergleichbare Trennschärfe mit den Erwachsenen hat).

Eine Berechnung der Reliabilität nach dem RM ergibt für die drei Stichproben .86, .87 und .81 (Jugendlichen-, Norm- und Älterenstichprobe). Die Qualität der einzelnen Items soll hier nicht weiter bewertet werden. Zur Verfügung stünde der Item-Q-Index, doch entsteht damit das Problem, dass er beim RM in den großen Stichproben sehr häufig hoch signifikant ist, so dass lediglich die Unterscheidung zwischen nicht so trennscharfen und "zu trennscharfen" Items erkennbar ist. Zudem korrespondiert der Q-Index stark mit der Trennschärfeberechnung nach der KTT (Rost et al. 1999).

Die Analyse der Klassenstruktur in der Jugendlichenstichprobe ergibt, dass die Klassen gemäß ihren Itemprofilen wie auch der sich ergebenden ADS-Summenwerte eine recht gute Replikation der ersten Ergebnisse aus Keller et al. (2008) darstellen. Geht man von einer 5-LC aus, kann wieder eine Klasse (Cluster 4) identifiziert werden, die geringe Erwartungswerte bei den negativ formulierten und hohe Erwartungswerte bei den positiv formulierten Items aufweist. Die anderen vier Klassen weisen dagegen geordnete Profile auf. Dieses Ergebnis deckt sich mit demjenigen für die ursprünglichen Stichproben, und die Klassengrößen und die sich ergebenden Summenwerte sind ebenfalls recht ähnlich.

Bei der Analyse der großen Normstichprobe zeigt sich ebenfalls dieses Muster von vier geordneten Klassen und einer "überschneidenden" Klasse, die knapp 19% der Personen umfasst. Diese Größenordnung ist vergleichbar mit den Ergebnissen von

Matschinger et al. (2000), die mittels LCA eine Klasse von etwa 18% der Probanden identifiziert haben, die (nach ihrer Interpretation) anscheinend die Polung übersehen hat. Im Mittel erhielten diese Probanden einen ADS-Summenwert von 12.3; in der vorliegenden Normstichprobe beträgt der Summenwert 10.2. In der Jugendlichenstichprobe ist er höher (15.6) und Cluster 4 kleiner (9%). In der Älterenstichprobe, bei der nur die Kurzversion der ADS eingesetzt wurde, kann dieses Muster hingegen nicht beobachtet werden. Bei der 5-LC ergeben sich keine Überschneidungen in den Profilen und bei der 6-LC allenfalls eine Tendenz dazu.

Zu diskutieren ist sicherlich die Festlegung auf die optimale Anzahl von Klassen in jeder Stichprobe. Gemäß BIC ergeben sich sehr unterschiedliche Klassenzahlen in den einzelnen Stichproben, die vermutlich von der Stichprobengröße beeinflusst werden (obwohl diese ja in der Formel berücksichtigt wird), aber mehr noch von ihrer unterschiedlichen "Komplexität". Die Norm- und die Älterenstichprobe basieren im Unterschied zur Jugendlichenstichprobe auf bundesweiten Erhebungen und sie decken vor allem einen (weitgehend) anderen und viel größeren Altersbereich ab. Dennoch scheint es einige übergreifende stabile Muster zu geben. Die beiden Klassen mit mittlerer und hoher Depressivität sind etwa gleich groß, ebenso ergeben sich jeweils ein (oder zwei) symptomarme Klassen. Beim Übergang von der 4-LC auf die 5-LC ergibt sich in der Jugendlichen- wie auch in der Normstichprobe Cluster 4 mit seinem aus der Geordnetheit der Profile herausfallenden Muster.

Bei Anwendung der ADS-K, in der einige "schlechtere" Items entfernt sind, war in der Älterenstichprobe das überschneidende Cluster 4 hingegen nicht zu erkennen und nur tendenziell in der 6-LC. Es stellt sich daher die Frage, ob das Entfernen dieser Items, insbesondere des Items 8 (und evtl. 4) dafür gesorgt hat, oder aber die andere Vorgabemodalität: Bei der Normstichprobe fand die übliche Fragebogenerhebung statt. Obwohl bei der Älterenstichprobe die Erhebung in ein mündliches Interview eingebettet war, wurde die ADS-K aber vermutlich trotzdem in Papierform zum Ausfüllen überreicht (dies wird nicht ganz klar im Methodenbericht der DEAS). Jedenfalls war der Interviewer normalerweise immer anwesend. Eine exploratorische LCA der Normstichprobe, in die nur die 15 Items der ADS-K eingingen (nicht dargestellt im Ergebnisteil), ergibt wieder eine Klasse mit überschneidendem Muster, ganz analog zu dem in Kapitel 3.3 bzw. Abbildung 2 dargestellten Ergebnis. Das Entfernen der "schlechten" Items allein scheint also nicht zu einer Geordnetheit der Antwortprofile zu führen.

Zur Frage danach, ob zwischen Cluster 4 und dem (deterministisch berechneten) Lügenwert, der eine auszuschließende Gruppe über das Lügenkriterium definieren soll, Deckungsgleichheit besteht: Es gibt erwartungsgemäß eine starke Überlappung beider Gruppen, aber bei weitem keine Deckungsgleichheit. Im Gegenteil treten auch in den anderen Klassen einige Personen auf, die das Lügenkriterium erfüllen; in der Normstichprobe relativ viele in der Klasse mit mittlerer Depressivität. Gerade diese Gruppe würde den hypothetischen Subtyp mit noch geringer negativer Symptomatik und gleichzeitig schon wenig positivem Erleben nahe legen. Mit einem mittleren ADS-Summenwert von 18.9 läge dieser Subtyp bereits über dem

für Screening-Zwecke häufig als Cut-off-Kriterium verwendeten Wert von 16 (vgl. Hautzinger et al. 2012).

Mit großer Wahrscheinlichkeit ist Cluster 4 auch keine homogene Klasse. Auf der einen Seite dürfte es (einen kleineren Teil von) Personen enthalten, die unaufmerksam ankreuzen (Personen, die durchgängig 0 ankreuzen und damit den maximalen Lügenwert von -48 erreichen, sind allerdings in der Normstichprobe nicht aufgetreten!; sie wären sicherlich auffällig und verdächtig, mit hoher Wahrscheinlichkeit unmotiviert und ohne allzu genaues Lesen geantwortet zu haben). Auf der anderen Seite dürfte es eine Reihe von Personen geben, bei denen sich, gemäß einer vorläufigen Erklärung, eine leichte oder beginnende Depressivität nur im verminderten positiven Erleben abzeichnet, aber noch kaum im Auftreten negativer Symptome. Gerade bei jüngeren männlichen Probanden ist zudem vermehrt mit Reaktanz und Abwehr zu rechnen, so dass die negative Seite nicht zugegeben wird, wohl aber (schon) das Fehlen der positiven. Zur Untermauerung dieser Vermutung sind Längsschnittbeobachtungen nötig, wobei die Ergebnisse in Keller et al. (2008) bisher wenig Anhaltspunkte dafür lieferten. Das auffällige Cluster 4 ließ zu den vier Erhebungszeitpunkten, die etwa 3 Monate Abstand voneinander hatten, keine gehäuften Wanderungsbewegungen erkennen: Ein bis zwei Drittel verblieben jeweils in Cluster 4, und sehr wenige wechselten in die Klasse mit hoher Depressivität. Eventuell waren die zeitlichen Intervalle ungünstig für die Fragestellung und zu berücksichtigen ist auch die Zahl der drop-outs, die sich insgesamt zwar im normalen Rahmen solcher Studien bewegte, aber möglicherweise könnte gerade bei dem sensiblen Übergang in Klasse 5 eine Häufung der drop-outs vermutet werden. Die relativ große Stabilität, in Cluster 4 zu bleiben, spricht ebenfalls für ein überdauerndes Erleben und weniger für andauernde Lustlosigkeit beim Ausfüllen, wäre aber auch vereinbar mit der Annahme zweier zeitlich stabiler Response-Style-Faktoren, wie sie von Marsh et al. (2010) gefunden wurden.

Neben einer solchen längsschnittlichen Betrachtung der Klassen könnte auch eine Beschreibung des Clusters 4 durch weitere Variablen im Datensatz vorgenommen werden (was aus Platzgründen hier nicht dargestellt wird). Ebenso überlegenswert für weitere Studien wäre eine direkte Frage nach diesem Typ von Erleben, die altersabhängig formuliert werden müsste: Bei Jugendlichen vermuten wir, dass es mit dem Phänomen des Gelangweilt-seins erfragbar sein könnte; bei Älteren dürfte dieses Erleben eine realistische und nicht weiter pathologische Lebenssituation darstellen. In jedem Fall identifiziert Cluster 4 eine interessante Subgruppe von Personen, deren ADS-Summenwert wegen der qualitativen Besonderheit zwar nicht unbesehen übernommen werden sollte, die man aber auch nicht pauschal als fraglich valide herausnehmen sollte. Ebenso wenig sollte man auf der Itemebene resignativ handeln im Sinne von "overall the positive items simply provided more noise than information in this sample" (Stansbury et al. 2006, S. 19) und die positiv formulierten Items einfach weglassen.

Die Auswertungen dieses Beitrags lassen sich dahingehend zusammenfassen, dass mit dem verwendeten probabilistischen Ansatz eine sinnvoll erscheinende Subgruppe herausgearbeitet werden kann. Dabei kann momentan noch nicht ent-

schieden werden, ob es sich um eine inhaltlich erklärbare Subgruppe mit beginnender oder subklinischer Depression handelt oder ob sich die zwei Response-Style-Methodenfaktoren von Marsh et al. (2010) darin widerspiegeln. Auf alle Fälle kann die pauschalisierende Annahme einer potenziellen "Lügengruppe" mit fraglicher Datenqualität verworfen werden, und daher erscheinen weitere Untersuchungen lohnenswert, um diese spezielle Subgruppe näher zu charakterisieren.

Zur Skalierung der Lebensqualität: Probabilistische Testmodelle bei der Bestimmung von Nutzwerten für die gesundheitsökonomische Evaluation

Thomas Kohlmann, You-Shan Feng, Bas Janssen & Simon Pickard

1 Gesundheitsbezogene Lebensqualität: Konzepte und Messmethodik

1.1 Lebensqualitätsmessung in der Medizin

Es ist das allgemeine Ziel der medizinischen Versorgung, die Gesundheit der von ihr betreuten Menschen zu erhalten und zu fördern, bei eingetretener Erkrankung diese nach Möglichkeit wiederherzustellen oder ein Fortschreiten der Erkrankung zu verhindern oder zumindest zu verlangsamen. Um bestimmen zu können, in welchem Grad dieses Ziel erreicht wird, müssen geeignete Indikatoren definiert und gemessen werden. Diese Indikatoren ermöglichen es nicht nur, den Erfolg medizinischer Versorgungsmaßnahmen insgesamt zu beurteilen, z. B. die Verkürzung der Krankheitsdauer, die Erhöhung der Überlebenschancen oder die Verbesserung des Gesundheitszustands, mit ihrer Hilfe kann auch die Effektivität verschiedener Behandlungsalternativen vergleichend untersucht werden.

Als Ergänzung der traditionellen "objektiven" Indikatoren (Überlebenszeit, klinische Befunde) hat in den letzten Jahren die Beurteilung des Gesundheitszustands durch die Patienten selbst in vielen Bereichen der Medizin ein ständig wachsendes Interesse gefunden. Diese mit Begriffen wie subjektive Gesundheit, gesundheitsbezogene Lebensqualität oder patient-reported outomes bezeichneten Indikatoren erweitern das Spektrum der medizinischen Evaluationskriterien in wesentlichem Umfang. Der Begriff Lebensqualität als solcher ist in Deutschland sogar in die Gesetzgebung zur Bestimmung des Nutzens von medizinischen Maßnahmen aufgenommen worden (z. B. in die Arzneimittel-Nutzenbewertungsverordnung, AM-NutzenV).

Zur Messung von Indikatoren der subjektiven Gesundheit, insbesondere der gesundheitsbezogenen Lebensqualität, stehen zahlreiche Erhebungsverfahren zur Verfügung (Schumacher et al. 2003). Zumeist handelt es sich dabei um standardisierte Fragebögen, mit denen z. B. die Wahrnehmung der körperlichen Gesundheit, das psychische Befinden und die Einflüsse der Krankheit auf die soziale Situation erfasst werden.

Die Entwicklung dieser Befragungsinstrumente stützte sich in vielen Fällen auf moderne psychometrische Verfahren, wie sie in zahlreichen Bereichen der Fragebogenkonstruktion angewandt werden. Während allerdings in der Anfangsphase der Lebensqualitätsmessung in der Medizin noch die Methoden auf der Basis der klassischen Testtheorie ganz im Vordergrund standen, haben das Rasch-Modell und weitere aus der Item-Response-Theorie stammende Modelle gerade in den letzten Jahren eine stark zunehmende Beachtung gefunden. Die in diesem Beitrag angewandten probabilistischen Testmodelle zur Skalierung von Items eines Lebensqualitätsfragebogens fügen sich demnach lückenlos in eine aktuelle Entwicklungslinie der Forschung in der Medizin und in den Gesundheitswissenschaften ein.

1.2 Deskriptive und präferenzbasierte Lebensqualitätsmessung

Die Erhebungsverfahren zur Messung der gesundheitsbezogenen Lebensqualität lassen sich – sowohl inhaltlich als auch methodisch – in zwei Gruppen untergliedern:

- Die deskriptiven Verfahren folgen den traditionellen Prinzipien der psychometrischen Testkonstruktion: Zur Messung einer nicht direkt beobachtbaren (latenten) Eigenschaft einer Person, z. B. der körperlichen Funktionsfähigkeit oder des psychischen Wohlbefindens, werden Fragen formuliert, deren Beantwortung Rückschlüsse auf die Ausprägung der latenten Eigenschaft erlauben. Der Grad, in dem bestimmte Fragen derartige Rückschlüsse ermöglichen, also die latente Eigenschaft auf der Ebene der Beobachtungsdaten repräsentieren, kann mit psychometrischen Modellen, wie denen der klassischen oder der probabilistischen Testtheorie, bestimmt werden. Bei der praktischen Anwendung dieser deskriptiven Verfahren ergeben sich in der Regel Merkmalsprofile, mit denen die Ausprägungen der verschiedenen Dimensionen der Lebensqualität beschrieben werden.
- Die präferenzbasierten Verfahren weisen bei oberflächlicher Betrachtung große Ähnlichkeit mit den eben dargestellten deskriptiven Verfahren auf: Wie bei jenen wird der subjektive Gesundheitszustand multidimensional mit standardisierten Fragen erfasst. Die präferenzbasierten Verfahren stehen jedoch in einem anderen, einem gesundheitsökonomischen Anwendungskontext und folgen einem völlig anderen Konstruktionsprinzip. In gesundheitsökonomischen Kosten-Nutzwert-Analysen (cost-utility analysis) wird die gesundheitsbezogene Lebensqualität den Kosten medizinischer Maßnahmen gegenübergestellt, um damit z. B. die Kosten für einen bestimmten Gewinn an Lebensqualität errechnen zu können. Hierzu müssen die erhobenen multidimensionalen Lebensqualitätsdaten zu einer eindimensionalen Maßzahl, einem Nutzwert, verdichtet werden.

Beim Vergleich der deskriptiven und der präferenzbasierten Lebensqualitätsmessung ist ganz besonders herauszustellen, dass sich beide in psychometrischer Hinsicht fundamental unterscheiden. Während bei der Entwicklung deskriptiver Erhebungsverfahren problemlos die traditionellen Methoden der Testkonstruktion angewandt werden können, beruht die Berechnung eines Nutzwertes bei der präferenzbasierten Lebensqualitätsmessung auf Ergebnissen von Standard-Gamble-

oder Time-Trade-Off-Techniken, wie sie am Beispiel des EQ-5D-Fragebogens in Kapitel 1.3 erläutert werden.

1.3 Der EQ-5D-Fragebogen

Die im Jahre 1987 gegründete EuroQol-Gruppe ist ein internationales Netzwerk von Wissenschaftlern aus unterschiedlichen Fachdisziplinen, die sich mit Fragen der Messung subjektiver Gesundheitsparameter, insbesondere für die Anwendung in gesundheitsökonomischen Analysen, beschäftigen (www.euroqol.org). Der von dieser Gruppe entwickelte EQ-5D-Fragebogen gehört zu den international am häufigsten eingesetzten Erhebungsverfahren zur präferenzbasierten Messung der gesundheitsbezogenen Lebensqualität (Rabin & de Charro 2001). Er besteht im Kern aus fünf Einzelfragen, die sich auf körperliche und psychische Beeinträchtigungen beziehen und die in einem dreistufigen Antwortformat (keine Probleme, mäßige Probleme, extreme Probleme) von den Befragten beantwortet werden sollen. Die Items und Antwortvorgaben des EQ-5D bilden ein deskriptives System von Gesundheitszuständen, das durch die 3^5 (= 243) möglichen Antwortmuster definiert ist (vgl. Abb. 1).

Dieser Teil des Fragebogens enthält fünf Gruppen (A-E) mit jeweils drei Aussagen.
Bitte lesen Sie jede Gruppe sorgfältig durch und kreuzen Sie in jeder Gruppe das *eine* Kästchen [ ] an, das *am besten* auf Sie zutrifft.

A. Beweglichkeit/Mobilität:
Ich habe keine Probleme herumzugehen . . . . . [ ]
Ich habe einige Probleme herumzugehen . . . . . [ ]
Ich bin ans Bett gebunden . . . . . [ ]

B. Für sich selbst sorgen:
Ich habe keine Probleme, für mich selbst zu sorgen . . . . . [ ]
Ich habe einige Probleme, mich selbst zu waschen oder mich anzuziehen . . . . . [ ]
Ich bin nicht in der Lage, mich selbst zu waschen oder mich anzuziehen . . . . . [ ]

C. Allgemeine Tätigkeiten:
(z. B. Arbeit, Studium, Hausarbeit, Familien- oder Freizeitaktivitäten)
Ich habe keine Probleme, meinen alltäglichen Tätigkeiten nachzugehen . . . . . [ ]
Ich habe einige Probleme, meinen alltäglichen Tätigkeiten nachzugehen . . . . . [ ]
Ich bin nicht in der Lage, meinen alltäglichen Tätigkeiten nachzugehen . . . . . [ ]

D. Schmerzen/Körperliche Beschwerden:
Ich habe keine Schmerzen oder Beschwerden . . . . . [ ]
Ich habe mäßige Schmerzen oder Beschwerden . . . . . [ ]
Ich habe extreme Schmerzen oder Beschwerden . . . . . [ ]

E. Angst/Niedergeschlagenheit:
Ich bin nicht ängstlich oder deprimiert . . . . . [ ]
Ich bin mäßig ängstlich oder deprimiert . . . . . [ ]
Ich bin extrem ängstlich oder deprimiert . . . . . [ ]

Abbildung 1: Die Originalversion des EQ-5D-Fragebogens mit dreistufiger Antwortskalierung

Wie bereits ausgeführt, besteht die Besonderheit präferenzbasierter Instrumente darin, dass aus den Angaben der Befragten keine testtheoretischen Scores (z. B. Faktorscores, Personenparameter) geschätzt werden. Die Berechnung des für einen bestimmten Gesundheitszustand geltenden Nutzwerts erfolgt anhand eines Tarifs, der im Rahmen von Valuierungsstudien (zumeist in der Allgemeinbevölkerung) festgelegt wurde. In diesen Valuierungsstudien werden die Befragten gebeten, verschiedene Gesundheitszustände, die durch das deskriptive System des Fragebogens vorgegeben sind, vergleichend zu bewerten – deshalb die Bezeichnung präferenzbasiert. Diese Vergleiche werden anhand verschiedener Befragungsmethoden durchgeführt, von denen das Standard Gamble und die Time-Trade-Off-Methode die am häufigsten angewandten sind (Schöffski & von der Schulenburg 2007). Es wird dabei ein Szenario vorgegeben, in dem die Befragten aus zwei Möglichkeiten auswählen sollen: Sie können entweder im gegenwärtigen (beeinträchtigten) Gesundheitszustand verbleiben (Wahl B) oder sich einer Therapie unterziehen (Wahl A), die allerdings mit dem Risiko des sofortigen Todes (Standard Gamble) oder dem Verlust an Restlebenszeit (Time Trade-Off) verbunden ist. Bei der Standard-Gamble-Methode müssen die Probanden angeben, mit welcher Wahrscheinlichkeit p des sofortigen Todes eine zu kompletter Heilung führende Therapie verbunden sein darf, um für sie noch akzeptabel zu sein. Bei der Time-Trade-Off-Methode wird der noch akzeptable Verlust an Lebenszeit t durch eine zu kompletter Heilung führende Therapie erfragt (vgl. Abb. 2). Der gesuchte Nutzwert eines beeinträchtigten Gesundheitszustands wird dann als Funktion von p oder t bestimmt und auf den Bereich 1 (= vollständige Gesundheit) und 0 (= Tod) normiert. Je größer die akzeptierte Wahrscheinlichkeit des sofortigen Todes bzw. der akzeptierte Verlust an verbleibender Lebenszeit ist, desto schlechter wird der Nutzwert des Gesundheitszustandes bewertet.

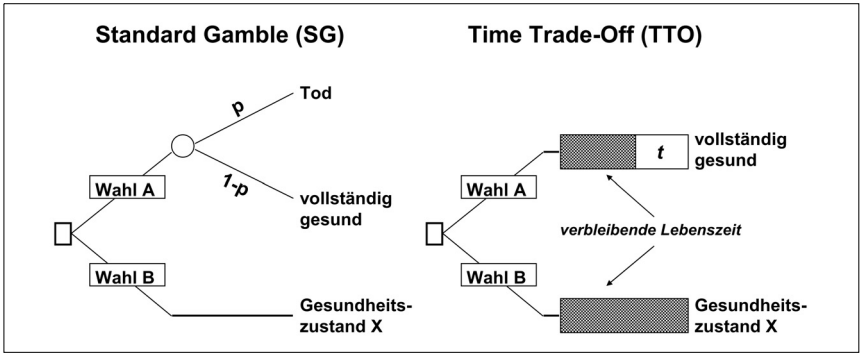


Abbildung 2: Schematische Darstellung der Standard-Gamble- und der Time-Trade-Off-Methode zur Bestimmung von Nutzwerten für definierte (tatsächlich bestehende oder fiktive) Gesundheitszustände

Im Prinzip wäre es möglich, für alle 243 durch den EQ-5D-Fragebogen definierten Gesundheitszustände den entsprechenden Nutzwert mit der Standard-Gamble- oder der Time-Trade-Off-Methode zu bestimmen. Ein solches Vorgehen wäre al-

lerdings sehr aufwändig und würde eine enorme Stichprobengröße erforderlich machen, da erfahrungsgemäß von einer in einem Interview befragten Person nicht mehr als 15 bis 20 Gesundheitszustände beurteilt werden können und pro Gesundheitszustand wiederum insgesamt etwa 20 Beurteilungen vorliegen sollten, um eine ausreichend stabile Schätzung der Nutzwertes zu erhalten. In der Praxis wird deshalb für die Valuierungsstudie eine Untermenge aller Gesundheitszustände ausgewählt. Die dafür erhaltenen Nutzwerte werden als abhängige Variable in ein Regressionsmodell aufgenommen, in dem die Ausprägungen der die Gesundheitszustände definierenden Fragebogenitems als Prädiktoren verwendet werden. Durch geeignete Kodierung der Effekte können so die Abschläge (Disutilities) geschätzt werden, die z. B. aus dem Übergang von "keine Probleme beim Herumgehen" zu "einige Probleme beim Herumgehen" resultieren. Aus der Regression von den mit der Standard-Gamble- oder der Time-Trade-Off-Methode bestimmten Nutzwerten auf die Items des Fragebogens ergibt sich damit ein Algorithmus (Tarif), mit dem für jedes mögliche Antwortmuster ein Nutzwert berechnet werden kann.

Es ist unmittelbar zu erkennen, dass diese Berechnung eines eindimensionalen Scores keine Gemeinsamkeit mit der psychometrischen Testkonstruktion hat. Die Items im Fragebogen werden nicht als reflektive Indikatoren einer latenten Eigenschaft betrachtet, und es ist unerheblich, ob ein Modell der klassischen oder probabilistischen Testtheorie gilt. Entscheidend ist vielmehr, ob die Fragebogenitems, die das deskriptive System bilden, die relevanten Gesundheitszustände ausreichend gut definieren und ob die mit der Standard-Gamble- oder Time-Trade-Off-Methode bestimmten Nutzwerte in der Regressionsanalyse durch die Antwortmuster im statistischen Sinne ausreichend gut erklärt werden können.

Der EQ-5D-Fragebogen wird bereits weltweit in zahlreichen Sprachen verwendet, in vielen Ländern wurden für die dreistufige Version Valuierungsstudien durchgeführt und entsprechende Nutzwert-Tarife bestimmt (Szende et al. 2007). Diese Valuierungsstudien mussten selbst bei Beschränkung auf ausgewählte Gesundheitszustände mit großen Stichproben durchgeführt werden, um stabile Ergebnisse liefern zu können und waren sehr kosten- und zeitaufwändig, da die Befragungen in der Regel in der Form persönlicher Interviews erfolgen mussten.

1.4 Der Übergang von der dreistufigen zur fünfstufigen Antwortskala

Angesichts der besonderen Art, mit der Nutzwert-Tarife bestimmt werden, wird sofort deutlich, dass jede nachträgliche Änderung eines präferenzbasierten Fragebogens neuerliche Valuierungsstudien notwendig macht. Eine solche Situation ist im Jahre 2005 für den EQ-5D-Fragebogen eingetreten. Nachdem in zahlreichen Studien beobachtet wurde, dass das auf einer nur dreistufigen Antwortskala beruhenden deskriptive System zu grob ist, um kleine, aber relevante Veränderungen des Gesundheitszustands sicher abbilden zu können, hat die EuroQol-Gruppe beschlossen, einen Fragebogen mit fünf Antwortstufen (EQ-5D 5L; Herdman et al. 2011) zu entwickeln. Die fünf Items wurden beibehalten. Zusätzlich zu den drei

ursprünglichen Antwortstufen wurden zwei Zwischenstufen in die Vorgaben aufgenommen (vgl. Abb. 3).

Abbildung 3: Das neue fünfstufige Antwortformat im EQ-5D-5L-Fragebogen am Beispiel der Dimension Schmerzen/Körperliche Beschwerden (im Text kursiv hervorgehoben sind die eingefügten Zwischenkategorien)

Die Entscheidung für eine fünfstufige Fragebogenversion impliziert die Notwendigkeit, für jede Sprachversion bzw. für jedes Land eigene Valuierungsstudien durchzuführen, was mehrere Jahre in Anspruch nehmen wird. Um für Anwender des Fragebogens, die bereits jetzt die neue Version einsetzen wollen, eine Zwischenlösung anbieten zu können, wurde untersucht, ob es nicht möglich ist, durch statistische Modellierung eine vorläufige (intermediäre) Version des neuen Tarifs zu gewinnen. Erforderlich sind hierzu Befragungsdaten, in denen Angaben sowohl für die drei- als auch für die fünfstufige Version des EQ-5D enthalten sind. Durch geeignete statistische Modelle könnten dann aus den mit dem dreistufigen Fragebogen erzielten Ergebnissen die für die fünfstufigen Antwortvorgaben verwendbaren Disutilities berechnet werden.

2 Methodik und Datengrundlage

2.1 Statistische Methoden zur Bildung eines intermediären Tarifs

Für derartige Modellierungen kommen verschiedene Methoden in Betracht. Zwei dieser Methoden wurden von Mortimer & Segal (2008) als transfer to utility und response mapping bezeichnet. Bei der Transfer-to-Utility-Methode werden zunächst die Nutzwerte aus den Daten der dreistufigen Version nach dem dafür verfügbaren Tarif berechnet. In einem zweiten Schritt werden in einer linearen Regression mit diesen Nutzwerten als abhängige Variable die Regressionskoeffizienten für die dummy-kodierten fünfstufigen Antworten geschätzt (vgl. Gleichung 1). Es ergibt sich dadurch eine Schätzfunktion, mit der die Nutzwerte aus den Antwortmustern der fünfstufigen Version bestimmt werden können.

$$\hat{y}_i = a + \sum_{j=1}^{J} \sum_{c=2}^{m_j} b_{jc} X_{jc} \,. \qquad (1)$$

Die abhängige Größe $\hat{y}_i$ bildet der für Person i (i = 1,…,N) gemäß einem existierenden Tarif berechnete Nutzwert. Die rechte Seite der Gleichung enthält ein Set

von Regressionskoeffizienten und Dummy-Variablen für die Ausprägungen der Fragebogenitems: Die Regressionskonstante a sowie Regressionskoeffizienten b_{jc} und Dummies X_{jc} für die Kategorie c von Item j (j = 1,…,J; c = 2,…,m_j).

Bei der Response-Mapping-Methode werden die Angaben aus der fünfstufigen Fragebogenversion ebenfalls als Prädiktorvariablen in ein Regressionsmodell einbezogen. Die abhängige Variable bildet jedoch nicht der bereits unter Anwendung eines bestehenden Tarifs berechnete Nutzwert, sondern die Antworten der dreistufigen Fragebogenversion. Für jedes Item dieser Version werden in einer multinominalen logistischen Regression die für die drei Antwortstufen unter dem Modell erwarteten Wahrscheinlichkeiten bestimmt. Um für einen bestimmten Befragten einen diskreten erwarteten Wert für das betreffende Item zu erhalten, wird ein ganzzahliger Zufallswert aus einer Multinomialverteilung gezogen, deren Verteilungsparameter die im Regressionsmodell ermittelten Wahrscheinlichkeiten sind.

Diese beiden Methoden zur Bildung eines intermediären Tarifs für die fünfstufige Version des EQ-5D-Fragebogens wurden bereits in verschiedenen Stichproben erfolgreich angewandt (van Hout et al. 2011). In diesem Kontext stellte sich die Frage, ob und wenn ja, in welcher Weise auch probabilistische Testmodelle zu solchen Schätzungen herangezogen werden können. Ausgangspunkt war die Überlegung, dass für jedes der jeweils aus dem drei- und dem fünfstufigen Item bestehenden Paar von Items angenommen werden kann, dass die zu dem selben Paar gehörenden Items die selbe latente Eigenschaft repräsentieren (Mobilität, Selbstpflege, allgemeine Tätigkeiten, Schmerzen/körperliche Beschwerden, Angst/Niedergeschlagenheit; vgl. Tab. 1). Aus den Schätzungen der Personenparameter könnte dann eine Berechnungsvorschrift entwickelt werden, mit der ein Tarif für die fünfstufige Fragebogenversion gebildet werden kann. Ein besonderer Vorteil im Vergleich zu den Transfer-to-Utility- oder den Response-Mapping-Methoden bestünde darin, dass nicht nur eine empirisch gestützte Abbildung des fünfstufigen in das dreistufige System gefunden werden kann, sondern dass gleichzeitig weitere Informationen über die Messcharakteristik der Items im EQ-5D-Fragebogen gewonnen werden können.

Aus der Familie der probabilistischen Testmodelle ist das Rasch-Modell für mehrkategorielle Items (Partial-Credit-Modell, Masters 1982) in besonderer Weise für diese Fragestellung geeignet. Da es sich dabei um ein sehr strenges messtheoretisches Modell handelt, wurden für die hier dargestellten Analysen vergleichende Auswertungen mit einem weniger strengen Testmodell durchgeführt, in dem bei jedem Itempaar ein für beide Items identischer Diskriminationsparameter α geschätzt wird (vgl. Gleichung 2).

$$\text{prob}\left(x_{jk}\middle|\theta\right) = \frac{\exp\left(\sum_{h=1}^{k}\alpha_j\left(\theta-\beta_{jh}\right)\right)}{\sum_{c=1}^{m_j}\exp\left(\sum_{h=1}^{c}\alpha_j\left(\theta-\beta_{jh}\right)\right)} \qquad (2)$$

Modelliert wird in Abhängigkeit von der Ausprägung der latenten Variable θ die Wahrscheinlichkeit prob($x_{jk}|\theta$), dass bei Item j (j = 1,...,J) die k-te Kategorie (k = 1,...,m_j) gewählt wird. Neben den Kategorienparametern β_{jh} enthält das Modell die Diskriminationsparameter α_j. Unter der Restriktion $\alpha_j = 1$; j = 1,...,J ergibt sich das Rasch-Modell für mehrkategorielle Items, aus der Restriktion $\alpha_j = \alpha$; j = 1,...,J resultiert das hier verwendete Modell mit einem Diskriminationsparameter (identisch für beide Items). Werden die Diskriminationsparameter für die Items frei geschätzt, entspricht Gleichung 2 dem Generalized-Partial-Credit-Modell von Muraki (1992). Zur Identifikation der Modellparameter sind geeignete Normierungsbedingungen zu definieren (z. B. $\beta_{j1} = 0$).

Die Parameter des Rasch-Modells bzw. des Modells mit einem Diskriminationsparameter wurden mit dem für die Programmiersprache R entwickelten Programmpaket ltm unter Verwendung der Marginal-Maximum-Likelihood-Methode (MML) geschätzt (Rizopoulos 2006). Dabei wurden mehrere zufällige Sets von Startwerten verwendet. Die Schätzung der Personenparameter erfolgte mit drei unterschiedlichen Ansätzen, der Empirical-Bayes(EB)-Schätzung, der Expected-a-Posteriori(EAP)-Schätzung und der Schätzung mit multipler Imputation (MI) der Personenparameter (Rizopoulos 2010).

Um einen Schätzwert für die Lage jeder der drei- bzw. fünfstufigen Antwortvorgaben auf dem latenten Kontinuum zu erhalten, wurden für jedes Item und jede Antwortskalierung die kategorienspezifischen Mittelwerte der Personenparameter berechnet. Die Disutilities der fünfstufigen Antwortskala wurden anhand der zwischen den kategorienspezifischen Mittelwerten bestehenden Abstände proportional bestimmt (Pickard et al. 2007; vgl. Gleichung 3.1-3.5), wobei in beiden Antwortskalen die Kategorie "keine Probleme" auf den Wert 0 festgelegt wurde und die Disutilities für die Antwortkategorien "einige Probleme" ($D_{3.2}$) bzw. "extreme Probleme" ($D_{3.3}$) aus dem Dolan-Tarif (Dolan 1997) für die dreistufige Fragebogenversion verwendet wurden.

$$D_{5.1} = D_{3.1} = 0 \tag{3.1}$$

$$D_{5.2} = D_{5.3}\frac{\overline{\theta}_{5.1} - \overline{\theta}_{5.2}}{\overline{\theta}_{5.1} - \overline{\theta}_{5.3}} \tag{3.2}$$

$$D_{5.3} = D_{3.2}\frac{\overline{\theta}_{5.1} - \overline{\theta}_{5.3}}{\overline{\theta}_{3.1} - \overline{\theta}_{3.2}} \tag{3.3}$$

$$D_{5.4} = D_{5.3}\frac{\overline{\theta}_{5.4} - \overline{\theta}_{5.5}}{\overline{\theta}_{5.3} - \overline{\theta}_{5.5}} + D_{5.5}\frac{\overline{\theta}_{5.3} - \overline{\theta}_{5.4}}{\overline{\theta}_{5.3} - \overline{\theta}_{5.5}} \tag{3.4}$$

$$D_{5.5} = D_{3.3}\frac{\overline{\theta}_{5.3} - \overline{\theta}_{5.5}}{\overline{\theta}_{3.2} - \overline{\theta}_{3.3}} \tag{3.5}$$

In den Gleichungen 3.1 bis 3.5 werden die Disutilities für die fünfstufige Fragebogenversion ($D_{5.1}$–$D_{5.5}$) auf der Basis der im Dolan-Tarif enthaltenen Disutility für die mittlere Kategorie des dreistufigen Fragebogens ($D_{3.2}$) und der Abstände zwischen den kategorienspezifischen Mittelwerten der geschätzten Personenparameter $\overline{\theta}_{j.k}$ für Item j und dessen Kategorie k berechnet.

2.2 Untersuchte Stichprobe

		5L				
	Mobilität	Keine Probleme	Leichte Probleme	Mäßige Probleme	Große Probleme	Nicht in der Lage
3L	Keine Probleme	905	66	9	1	3
	Einige Probleme	16	353	389	265	19
	Ans Bett gebunden	1	1	4	24	76
	Selbstpflege	Keine Probleme	Leichte Probleme	Mäßige Probleme	Große Probleme	Nicht in der Lage
3L	Keine Probleme	1347	52	10	5	0
	Einige Probleme	27	268	210	71	4
	Nicht in der Lage	1	3	5	27	96
	Allgemeine Tätigkeiten	Keine Probleme	Leichte Probleme	Mäßige Probleme	Große Probleme	Nicht in der Lage
3L	Keine Probleme	655	102	13	7	0
	Einige Probleme	23	416	421	191	12
	Nicht in der Lage	2	4	17	92	163
	Schmerzen/Beschwerden	Keine Schmerzen	Leichte Schmerzen	Mäßige Schmerzen	Starke Schmerzen	Extreme Schmerzen
3L	Keine Schmerzen	579	114	16	5	1
	Mäßige Schmerzen	33	501	532	155	5
	Extreme Schmerzen	0	2	10	109	58
	Angst/Niedergeschlagenheit	Nicht ängstlich	Ein wenig ängstlich	Mäßig ängstlich	Sehr ängstlich	Extrem ängstlich
3L	Nicht ängstlich	780	116	10	6	2
	Mäßig ängstlich	15	485	425	109	4
	Extrem ängstlich	1	2	11	106	53

Tabelle 1: Bivariate Kreuztabellen der Antwortmuster in den drei- und fünfstufigen Items des EQ-5D-Fragebogens

Die empirische Basis der in diesem Beitrag dargestellten Ergebnisse bildet eine nach einem gemeinsamen Studienprotokoll in sechs Ländern (Dänemark, England, Italien, Niederlande, Polen und Schottland) erfasste Patientenstichprobe (für ge-

nauere Angaben zur Stichprobenziehung vgl. van Hout et al. [2011]), in die Patienten aus mehreren Diagnosegruppen einbezogen wurden, darunter Patienten mit Erkrankungen der Atemwege, Stoffwechsel-, Leber- oder Nierenerkrankungen, Krankheiten des Herz-Kreislaufsystems sowie mit muskuloskelettalen, neurologischen oder psychischen Erkrankungen. Eine ebenfalls in dieser multinationalen Studie zu Vergleichszwecken erhobene Stichprobe von Studierenden wurde in die hier dargestellten Analysen nicht einbezogen.

Den Befragten in allen Ländern wurde ein einheitlicher Fragebogen vorgelegt, der sowohl die drei- als auch die fünfstufige Version des EQ-5D enthielt. Am Anfang stand die fünfstufige Version, der eine visuelle Analogskala zur Einschätzung des allgemeinen Gesundheitszustandes sowie einige soziodemographische Fragen folgten. Im letzten Teil war die dreistufige Variante des EQ-5D zu beantworten.

Um die methodische Güte mehrerer Ansätze zur Berechnung eines intermediären Tarifs besser vergleichen zu können, wurde der Gesamtdatensatz von n = 3.248 Fällen zufällig im Verhältnis von 2:1 in einen Lern- und einen Testdatensatz aufgeteilt. Die hier beschriebenen Ergebnisse stützen sich mit einer Ausnahme auf die mit dem größeren Lerndatensatz (n = 2.165) durchgeführten Analysen. Die bivariaten Kreuztabellen der fünf Itempaare sind in Tabelle 1 wiedergegeben.

	Gesamt-stichprobe	davon weiblich		Alter		Gesundheitszu-stand*)	
	n	n	%	Mittelwert	Streuung	Mittelwert	Streuung
Schottland	340	205	60,3	70,0	10,3	57,5	21,0
Polen	346	168	48,6	69,8	13,1	50,8	26,1
Nieder-lande	304	198	65,1	35,5	17,1	60,3	18,2
Dänemark	247	110	44,5	48,9	17,7	73,6	22,0
England	652	295	45,3	51,4	14,1	61,2	20,9
Italien	276	86	31,2	56,2	12,9	71,3	20,6
Total	2165	1062	49,1	55,3	18,3	61,5	22,6

*) gemessen auf einer Ratingskala von 0 (schlechtester Gesundheitszustand) bis 100 (bester Gesundheitszustand)

Tabelle 2: Fallzahlen, Geschlecht, Alter und Gesundheitszustand in der untersuchten Stichprobe

Tabelle 2 zeigt verschiedene Basismerkmale der untersuchten Stichprobe. Der Anteil weiblicher Patienten beträgt in der Gesamtstichprobe rund 50%, er variiert zwischen 31% (Italien) und 65% (Niederlande). Das Durchschnittsalter ist in den Teilstichproben aus Schottland und Polen am höchsten, da in diese überwiegend Patienten mit chronischen Krankheiten, die im höheren Lebensalter auftreten (z. B. Schlaganfall, kardiovaskuläre Erkrankungen), aufgenommen wurden. Am niedrigsten ist das Durchschnittsalter in der niederländischen Teilstichprobe, in der überwiegend Patienten mit psychischen Erkrankungen erfasst wurden. Zur Be-

schreibung des Gesundheitszustands zeigt die Tabelle 2 zusätzlich die Mittelwerte einer visuellen Analogskala – ähnlich einer Thermometerskala –, mit der die Patienten in einem Wertebereich von 0 (schlechteste Gesundheit) bis 100 (beste Gesundheit) ihren Gesundheitszustand einschätzen sollten. Die Mittelwerte der Angaben auf der visuellen Analogskala verdeutlichen, dass in die Gesamtstichprobe Patienten mit unterschiedlich gravierenden gesundheitlichen Beeinträchtigungen eingeschlossen waren. Während sich die Patienten aus der dänischen Teilstichprobe als eher geringgradig eingeschränkt darstellen (überwiegend Diabetes-Patienten), ist für die Teilstichprobe aus Polen eine erhebliche Beeinträchtigung des Gesundheitszustands zu verzeichnen (überwiegend Schlaganfall-Patienten).

3 Ergebnisse

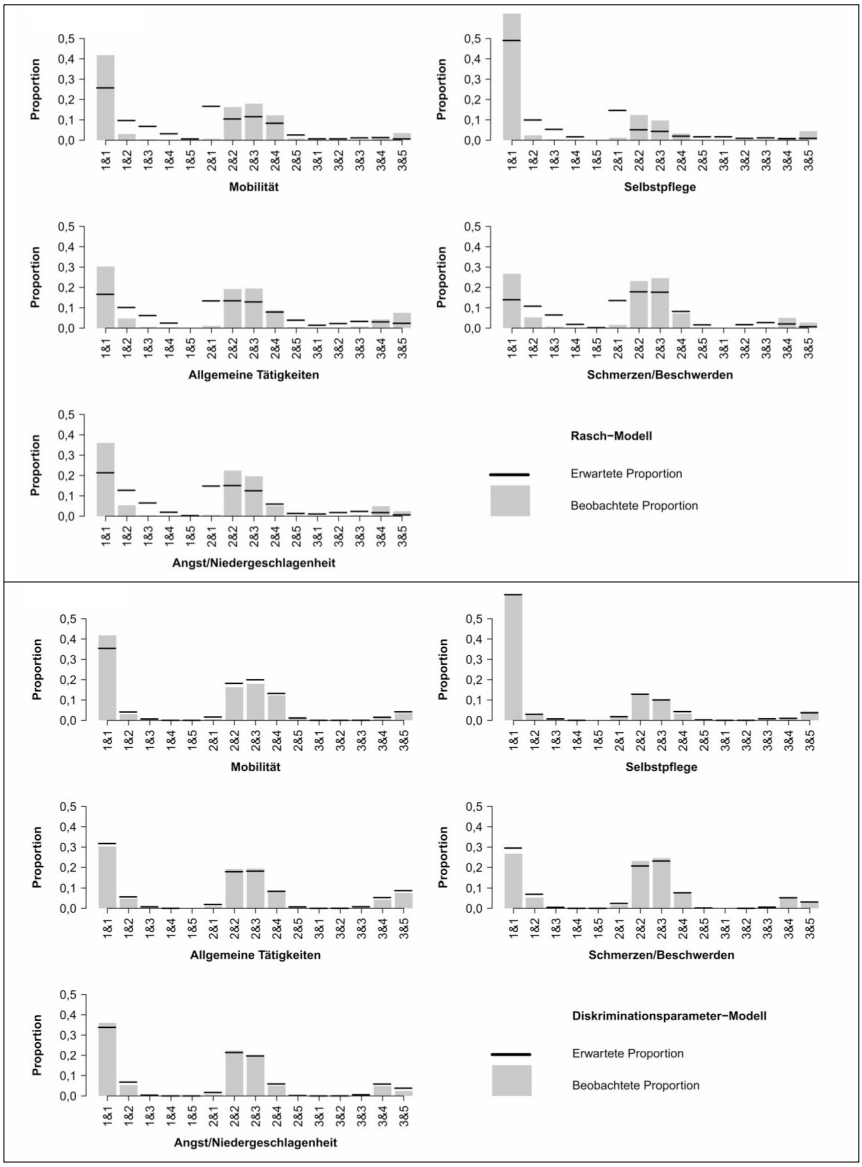
3.1 Modellgüte und geschätzte Parameter

	Rasch-Modell				
	Log-Likelihood	AIC	BIC	Dissimilarity Index	p-Wert*)
Mobilität	-4364,72	8741.44	8775.52	0.364	<0.001
Selbstpflege	-3583.45	7178.90	7212.98	0.322	<0.001
Allgemeine Tätigkeiten	-4798.40	9608.81	9642.89	0.347	<0.001
Schmerzen/Beschwerden	-4524.03	9060.05	9094.13	0.313	<0.001
Angst/Niedergeschlagenheit	-4450.05	8912.10	8946.18	0.357	<0.001
	Diskriminationsparameter-Modell				
	Log-Likelihood	AIC	BIC	Dissimilarity Index	p-Wert*)
Mobilität	-3610.22	7234.44	7274.20	0.077	<0.001
Selbstpflege	-2853.39	5720.78	5760.55	0.027	<0.001
Allgemeine Tätigkeiten	-4082.96	8179.92	8219.68	0.048	<0.001
Schmerzen/Beschwerden	-3915.37	7844.74	7884.50	0.056	<0.001
Angst/Niedergeschlagenheit	-3745.98	7505.97	7545.73	0.049	<0.001

*) auf der Grundlage eines parametrischen Bootstrap-Verfahrens für die Pearson-Chiquadrat-Prüfgröße, 1.000 Iterationen

Tabelle 3: Informationstheoretische, deskriptive und inferenzstatistische Fit-Maße für das Rasch-Modell und das Modell mit einem Diskriminationsparameter

Für jedes der fünf Itempaare wurden die Parameter für das Rasch-Modell sowie für das Modell mit einem Diskriminationsparameter geschätzt. Da bei den Berechnungen des generalisierten Partial-Credit-Modells häufig Konvergenzprobleme auftraten, wurde es nicht in die weiteren Analysen einbezogen. In Tabelle 3 und



Abbildungen 4a und 4b: Vergleich der beobachteten und unter dem Modell erwarteten relativen Häufigkeiten für die 15 möglichen Antwortmuster der drei- und fünfstufigen Versionen des EQ-5D-Fragebogens (Abb. 4a (oben): Rasch-Modell, Abb. 4b (unten): Diskriminationsparameter-Modell)

den Abbildungen 4a und 4b sind die Ergebnisse für die Prüfung der Modellanpassung für das Rasch- und das Diskriminationsparameter-Modell wiedergegeben. Es

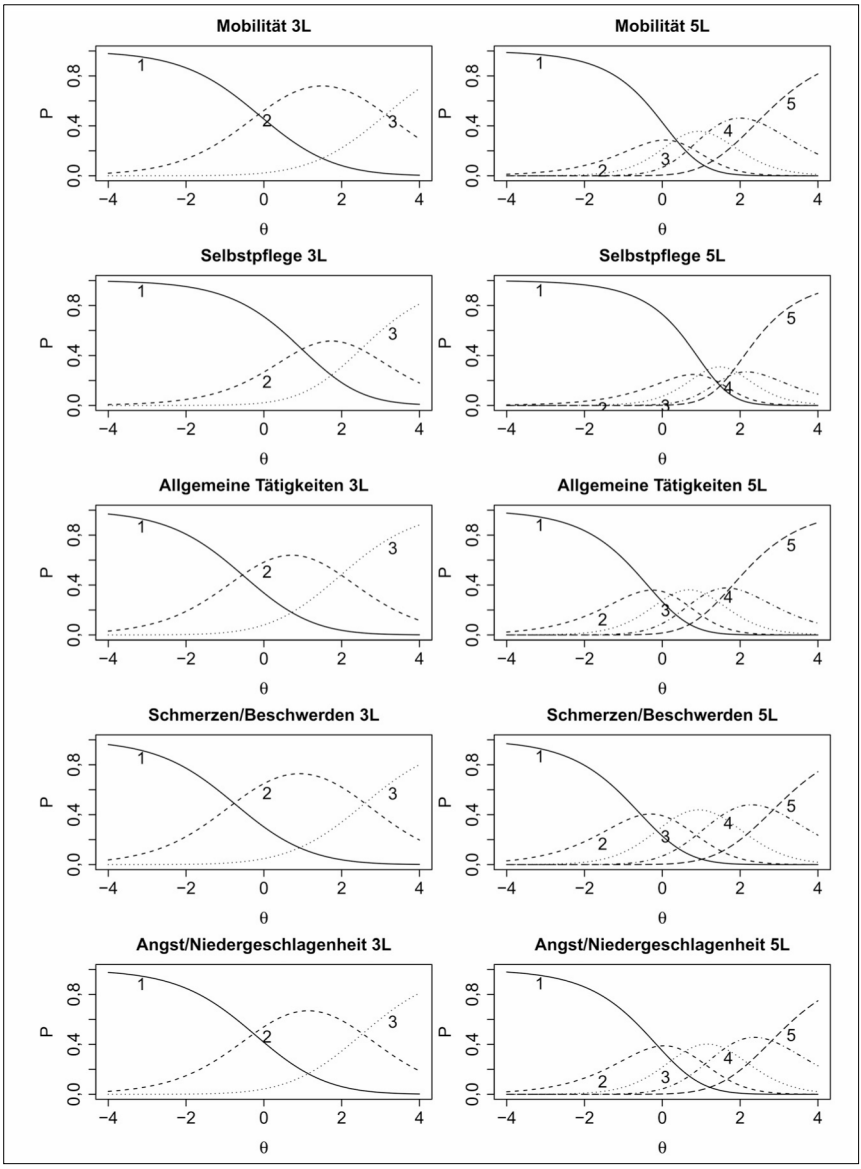


Abbildung 5a: Kategorienfunktionen der dreistufigen (3L) bzw. fünfstufigen (5L) Items des EQ-5D-Fragebogens: Rasch-Modell

zeigen sich bei den informationstheoretischen Gütekriterien AIC und BIC erhebliche Unterschiede zwischen den fünf Items des Fragebogens. Die beste Modellanpassung ergibt sich für das Item Selbstpflege, die schlechteste für allgemeine Tä-

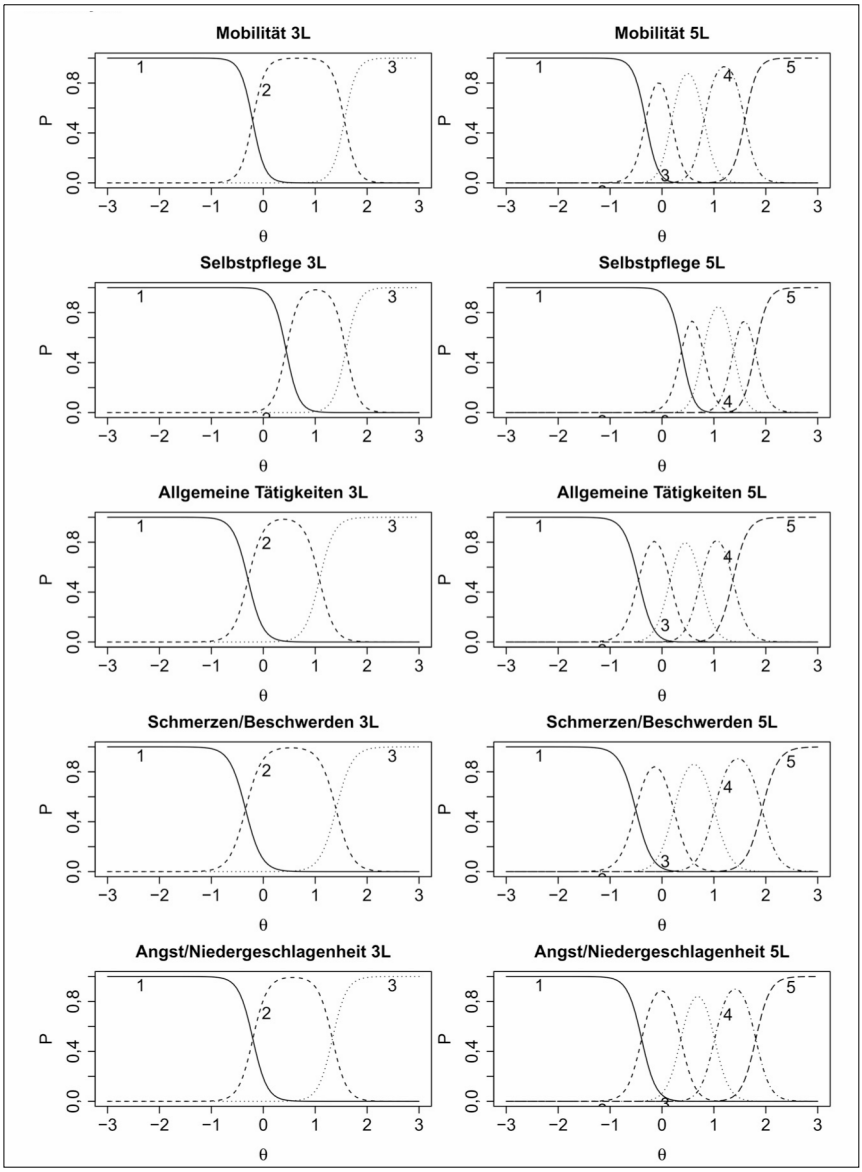


Abbildung 5b: Kategorienfunktionen der dreistufigen (3L) bzw. fünfstufigen (5L) Items des EQ-5D-Fragebogens: Diskriminationsparameter-Modell

tigkeiten. Ebenso deutlich ist der Gewinn an Modellanpassung beim Übergang vom Rasch-Modell auf das weniger restriktive Modell mit einem Diskriminationsparameter. Dieser Effekt ist im Dissimilarity-Index (Clogg 1995) und in der Gegenüber-

stellung der beobachteten und unter dem Modell erwarteten Häufigkeiten (vgl. Abb. 4a für das Rasch-Modell, Abb. 4b für das Diskriminationsparameter-Modell) besonders klar zu sehen. Die mit dem Bootstrap-Verfahren mit jeweils 1.000 Iterationen gewonnenen p-Werte zeigen allerdings, dass keines der geschätzten Modelle einen akzeptablen Fit erreicht. Es kann deshalb nur festgehalten werden, dass die Einführung eines einzigen weiteren (Diskriminations-)Parameters die Modellanpassung substanziell verbessert, ohne allerdings zu einem guten globalen Modellfit zu führen.

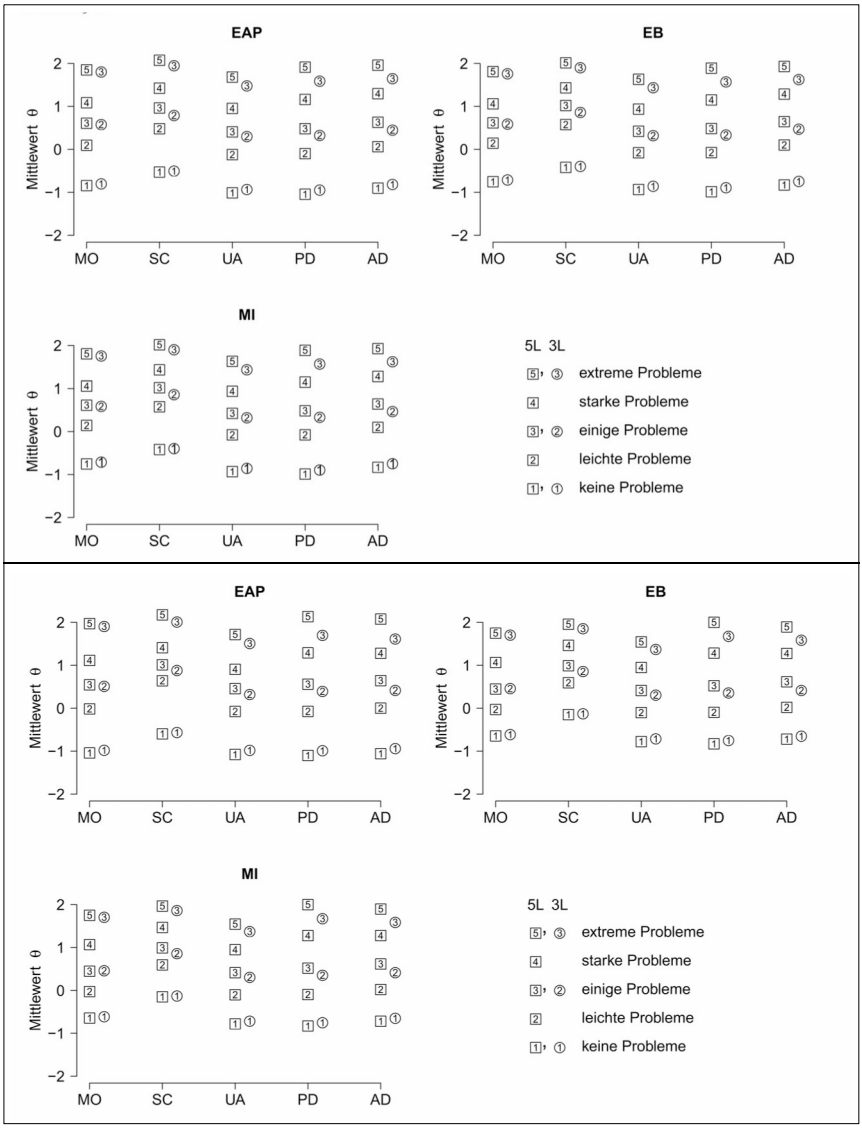
Die Kategorienfunktionen (vgl. Abb. 5a für das Rasch-Modell, Abb. 5b für das Diskriminationsparameter-Modell) zeigen für alle dreistufigen Items unabhängig vom Modell erwartungskonforme Verläufe. Bei den fünfstufigen Items ergeben sich jedoch für das Rasch-Modell beim Item Mobilität in der zweiten Antwortstufe und beim Item Selbstpflege in der zweiten und vierten Antwortstufe Überschneidungen der Kategorienfunktionen. Die Annahme geordneter Antwortkategorien scheint für diese beiden Items also nicht erfüllt zu sein. Der Übergang vom Rasch-Modell auf das weniger restriktive Diskriminationsparameter-Modell sorgt aber auch hier dafür, dass diese Inkonsistenzen aufgelöst werden und die Übergänge zwischen den Antwortkategorien den Erwartungen an eine Antwortskala mit geordneten Kategorien entsprechen (vgl. Abb. 5b).

3.2 Personenparameter und berechnete Disutilities

Die Abbildungen 6a und 6b zeigen, dass sich die item- und kategorienspezifischen Durchschnittswerte der nach den drei Methoden geschätzten Personenparameter nur unwesentlich unterscheiden. Als ebenso gering erscheinen die Unterschiede zwischen den Ergebnissen, die sich aus der Anwendung des Rasch- bzw. des Diskriminationsparameter-Modells ergeben haben.

Erwartungsgemäß weisen die kategorienspezifischen Mittelwerte der fünfstufigen Antwortskala eine größere Variationsbreite als die der dreistufigen Items auf. Die niedrigste ("keine Probleme") bzw. höchste Antwortkategorie ("extreme Probleme") der fünfstufigen Skala markiert einen besseren bzw. schlechteren Gesundheitszustand als die entsprechenden Randkategorien der dreistufigen. Die Mittelkategorien beider Antwortformate weisen geringe Abstände voneinander auf. Beim Vergleich der Ergebnisse des Rasch- bzw. des Diskriminationsparameter-Modells ist trotz der geringen Unterschiede von Interesse, dass die kategorienspezifischen Mittelwerte im Diskriminationsparameter-Modell für beide Antwortformate in den meisten Fällen nahezu äquidistant sind.

Tabelle 4 enthält eine Übersicht über die nach den Gleichungen 3.1 bis 3.5 berechneten Disutilities für die beiden Modellvarianten und die drei Methoden zur Schätzung der Personenparameter. In der ersten Spalte sind zum Vergleich die Disutilities für die dreistufige Antwortskala nach dem Dolan-Tarif wiedergegeben. Da die kategorienspezifischen Mittelwerte der Personenparameter sehr geringe Unterschiede aufweisen, ist auch hier nur mit unwesentlichen Unterschieden zwischen



Abbildungen 6a und 6b: Kategorienspezifische Mittelwerte der Personenparameter im Rasch-Modell (Abb. 6a, oben) und im Diskriminationsparameter-Modell (Abb. 6b, unten). Die Mittelwerte werden durch Quadrate (fünfstufige Items) bzw. Kreise (dreistufige Items) markiert. Die Zahlen in den Quadraten bzw. Kreisen bezeichnen die Ordnungsnummer der Antwortkategorien. Die Abbildungen zeigen die Mittelwerte für die fünf Items des EQ-5D-Fragebogens (MO = Mobilität, SC = Selbstpflege, UA = allgemeine Tätigkeiten, PD = Schmerzen/körperliche Beschwerden, AD = Angst/Niedergeschlagenheit) und für die drei unterschiedlichen Methoden zur Schätzung der Personenparameter (EAP = Expected a Posteriori, MI = Multiple Imputation, EB = Empirical Bayes)

				Modell und Schätzmethode für die Personenparameter					
	3L*)	Dolan-Tarif	5L*)	RM EAP	DP EAP	RM MI	DP MI	RM EB	DP EB
Mobilität			2	-0.047	-0.048	-0.048	-0.040	-0.047	-0.040
	2	-0.069	3	-0.073	-0.074	-0.073	-0.070	-0.073	-0.070
			4	-0.167	-0.172	-0.165	-0.193	-0.167	-0.193
	3	-0.314	5	-0.318	-0.321	-0.320	-0.328	-0.318	-0.328
Selbstpflege			2	-0.081	-0.089	-0.082	-0.078	-0.081	-0.078
	2	-0.104	3	-0.120	-0.116	-0.119	-0.120	-0.120	-0.120
			4	-0.155	-0.152	-0.155	-0.163	-0.155	-0.163
	3	-0.214	5	-0.205	-0.221	-0.205	-0.207	-0.205	-0.207
Allgemeine Tätigkeiten			2	-0.026	-0.028	-0.026	-0.024	-0.026	-0.024
	2	-0.036	3	-0.041	-0.042	-0.041	-0.042	-0.041	-0.042
			4	-0.067	-0.063	-0.067	-0.069	-0.068	-0.069
	3	-0.094	5	-0.102	-0.100	-0.102	-0.100	-0.103	-0.100
Schmerzen/ Beschwerden			2	-0.091	-0.091	-0.091	-0.082	-0.091	-0.082
	2	-0.123	3	-0.148	-0.148	-0.147	-0.150	-0.147	-0.150
			4	-0.285	-0.295	-0.285	-0.295	-0.286	-0.295
	3	-0.386	5	-0.437	-0.465	-0.438	-0.433	-0.441	-0.432
Angst/ Niedergeschlagenheit			2	-0.054	-0.056	-0.054	-0.049	-0.054	-0.049
	2	-0.071	3	-0.086	-0.089	-0.086	-0.089	-0.086	-0.089
			4	-0.174	-0.175	-0.174	-0.176	-0.175	-0.176
	3	-0.236	5	-0.262	-0.282	-0.263	-0.258	-0.263	-0.258

	Mittlere quadratische Abweichung der im Modell berechneten Nutzwerte von den Nutzwerten nach dem Dolan-Tarif												
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
	Lernstichprobe	0.04	0.09	0.05	0.09	0.04	0.08	0.05	0.09	0.04	0.09	0.05	0.09
	Teststichprobe	0.04	0.08	0.04	0.08	0.04	0.08	0.04	0.08	0.04	0.08	0.04	0.08

*) Kategoriennummern der drei- bzw. fünfstufigen Items

Tabelle 4: Disutilities im Dolan-Tarif und berechnete Disutilities für die fünfstufige Version des EQ-5D-Fragebogens nach Modell (RM = Rasch-Modell, DP = Diskriminationsparameter-Modell) und Schätzmethode für die Personenparameter (EAP = Expected a Priori, MI = Multiple Imputation, EB = Empirical Bayes). M = Mittelwert, SD = Standardabweichung

den Ergebnissen der Modelle bzw. Schätzmethoden zu rechnen. Wie zu sehen ist, entsprechen die Disutilities der Mittelkategorie im fünfstufigen Antwortformat sehr gut denen der dreistufigen. Mit wenigen Ausnahmen sind die Disutilities der höchsten fünfstufigen größer als die der höchsten dreistufigen Kategorie. Die im fünf-

stufigen System eingefügten Zwischenkategorien sind konsistent und fast gleichabständig zwischen den Werten der Rand- und Mittelkategorien lokalisiert.

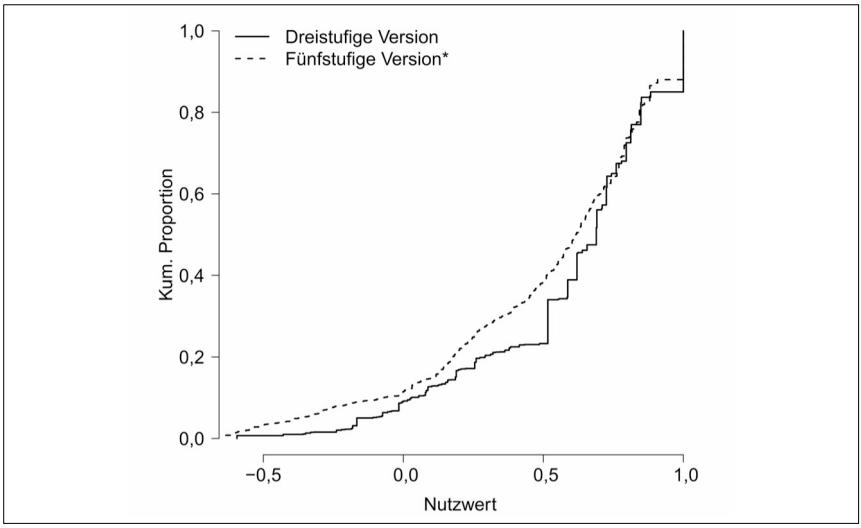
Im unteren Teil von Tabelle 4 sind die Mittelwerte der quadrierten Differenzen (mean squared error) zwischen den aus den fünfstufigen und aus den dreistufigen Disutilities berechneten Nutzwerten sowie die Standardabweichungen dieser Differenzen dargestellt. Dazu wurde zum einen die Lernstichprobe, zum anderen zu Validierungszwecken die Teststichprobe (n = 1.083) herangezogen. Auch beim Vergleich dieser Ergebnisse wird deutlich, dass zwischen den Modell- und Schätzvarianten nur sehr geringe Unterschiede bestehen.

Die große Übereinstimmung zwischen den Modellen auf der Ebene der Personenparameter und der berechneten Disutilities deutet darauf hin, dass trotz der manifesten Unterschiede z. B. im Modellfit bei Betrachtung höher aggregierter Maßzahlen, wie etwa der kategorienspezifischen Mittelwerte der Personenparameter, nur noch geringe Abweichungen der Ergebnisse in Abhängigkeit vom gewählten Modell auftreten. Diese hohe Konsistenz zeigt sich auch in den sehr hohen Korrelationen ($r > 0,95$) zwischen den mit den verschiedenen Modell- und Schätzvarianten berechneten Personenparametern.

3.3 Vorteile des intermediären Tarifs gegenüber der dreistufigen Antwortskalierung

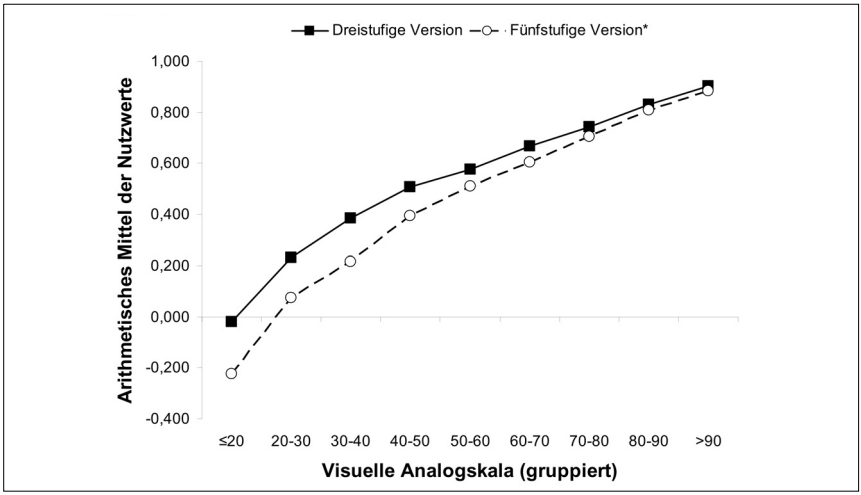
Der Entwicklung des fünfstufigen Antwortformats lag die Überlegung zugrunde, dass der bisherige, auf einer nur dreistufigen Antwortskala beruhende Fragebogen nicht ausreichend gut in der Lage ist, kleine, aber bedeutsame Unterschiede zwischen Gesundheitszuständen abzubilden. Darüber hinaus wurden bei der dreistufigen Variante häufig Deckeneffekte in dem Sinne beobachtet, dass ein besonders hoher Prozentsatz der Befragten über keinerlei Beeinträchtigungen des Gesundheitszustands berichtete. In diesem abschließenden Kapitel der Ergebnisdarstellung soll die Frage beantwortet werden, ob und wenn ja, in welchem Umfang der auf eine fünfstufige Antwortskala erweiterte Fragebogen in der Lage ist, diese Probleme abzuschwächen.

Zur Beurteilung der Verteilungseigenschaften der drei- und fünfstufigen Fragebogenversion sind in Abbildung 7 die kumulierten Häufigkeitsverteilungen der mit beiden Versionen erfassten Nutzwerte dargestellt. Während die beiden Verteilungsfunktionen im oberen Bereich der Nutzwertskala (0,70-1,00) kaum Unterschiede aufweisen, zeigt die Verteilungsfunktion der fünfstufigen Version in der unteren Skalenhälfte (< 0,50) eine deutlich größere Verteilungsmasse. Die fünfstufige Version verfügt damit bei ungünstigeren Gesundheitszuständen über einen stärkeren Differenzierungsgrad als die dreistufige. Wie die Verteilungsfunktionen im oberen Skalenbereich erkennen lassen, werden allerdings die Deckeneffekte nur unwesentlich verringert. Während bei der dreistufigen Variante 15 % der Stichprobe einen Nutzwert von 1,0 erzielt, vermindert sich dieser Deckeneffekt bei der fünfstufigen Version auf einen Anteil von nur 12%.



* Es wurde die Schätzung der Personenparameter nach dem Empirical-Bayes(EB)-Verfahren zugrunde gelegt.

Abbildung 7: Kumulierte Häufigkeitsverteilungen der Nutzwerte aus der dreistufigen Fragebogenversion (Dolan-Tarif) und dem berechneten intermediären Tarif für die fünfstufige Version



* Es wurde die Schätzung der Personenparameter nach dem Empirical-Bayes(EB)-Verfahren zugrunde gelegt.

Abbildung 8: Arithmetische Mittel der Nutzwerte aus der dreistufigen Fragebogenversion (Dolan-Tarif) und dem berechneten intermediären Tarif für die fünfstufige Version in Abhängigkeit von den Angaben zum Gesundheitszustand auf der visuellen Analogskala (0 = schlechteste Gesundheit, 100 = beste Gesundheit)

Die bessere Diskriminationsfähigkeit der fünfstufigen Version zeigt sich auch bei Betrachtung der Zusammenhänge der präferenzbasierten Nutzwerte mit der mit

Hilfe der visuellen Analogskala (VAS, schlechteste Gesundheit = 0, beste Gesundheit = 100) erfassten Einschätzung des Gesundheitszustands. Wie Abbildung 8 verdeutlicht, besteht – wiederum insbesondere bei schlechteren Gesundheitszuständen (VAS < 50) – ein stärkerer Abfall der Nutzwerte in Abhängigkeit von den Werten auf der visuellen Analogskala. Der Zugewinn ist aber auch hier nur gering, die Korrelation der visuellen Analogskala mit der dreistufigen Nutzwertskala beträgt r = 0,68, sie erhöht sich beim Übergang auf die fünfstufige Fragebogenversion nur sehr geringfügig auf r = 0,70.

Zugewinne in der Stärke von Zusammenhängen zeigen sich auch bei Betrachtung weiterer Kriterien, so zum Beispiel bei der Analyse von Unterschieden zwischen den einbezogenen Diagnosegruppen. Dort erhöht sich die durch die Gruppenzugehörigkeit erklärte Varianz von 16 % (dreistufige Version) auf 19 % (fünfstufige Version).

4 Resümee

In diesem Beitrag wurde untersucht, in welchem Umfang probabilistische Testmodelle, speziell das Rasch-Modell für mehrkategorielle Items und das weniger restriktive Diskriminationsparameter-Modell, zur Beantwortung einer sowohl methodisch als auch praktisch bedeutsamen Fragestellung geeignet sind. Es zeigte sich, dass die Anwendung dieser Modelle zu inhaltlich klaren und vor allem konsistenten Ergebnissen führt, die dazu beitragen, den Anwendern in der Praxis der gesundheitsökonomischen Evaluation bereits vor dem Abschluss großer Valuierungsstudien einen empirisch fundierten vorläufigen Tarif für die neuentwickelte fünfstufige Version des EQ-5D-Fragebogens zur Verfügung zu stellen.

Gegenüber anderen Verfahren der Bildung eines intermediären Tarifs, die z. B. auf linearen oder logistischen Regressionsmodellen basieren, hat die Anwendung von probabilistischen Testmodellen den Vorteil, dass diese auch Informationen über die Messcharakteristik der Items liefern. So zeigte sich, dass mit dem weniger strengen Diskriminationsparameter-Modell nicht nur eine substanzielle Verbesserung der Modellanpassung erzielt werden konnte, sondern dass sich auch erwartungskonforme Verläufe der Kategorienfunktionen ergaben.

Die hier dargestellten Ergebnisse korrespondieren sehr gut mit denen einer früheren Analyse, in der das Partial-Credit-Modell auf eine Vorläuferversion des fünfstufigen EQ-5D-Fragebogens angewandt wurde (Pickard et al. 2007). Diese Studie umfasste Daten von 423 Patienten mit einer Tumorerkrankung. Die in diesem Beitrag dargestellten Analysen stützen sich auf eine weitaus größere, multinationale Stichprobe. Neben der sehr guten Vergleichbarkeit der in diesen beiden Studien gewonnenen Ergebnisse ist die hohe Konsistenz der hier vorgelegten Ergebnisse besonders hervorzuheben. Die mit beiden Modellen erzielten Ergebnisse wiesen unabhängig von der Methode der Schätzung von Personenparametern eine große Übereinstimmung auf – ein Sachverhalt, der als deutlicher Hinweis auf die Belastbarkeit der Resultate gewertet werden kann.

Der Klarheit und der methodischen Strenge der Skalierungsergebnisse sind allerdings die nicht besonders eindrucksvollen Verbesserungen etwa im Hinblick auf die Verteilungseigenschaften und die Diskriminationsfähigkeit der fünfstufigen Fragebogenversion im Vergleich zur dreistufigen Originalfassung gegenüberzustellen. Es ist noch unklar, ob die neue Version tatsächlich die mit ihrer Entwicklung intendierte größere Trennschärfe und Änderungssensitivität erreicht. Dies muss sich erst in konkreten Anwendungsstudien zeigen. Für die medizinische Forschung ist jedoch nicht auszuschließen, dass selbst geringfügige Verbesserungen der methodischen Eigenschaften des Befragungsinstruments zu einer substanziell besseren Interpretierbarkeit der Messergebnisse führen.

Die in diesem Beitrag dargestellte Anwendung probabilistischer Testmodelle beschränkte sich auf die einfache Analyse von Itempaaren mit gleichem Inhalt, aber unterschiedlicher Antwortskalierung. Hierzu wurden getrennte Auswertungen für jede der fünf Dimensionen durchgeführt. Weiterführende Analysen könnten einen noch umfassenderen Ansatz verfolgen und die Mehrdimensionalität der Items explizit in die Modellierung einbeziehen (von Davier & Carstensen 2007).

In Kapitel 1 wurde bereits darauf hingewiesen, dass probabilistische Testmodelle in der Gesundheitsforschung der letzten Jahre sehr große Beachtung gefunden haben. In der psychologischen Methodenlehre jedoch waren sie bereits seit langer Zeit bekannt und in deutschsprachigen Publikationen zugänglich (Fischer 1974, Rost 1988b). Weiterentwicklungen und Anwendungen dieser Modelle konzentrierten sich zunächst mit wenigen Ausnahmen auf das Arbeitsgebiet der Sozialwissenschaften (Rost & Langeheine 1988, Langeheine & Rost 1997), ihre Vermittlung in der Lehre war noch weitgehend auf die psychologische Testtheorie und Testkonstruktion (Rost 1996) beschränkt. Obwohl eines der führenden Lehrbücher zur Entwicklung von Fragebögen in der Medizin (Streiner & Norman 1989) bereits in der ersten Auflage ein kurzes Kapitel *Latent-Trait Theory* enthielt (in der zweiten Auflage wurde dieses Kapitel in *Item Response Theory* umbenannt), fanden diese Testmodelle erst im Zuge der rapiden Entwicklung der Lebensqualitätsforschung in der Medizin ein breiteres Anwendungsfeld.

Die heutige Situation verdeutlicht eine systematische Übersichtsarbeit, in der bei einer Recherche in medizinischen Literaturdatenbanken (Medline, CINAHL, PEDro, EMBASE) über 200 wissenschaftliche Arbeiten identifiziert werden konnten, in denen das Rasch-Modell bei der Neu- oder Weiterentwicklung von Erhebungsinstrumenten angewandt wurde (Belvedere et al. 2010). Sowohl im Bereich der klinischen Assessments allgemein (Thomas 2010) als auch speziell bei der Entwicklung und Anwendung patientennaher Erhebungsinstrumente zur Messung der subjektiven Gesundheit und der gesundheitsbezogenen Lebensqualität (Hays et al. 2000) wird probabilistischen Modellen ein hohes Potenzial zugeschrieben. Dies ist sicherlich in weiten Teilen auf ihre besonders gute Eignung zur Analyse von Selbstbeurteilungsdaten und auf ihre messtheoretischen Eigenschaften zurückzuführen. Es ist zu erwarten, dass die durch diese Modelle mögliche methodische Absicherung der Erhebungsverfahren deren Stellenwert in der medizinischen und gesundheitswissenschaftlichen Forschung zukünftig erhöht und so die häufig geforderte stärkere Einbeziehung der Sichtweise von Patienten in der Medizin weiter fördert.

Was ist aus dem Integrierten Handlungsmodell geworden?

Thomas Martens

1 Das Integrierte Handlungsmodell

Mitte der 90er-Jahre sind Jürgen Rost und ich mit dem Anspruch angetreten, ein integratives Handlungsmodell zu validieren. Insbesondere sollten mit einer Bedingungsanalyse von kognitiven, affektiven und sozialen Faktoren Ansatzpunkte für die Förderung umweltgerechten Verhaltens identifiziert werden. Dieser Anspruch konnte in den nachfolgenden sechs Jahren durch insgesamt drei DFG-Projekte in wesentlichen Punkten nachgewiesen und erfüllt werden. In diesem Beitrag soll das Integrierte Handlungsmodell reflektiert, erweitert und auf Lernprozesse übertragen sowie beispielhaft empirisch überprüft werden.

Als das Integrierte Handlungsmodell entwickelt wurde (Martens 2000, Martens et al. 2008, Martens et al. 2009, Martens & Rost 1998, Rost 1996a), lagen die verschiedenen Theoriekomponenten eines möglichen Handlungsmodells noch unverbunden nebeneinander – die Integrationsleistung war namengebend. Es existierten Handlungsmodelle, die

- den normativen Charakter einer Handlung betont haben (Schwartz 1973, Schwartz & Howard 1981),
- das zu erwartende Handlungsergebnis in den Vordergrund gestellt haben (Heckhausen 1989) und
- die Handlungsausführung betrachtet haben (Achtziger & Gollwitzer 2007, Gollwitzer 1990).

Außerdem lagen kombinierte Modelle vor, die bereits zwei von drei Handlungsphasen integrieren konnten, etwa die Schutzmotivationstheorie (Rippetoe & Rogers 1987, Rogers 1983) oder das sozial-kognitive Prozessmodell gesundheitlichen Handelns (Schwarzer 1994).

Das Integrierte hat als erstes Modell alle drei Phasen (Motivierung, Intention und Volition) einbezogen (vgl. Abb. 1).

Eine zweite wichtige modelltheoretische Entscheidung für die Entwicklung des Integrierten Handlungsmodells war die Berücksichtigung von Prozessen. Sehr viele Modelle zur Vorhersage des Handelns und insbesondere des Umwelthandelns orientieren sich an der methodischen Logik von einfachen Regressions- oder Strukturgleichungsmodellen, um eine maximale Varianzaufklärung zu erzielen (etwa Bamberg 1996). Das Kriterium der Varianzaufklärung kann allerdings nicht das ein-

zige für eine theoretische Entwicklung sein (vgl. Martens 2000). Das Verständnis von Handlungsprozessen ist etwa die Voraussetzung für die Ableitung von pädagogischen Maßnahmen[1].

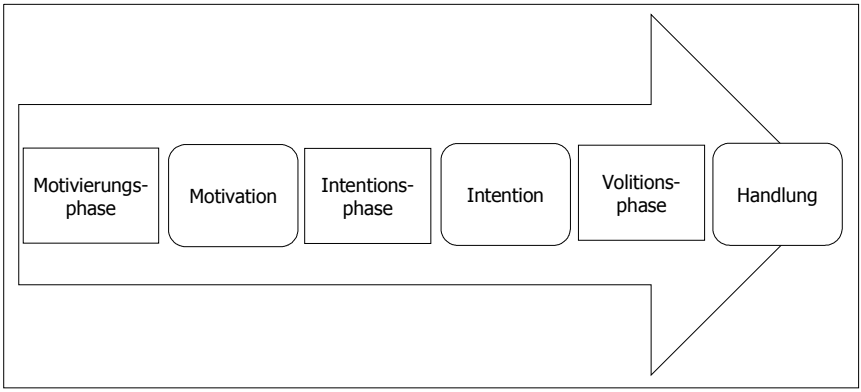


Abbildung 1: Die drei Hauptphasen des Integrierten Handlungsmodells: Motivierungsphase, Intentionsphase und Volitionsphase

Darüber hinaus erbringt das Handlungsmodell von Martens & Rost (1998) zwei weitere Integrationsleistungen: Es löst die Dichotomien zwischen unbewussten und bewussten sowie zwischen affektiven und kognitiven Prozessen auf. Während in der Vergangenheit unbewusste Prozesse in Handlungstheorien vor allem mit der Automatisierung von Handlungen (etwa Klöckner & Matthies 2004) in Verbindung gebracht wurden, so haben vor allem neurowissenschaftliche Verfahren neue Perspektiven auf unbewusste Prozesse der Handlungsgenese eröffnet (Kuhl 2001). Insbesondere den von Kuhl postulierten rechtshemisphärischen Makrosystemen, dem Extensionsgedächtnis und der Intuitiven Verhaltenssteuerung werden Funktionsprofile zugeschrieben, die sich auf ganzheitlich-assoziative Prozesse beziehen und bei denen eine bewusste Fokussierung stören würde[2]. Das Integrierte Handlungsmodell weist die Prozesse aus, die einen holistischen Abgleich mit Selbstschemata nötig machen[3]; insbesondere die Verantwortungsübernahme in der Motivierungsphase, die Selbstwirksamkeit in der Intentionsphase sowie die selbstkongruente Zielverfolgung in der Volitionsphase sind so konzipiert, dass sie mit einer Aktivierung des Makrosystems Extensionsgedächtnis einhergehen. Ein gelingender holistischer Abgleich mit den Selbstschemata verläuft hierbei im Wortsinn nicht unbedingt *un*bewusst, aber mindestens *vor*bewusst. Es kann einem bewusst sein, dass entsprechende Prozesse eingeleitet werden müssen[4].

1 Wenn der Mensch mehr sein soll als ein Verhaltensautomat im Sinne des Behaviorismus.
2 Dabei scheint es ein informationslogischer Zusammenhang zu sein, dass die holistische, assoziative und vor allem parallele Informationsverarbeitung *nicht* bewusst sein kann (vgl. Velmans 1991).
3 Diese Prozesse verlaufen parallel, damit eine hohe Geschwindigkeit des Abgleichs erreicht werden kann. Es kommt dabei eher auf einen summarischen Gesamtabgleich an als auf Detailgenauigkeit. Es muss nicht geprüft werden, ob jedes Subschema passt oder nicht.
4 Wichtig ist es, Zeit und Gelegenheit zu finden, um negative Affekte zu dämpfen, etwa durch systematische Entspannungspausen.

Die zweite Dichotomie, die mit dem Integrierten Handlungsmodell überwunden werden kann, ist die zwischen Kognition und Affekt. Die Konzeption von Kuhl (ebd.) zeigt eine funktionslogische Perspektive auf, wie die Aktivierung von Makrosystemen mit der Modulation von Affekten in Beziehung zur Genese von Handlungen steht. Das scheinbare Primat des bewussten Kognitivismus (Forgas et al. 2006, vgl. Rausch 2012) wird zugunsten einer durch Affekte vermittelten Prozessregulation aufgelöst. Die genauen Mechanismen werden in den nachfolgenden Kapiteln dargestellt. Generell ist die Verstärkung von negativen Affekten und die Abschwächung von positiven mit fokussierenden, linkshemisphärischen Prozessen verbunden: Die Verstärkung von negativen Affekten aktiviert das Makrosystem Objekterkennung und die Abschwächung von positiven Affekten aktiviert das Makrosystem Intentionsgedächtnis. Im ersten Fall wird die Aufmerksamkeit auf die Unstimmigkeiten (Objekterkennung), im zweiten auf das Ziel (Intentionsgedächtnis) fokussiert. Gedanken und Gefühle bilden nach Kuhl (2001) also einen funktionslogischen Zusammenhang für die Handlungsregulation[5]. Die Beteiligung der vier Makrosysteme an den in den Handlungsphasen ablaufenden Prozessen wird in Kapitel 3 detailliert aufgezeigt.

Viele Autoren, die insbesondere aus der Arbeitspsychologie kommen (vgl. Rausch et al. 2010), schreiben handlungstheoretischen Ansätzen das integrierende Potenzial einer allgemeinen Theorie (grand unifying theory) zu (Dörner 1985a, Frese & Zapf 1994, Rubinstein 1973/1981, Schurer 1984, Sonntag 2007). Die Wurzeln der Handlungstheorie können auf die Arbeiten von Lewin (1926a, 1926b), Miller (Miller et al. 1960/1973) sowie die von sowjetischen Forschern wie Rubinstein, Vygotzki, Galperin Leont'ev, Oschanin und anderen geprägte Tätigkeitspsychologie zurückgeführt werden (Frese & Zapf 1994, Rausch et al. 2010).

In dieser Tradition soll gezeigt werden, dass das Integrierte Handlungsmodell (Martens & Rost 1998) in sehr verschiedenen Kontexten angewendet werden kann, insbesondere zur Abbildung von Lernprozessen. Theoretisch folge ich dem Ansatz von Julius Kuhl (2000), der die PSI-Theorie auf hypothetische Lernprozesse der Volitionsphase anwendet, und erweitere diesen um die Motivierungs- und die Intentionsphase.

Einige der im Folgenden beschriebenen Prozesse sind empirisch gut belegt, insbesondere die der Volitionsphase, andere hingegen – etwa in der Motivierungsphase – nicht. Ein Beispiel für eine empirische Prüfung wird in Kapitel 4 dargestellt.

2 Ist es möglich, Lernen als Handlungsprozess zu verstehen?

Lernen und Lernprozesse werden besonders gern als Selbstregulationsprozesse konzipiert (vgl. etwa Zimmerman 1998), die in der Regel aus Handlungstheorien

5 Das Primat des logischen Denkens in der bisherigen psychologischen Handlungsforschung ist vermutlich auch der engen Verknüpfung von linkshemisphärischen Denk- und Sprachproduktionsprozessen geschuldet; das griechische Wort lógos spiegelt die semantische Vermischung von Sprache und Denken wider.

abgeleitet werden, wobei explizit auf die metakognitiven Prozesse der Volitionsphase referenziert wird (Landmann et al. 2009, Landmann & Schmitz 2007). Ein wenig verwunderlich ist es, dass der explizite Rückbezug auf handlungstheoretische Modelle in vielen Arbeiten nur am Rande thematisiert wird. Auch wenn in der Selbstregulationsliteratur die Gleichsetzung von Lernen und Handlung als weitgehend akzeptiert gilt (vgl. Landmann et al. 2009), soll diese Annahme im Folgenden noch einmal reflektiert und begründet werden.

Die Gleichsetzung von Lernen und Handlung wird von Dulisch (1986) weiter differenziert. Er unterscheidet

1. zielorientierte Handlungen, die als explizite Lernhandlungen vorgesehen sind, sowie
2. Aufbau und gedächtnismäßige Verfestigung von Handlungsdispositionen bzw. -komponenten.

Insbesondere die unter (1) definierte Lernhandlung kann als zielgerichtet im Sinne des Integrierten Handlungsmodells verstanden werden, das es erlaubt, eine zielgerichtete Lernhandlung als Erfüllung einer spezifischen Lernmotivation zu deuten. Handlungen ohne eine vorher spezifisch gebildete Motivation sind Spezialfälle, die später behandelt werden.

Die unter (2) definierte Verfestigung von Handlungsdispositionen kann als Vorbereitung für den ersten Typ einer Lernhandlung gelten und geht davon aus, dass es kontextfreie Handlungsdispositionen gibt, die erst später in einem größeren Handlungszyklus kontextualisiert werden.

Dieser zweite Handlungstyp erscheint aus motivationspsychologischer Sicht hoch problematisch. In der kindlichen (vgl. Gudjons 2008) und phylogenetischen (vgl. Klix 1980) Entwicklung scheint es ein natürlicher Prozess zu sein, in ganzheitlichen Handlungszyklen zu agieren, die, wie bereits dargestellt, unbewusst oder bewusst verkürzt sein können. Im Gegensatz hierzu stehen Handlungsdispositionen, die als isoliert und dekontextualisiert definiert werden. Sicherlich wäre es für einen sehr fortgeschrittenen Lerner möglich, ein hierarchisches Lernsystem selber zu definieren[6]. In der Regel aber wird eine Lehrperson die entsprechende Aneignung von Handlungsdispositionen vorgeben. Ein solches Lernen verleitet zu sehr kleinen und voneinander isolierten Lernschritten und nachfolgend zu einer verstärkten Anwendung von behavioristischen Lernprinzipien, etwa der Wiederholung. Eine Dekontextualisierung von Lerninhalten stellt deshalb eine große motivationale Herausforderung dar (Brown et al. 1989, Mandl & Kopp 2006), und die Anwendung der so gelernten Handlungsdispositionen in neuen Kontexten kann äußerst schwierig sein (vgl. etwa Gruber & Renkl 2000).

Es wird also im Folgenden davon ausgegangen, dass das Lernen vom Typ 2, insbesondere wenn die Gedächtnisinhalte von außen definiert und in einer sehr feinen

6 Hacker (1978) hat ein Modell der hierarchisch-sequenziellen Handlungsregulation entwickelt (vgl. Rausch 2012).

Granularität definiert sind, nur in begründeten Ausnahmefällen sinnvoll ist. Dies mag sogar dann gelten, wenn ein sehr kompetenter Lerner seine eigenen Lernpläne als hierarchische Lernhandlungen strukturiert. Wir folgen also den oft unausgesprochenen Annahmen der Selbstregulationsliteratur (vgl. Landmann et al. 2009) und begreifen Lernen als Lerntätigkeit im Sinne einer Lernhandlung.

3 Das Integrierte Handlungsmodell als Lernmodell

Gemäß dem Integrierten Handlungsmodell werden drei Prozessphasen einer vollständigen Lernhandlung unterschieden, die für tiefgreifende Lernprozesse notwendig sind:

- die Motivierungsphase,
- die Intentionsphase und
- die Volitionsphase.

In der Motivierungsphase wird eine Lernmotivation ausgebildet, d. h. es entsteht ein Bedürfnis, eine lernbezogene Soll-Ist-Diskrepanz zu reduzieren.

In der Intentionsphase wird eine Lernintention gebildet, die die Lernmotivation erfüllen kann.

In der Volitionsphase wird schließlich eine Lernintention in eine tatsächliche Lernhandlung umgesetzt (Hacker 1978).

3.1 Die Motivierungsphase

3.1.1 Soll-Ist-Diskrepanz

Am Anfang der Motivierungsphase (vgl. Abb. 2) steht die Wahrnehmung einer lernbezogenen Soll-Ist-Diskrepanz (vgl. Hofer 1981)[7], d. h. was kann ausgehend von der aktuellen Situation ohne eingreifendes Handeln[8] für eine zukünftige Situation erwartet werden? Diese Diskrepanz kann als Herausforderung oder als Bedrohung wahrgenommen werden (Lazarus & Launier 1981) und wird zunächst als Bedrohung diskutiert.

In der Umwelt- und Sozialpsychologie gibt es verschiedene Konzeptualisierungen von Bedrohung (etwa Rippetoe & Rogers 1987, Rogers 1983). So wird z. B. zwischen Schadenshöhe und der Eintrittswahrscheinlichkeit des Schadens unterschieden. Kuhl (2001) folgend, funktioniert die Wahrnehmung einer möglichen Bedrohung über die Aktivierung des Makrosystems Objekterkennung. Diese Aktivierung ist mit einem negativen Affekt verbunden, der die hemmende Wirkung des Exten-

7 Diese Diskrepanz kann als Interaktion zwischen Person und Situation verstanden werden (vgl. Grundmodell der klassischen Motivationspsychologie, Rheinberg 2008)

8 Dieser zukünftige Zustand kann auch als die Fortführung bestehender Handlungen und Handlungsmuster verstanden werden. In dieser Logik wird die bewusste Unterlassung einer nicht-funktionalen Handlung als eingreifendes Handeln verstanden (Martens 2000).

sionsgedächtnis auf die Objekterkennung aufhebt. Dies bedeutet unter anderem, dass der negative Affekt eine Zeit lang ausgehalten werden muss, um eine angemessene Risikoanalyse zu ermöglichen.

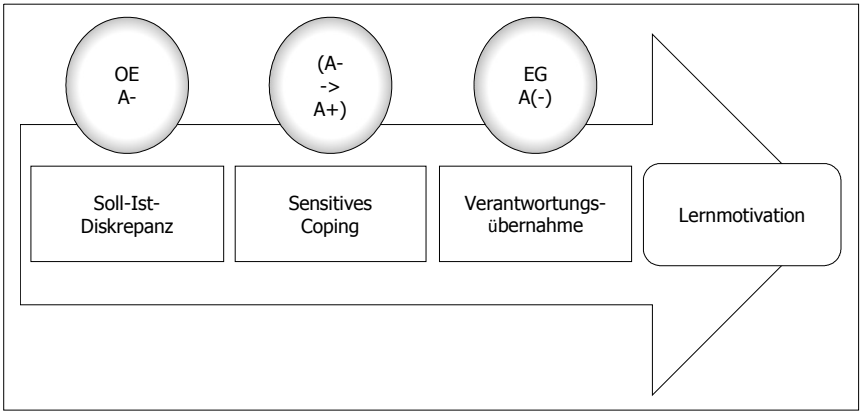


Abbildung 2: Die Motivierungsphase im Integrierten Handlungsmodell und die beteiligten Makrosysteme und Affekte. OE = Objekterkennung, EG = Extensionsgedächtnis, A- = negativer Affekt, A+ = positiver Affekt, A(-) = Dämpfung von negativem Affekt

3.1.2 Sensitives Coping

Die Genese der Motivation kann insbesondere durch Verdrängungsprozesse unterbrochen werden, etwa durch die Umwandlung von negativen in positive Affekte. Solche Verdrängungsprozesse sind theoretisch lange erwartet worden (vgl. etwa Krone 1974), konnten aber erst in jüngerer Zeit mit Hilfe von EEG-Ableitungen nachgewiesen werden (Sander et al. 2003). Sie können potenziell alle möglichen bedrohlichen Einzelinformationen betreffen, bestimmte Themenbereiche (siehe Martens 2000), oder sie können sich auch als generalisierter Persönlichkeitsstil manifestieren, etwa als Represser vs. Sensitizer sensu Krone (1974). Die energetisierende Wirkung einer als bedrohlich wahrgenommenen Soll-Ist-Abweichung kann also nur wirksam werden, wenn Verdrängungsprozesse dies *nicht* verhindern. In Abbildung 2 ist ein sensitives Coping durch (A- -> A+) symbolisiert, d. h. die automatische Umwandlung von einem negativen in einen positiven Affekt wird gedämpft.

3.1.3 Verantwortungsübernahme

Entscheidend für die Ausbildung einer tragfähigen Lernmotivation ist die Verantwortungsübernahme. Hiermit ist insbesondere die Bereitschaft gemeint, persönliche Verantwortung für die Lösung der wahrgenommenen Soll-Ist-Diskrepanz zu übernehmen.

Die Verantwortungsübernahme wird u. a. in der Sozialpsychologie (Schwartz 1977, Schwartz & Howard 1981; Stern et al. 1993) umfassend untersucht. Unter dem spezifischen Blickpunkt der Genese einer Lernhandlung kann wiederum der Rück-

bezug auf die von Kuhl postulierten Makrosysteme einen wesentlichen Beitrag zum Verständnis der Verantwortungsübernahme liefern: So kann diese als Abgleich mit dem Extensionsgedächtnis verstanden werden, d. h. als Passung zwischen Anforderung von außen und den unterschiedlichen Instanzen des Selbst. Sie ist umso tiefer verankert, je mehr es gelingt, Anknüpfungspunkte zu den Selbstschemata zu finden. Die erfolgreiche Verantwortungsübernahme führt zur Ausbildung eines Lern- oder Handlungsmotivs.[9]

Im Rahmen dieser theoretischen Konzeption kann angenommen werden, dass eine Soll-Ist-Diskrepanz dann als unmittelbare Herausforderung wahrgenommen wird, wenn keine Verdrängungsprozesse (sensitives Coping) stattfinden und eine direkte Verantwortungsübernahme erfolgt. Die bei der Verantwortungsübernahme ablaufenden Prozesse, von denen hier angenommen wird, dass sie vor allem über das Extensionsgedächtnis vermittelt werden, können sehr schnell und vor allem weitgehend unbewusst stattfinden, so dass das subjektive Gefühl einer unmittelbaren Herausforderung entsteht. Es kann also davon ausgegangen werden, dass unmittelbare Herausforderungen vor allem dann wahrgenommen werden, wenn die Art der Soll-Ist-Diskrepanz schon vorher bekannt war, etwa bei der Wiederholung einer Lernaufgabe in einem gewohnten Lernraum. Der unmittelbare Herausforderungscharakter einer Soll-Ist-Diskrepanz kann allerdings auch unerwünschte Nebenwirkungen haben – wenn etwa die Prozesse einer tiefgreifenden Verantwortungsübernahme einfach übergangen werden. Dies kann zu einer schwachen oder zu einer nicht erfüllbaren Motivation führen.

Die Stärke der Lern- oder Handlungsmotivation kann unterschiedlich sein. Je stärker sie ausgeprägt ist, desto größer wird die Energetisierung des nachfolgenden Handlungsprozesses sein. Diese Energetisierung kann sich sowohl in der zeitlichen Persistenz der Motivation zeigen als auch in der Motivationsstärke, die Voraussetzung für die Durchführung besonders schwieriger Handlungen ist (vgl. Martens 2000)[10].

3.2 Intentionsphase

3.2.1 Handlungssuche

Die Intentionsphase (vgl. Abb. 3) beginnt mit der Suche nach einer geeigneten Lernhandlung, wobei entweder auf den eigenen Erfahrungsschatz zurückgegriffen, die Lernhandlung durch die Lernumgebung impliziert oder durch die Lehrperson vorgeschlagen wird.

Die Suche nach geeigneten Handlungen gewinnt vor allem nach möglichen Revisionsprozessen, wenn z. B. die intendierte Handlung in der Volitionsphase nicht

9 Verschiedene Lernmotive können sich zu einem interessenthematischen Person-Gegenstands-Bezug verdichten (Krapp 1999).

10 Es kann auch zu einer zu starken Dominanz eines Motivs kommen – wenn etwa immer wieder vergeblich versucht wird, entsprechende Lernhandlungen zu finden. Eine übermächtige Motivation kann Ersatzhandlungen auslösen, die die eigentliche Soll-Ist-Diskrepanz nicht reduzieren können.

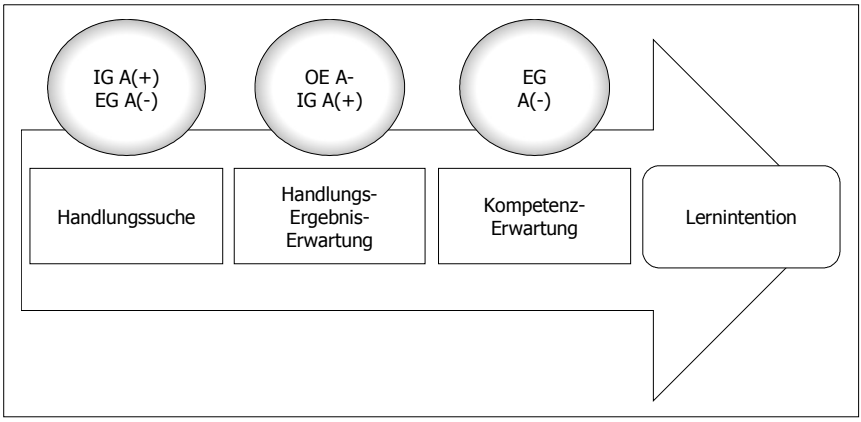


Abbildung 3: Die Intentionsphase im Integrierten Handlungsmodell und die beteiligten Makrosysteme und Affekte. IG = Intentionsgedächtnis, EG = Extensionsgedächtnis, OE = Objekterkennung, A(+) = Dämpfung von positivem Affekt, A(-) = Dämpfung von negativem Affekt, A- = negativer Affekt, A+ = positiver Affekt

umgesetzt werden konnte, eine steigende Bedeutung. Nun müssen alternative Handlungen identifiziert werden, die ebenfalls eine Erfüllung der Lernmotivation versprechen. Gerade bei einer sehr anspruchsvollen Motivation ist es nötig, neue Handlungen zu finden, die ggf. über den bereits bekannten Handlungsraum hinausgehen. Für eine solch kreative Suche ist der assoziative Zugang zum Extensionsgedächtnis unbedingt hilfreich. Prozesse, die diesen Zugang hemmen, führen umgekehrt zu einer Verengung des Handlungsraumes. Es besteht also die Gefahr, dass die Handlungsintention – trotz Misserfolgsinformation – beibehalten wird und wider besseres Wissen die gleiche Handlung noch einmal intendiert wird.

3.2.2 Handlungs-Ergebnis-Erwartung

Für die Ausprägung einer spezifischen Lernintention muss geprüft werden, ob die Lernhandlung das Potenzial hat, die Lernmotivation wirklich zu befriedigen, d. h. ob erwartet werden kann, dass sie eine Soll-Ist-Diskrepanz reduziert. Solche kognitiven Inferenzen müssen sicherlich mit Beteiligung der Objekterkennung und des Intentionsgedächtnisses durchgeführt werden.

3.2.3 Kompetenzerwartung

Neben den kognitivistischen Erwartungen, ob eine Handlung erfolgreich sein kann, ist der zentrale Prozess der Intentionsphase die Selbstwirksamkeit sensu Bandura (1977). Die von ihm konzipierte Erwartung, ob eine gewünschte Handlung aufgrund der eigenen Kompetenzen erfolgreich ausgeführt werden kann, soll an dieser Stelle etwas erweitert werden. Die Selbstwirksamkeit kann als Abgleichprozess verstanden werden, welche Handlungsalternative besonders gut zum Selbst passt. Diese Prüfung der Kongruenz zwischen Handlung und verschiedenen Aspekten des Selbst ist nur teilweise bewusst und verlangt die starke Beteiligung des Extensi-

onsgedächtnisses. Kuhl (2000) beschreibt den Vorgang als self compatibility check. Es kann erwartet werden, dass dieser Selbstkongruenzprozess unmittelbar vom psychischen Bedürfnis nach Autonomie beeinflusst wird, das in der Selbstbestimmungstheorie von Deci & Ryan (1993) eine entscheidende Rolle spielt. Nur wenn dem Lernenden ein autonomer Entscheidungsspielraum eingeräumt wird, kann der erforderliche Prozess eingeleitet werden, der dann eine doppelte Funktion hat:

1. Zum einen trägt dies zu einer Verstärkung der Intention bei, die sowohl in der Durchsetzung gegenüber anderen Handlungsalternativen als auch in der weiteren Abschirmung gegenüber Ablenkungen wirkt (siehe unten).
2. Zum anderen kann die Handlungsausführung durch eine weit verzweigte Verankerung der Handlungsintention in das Selbst flexibler werden, etwa indem besser auf sich verändernde Umweltbedingungen reagiert werden kann. Eine gründliche Exploration der Selbstkongruenz legt somit schon in der Intentionsphase die Grundlage für eine demokratische Selbststeuerung (Kuhl 2001), die für eine funktionale Emotions- und Motivationsregulation in der Volitionsphase sorgt.

Die Intentionsphase mündet in die Ausbildung einer Handlungsintention, die mit anderen um eine Realisierung in der Volitionsphase wetteifert. Die stärkste Handlungsintention überschreitet den Rubikon und muss sich hierbei nicht nur gegen anderen Handlungsintentionen durchsetzen (vgl. Heckhausen 1989), die die gleiche Motivation erfüllen wollen, sondern auch gegen solche, die mit dem Versprechen locken, ganz andere und möglicherweise attraktivere Bedürfnisse zu erfüllen. Natürlich ist es auch möglich, mit einer einzigen Handlung mehrere Motivationen zu befriedigen.

3.3 Volitionsphase

3.3.1 Persistente Zielverfolgung

Mit der erfolgreichen Überschreitung des Rubikon setzt sich eine Lernintention durch und die Volitionsphase beginnt (vgl. Abb. 4). Dies bedeutet, dass die Zielverfolgung gegen die Störung durch andere Intentionen abgeschirmt wird (vgl. Achtziger & Gollwitzer 2007). Die Konzentration auf eine einzelne Intention ist mit einer Aktivierung des Intentionsgedächtnisses (sensu Kuhl 2000) und der Dämpfung von positiven Affekten verbunden. Neben der Abschirmung wird dem Intentionsgedächtnis in dieser Phase die Aufrechterhaltung einer abstrakten Gedächtnisrepräsentation der Intention zugeschrieben (ebd.). Dies dient vor allem dazu, schwierige Intentionen für eine längere Zeit aufrecht zu erhalten[11].

11 Eine Intention wird dann schwierig, wenn sie nicht sofort in eine Handlung umgesetzt werden kann, etwa wenn situative Barrieren dies verhindern oder wenn die passenden Gelegenheiten fehlen.

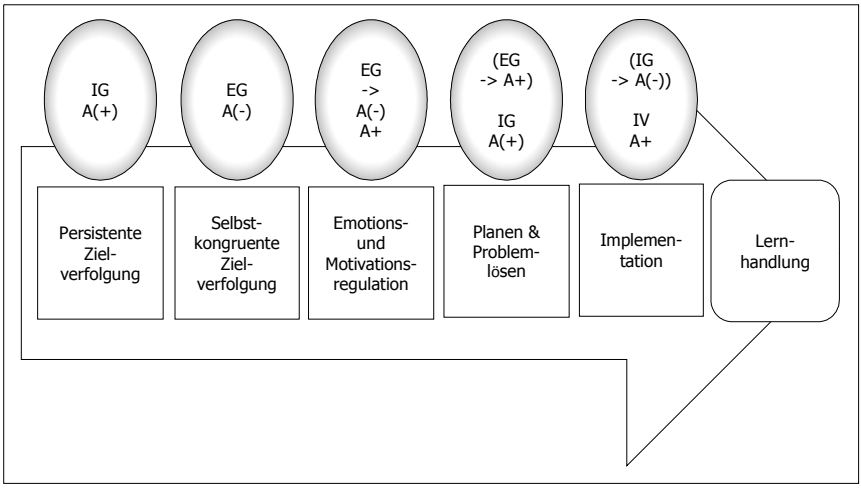


Abbildung 4: Die Volitionsphase im Integrierten Handlungsmodell in Anlehnung an Kuhl (2000) und die jeweils beteiligten Makrosysteme und Affekte. IG = Intentionsgedächtnis, EG = Extensionsgedächtnis, IV = Intuitive Verhaltenssteuerung, A(+) = Dämpfung von positivem Affekt, A(-) = Dämpfung von negativem Affekt, A+ = positiver Affekt

3.3.2 Selbstkongruente Zielverfolgung

Als zweiter Schritt in der Volitionsphase ist ein zielkongruentes Monitoring der internen und externen Umwelt nötig (Kuhl 2000). Dabei muss geprüft werden, ob die Handlungsverfolgung mit den eigenen Werten, Vorstellungen und Gefühlen im Einklang steht. Gleichzeitig muss konstant überwacht werden, ob günstige Gelegenheiten für eine Umsetzung der Lernhandlung auftreten. Für beide Überwachungsprozesse sind keine bewussten Routinen nötig – sie können im Hintergrund des Bewusstseins mitlaufen. Diese Vigilanzprozesse werden von Kuhl dem Extensionsgedächtnis zugeschrieben und gehen mit einer Dämpfung des negativen Affekts einher. Monitoring- und Vigilanzprozesse sind unbedingt nötig, um flexibel auf sich verändernde Umwelt- oder Selbstbedingungen reagieren zu können[12]. Die äußere Umwelt im Kontext von Lernen kann sich hierbei aus bestimmten Lernobjekten oder aus den Interaktionen mit anderen Lernenden bestehen. Gerade in kooperativen Lernszenarien ist ein ausreichendes Maß an Vigilanz nötig, um die Bedürfnisse der anderen Gruppenmitglieder wie auch die eigenen zu erkennen und aufeinander abzustimmen (vgl. Martens 2009).

12 Eine Zielverfolgung, die auf einer sehr starren und rigiden Kontrolle beruht, kann bei sich verändernden Umwelt- oder Selbstbedingungen zum Handlungsabbruch führen. Selbst bei einem Handlungserfolg durch starre Handlungskontrolle kann es zu unerwünschten Nebenwirkungen kommen, etwa zur Aufrechterhaltung von sehr ineffizienten Lernhandlungen, die dann andere Lernprozesse zeitlich verzögern, oder zu stressbedingten Belastungsreaktionen bis hin zum Burn-Out (vgl. Kuhl & Stolzenbacher 2010).

3.3.3 Emotions- und Motivationsregulation

Die wichtigsten Prozesse der Volitionsphase beziehen sich auf die Regulation von Aufmerksamkeit, Motivation und Emotion. Es kann etwa nötig sein, den negativen Affekt, der mit der Speicherung der Lernabsichten im Intentionsgedächtnis entsteht, zu dämpfen. Wird diese Regulationsform nicht beherrscht, kann das zur chronischen Vermeidung von schwierigen Lernprozessen führen, um den damit einhergehenden negativen Affekt zu vermeiden (vgl. Kuhl 2000)[13].

Idealtypisch könnte eine Regulation nach Kuhl wie folgt ablaufen: Positive Selbstreferenzen, die mit der Lernhandlung verknüpft sind, werden durch die Aktivitäten des Extensionsgedächtnis abgerufen, etwa die Erinnerung, ähnliche Lernhandlungen erfolgreich bewältigt zu haben. Damit kann ein positiver Affekt generiert werden, der bei Bedarf Selbstentspannungsprozesse einleiten und damit zur Dämpfung von negativem Effekt beitragen kann.

Eine wichtige Funktion kann das kollaborative Lernen in Gruppen erfüllen (Martens 2009) und für positive Gegenregulation sorgen, wenn die Gruppenmitglieder einander aufheitern und motivieren.

3.3.4 Planen und Problemlösen

Insbesondere für die Umsetzung von schwierigen Handlungen, die mit einem hohen Zeit- und Energieaufwand verbunden sind, ist es nötig, Planungs- und Problemlöseprozesse zu initiieren, die mit einer Aktivierung des Intentionsgedächtnisses und einer Dämpfung von positivem Affekt sowie generell mit einer Herabregulierung des Extensionsgedächtnisses verbunden sind. Eine große Toleranz von negativen Affekten und von Frustration ist unter Umständen für die Erfüllung von besonders langfristigen Lernmotiven hilfreich.

Allerdings muss an dieser Stelle betont werden, dass es kein Patentrezept für die Umsetzung von besonders schwierigen Lernintentionen gibt. Es hängt von den persönlichen und situativen Bedingungen ab, ob etwa die Implementationsphase (vgl. 3.3.5) mit einer sehr großen Beteiligung des Intentionsgedächtnis erreicht werden kann oder ob – gerade bei sehr langen Lernperioden – weitere Regulierungen unter Rückbezug auf die selbstkongruente Zielverfolgung (vgl. 3.3.2) sowie die Emotions- und Motivationsregulation (vgl. 3.3.3) erfolgen müssen (Kuhl & Stolzenbacher 2010).

Eine einseitige Ausrichtung auf Optimismus oder Problemorientierung ist wahrscheinlich kontraproduktiv (vgl. Kuhl 2000). Vermutlich sind es gerade die auf den spezifischen Lernverlauf abgestimmten Regulationsprozesse, die erfolgreiches Lernen ermöglichen. Welche Regulationsmöglichkeiten zum Lerngeschehen passen, kann vermutlich der Lernende am besten selbst einschätzen.

13 Im Gegensatz hierzu können leichte Lernintentionen durch sofortige Lernhandlungen aus dem Intentionsgedächtnis entladen werden.

Die Möglichkeiten zu einem Eingriff von außen erscheinen sehr eingeschränkt, denn etwa Lehrpersonen können kaum beurteilen, in welcher spezifischen Lernphase sich der Lernende gerade befindet. Selbst unter der Annahme, dass dies zu leisten wäre, können die Lernphasen sehr schnell wechseln. In einer zeitlich stabilen Lernphase kann möglicherweise Hilfe von außen angeboten werden. Allerdings können die im Extensionsgedächtnis vor allem vor- oder unbewusst ablaufenden Prozesse nicht direkt adressiert, sondern wohl nur indirekt durch eine pädagogische Grundhaltung gefördert werden, etwa durch eine positive pädagogische Beziehung (vgl. Deci & Ryan 1993, Kuhl & Stolzenbacher 2010).

Es ist wichtig, die Planungs- und Problemlöseprozesse nicht voreilig abzuschließen (Impulsivität) und trotzdem einen minimalen Anteil an positivem Affekt und Impetus zu erhalten, damit es nicht zum Stillstand des Lernprozesses oder zur Grübelei kommt (vgl. Kuhl 2000).

3.3.5 Implementation

In der Implementationsphase muss schließlich der Wille gebahnt werden (vgl. Kuhl 2000), d. h. die Intentionen müssen nun in eine tatsächliche Lernhandlung umgesetzt werden. Der von Kuhl hierfür idealtypisch vorgesehene Prozess hemmt das Intentionsgedächtnis und aktiviert die Intuitive Verhaltenssteuerung über die Verstärkung von positiven Affekten. Die Generierung von positiven Affekten ist auf verschiedenen Wegen möglich, etwa als Freude darüber, eine realistische Lernhandlung identifiziert zu haben. Der fortgeschrittene Lerner wird den positiven Affekt möglicherweise aus einem Selbstregulationsprozess heraus entwickeln.

Entgegen früheren Darstellungen des Integrierten Handlungsmodells (Martens 2000) soll hier vor einer mechanistischen Anwendung von Handlungsplänen insbesondere beim Lernen gewarnt werden. Die Umsetzung einer Lernintention in eine Lernhandlung kann natürlich durch eine kognitivistische Vorsatzbildung mit Plänen zum wie, wann und wo befördert werden, wenn die Motivierungs- und die Intentionsphase mit einem hohen Energetisierungsgrad durchlaufen werden und die Volitionsphase eine ausreichende Balance zwischen Selbstkongruenz (vgl. 3.3.2) und Problemlösung (vgl. 3.3.4) erreicht hat. Wenn umgedreht Motiv- und Intentionsstärke zu gering ausgeprägt sind, etwa wegen mangelhafter Verankerung im Selbst, dann können Vorsätze nur unter sehr starrer Selbstkontrolle mit entsprechend negativem Affekt umgesetzt werden. Das kann zu einer Anstrengungsvermeidung in der Zukunft führen oder zu einer unflexiblen und rigiden Ausführung von Lernhandlungen. Rigide Lernhandlungen können sich besonders dann negativ auf die Lernqualität auswirken, wenn Kreativität gefragt ist. Es muss also darauf geachtet werden, dass die metakognitive Zielführung einer Lernhandlung nicht notwendige kreative und andere vorbereitende Lernprozesse überlagert.

3.4 Rückkopplungs- oder Feedbackprozesse

Zu einem Integrierten Handlungsmodell gehören natürlich unbedingt Feedback- und Rückkopplungsprozesse (Ilgen et al. 1979; vgl. Rausch 2012), die allerdings

bisher nur unzureichend erforscht sind. Die in den Kapiteln 3.1-3.3 vorgestellte Phasenstruktur impliziert, dass es mindestens drei Rückmeldeschleifen zu allen Hauptphasen des Lernens geben muss: zum Ausgangspunkt der ursprünglichen Handlungsgenese und jeweils zum Beginn der Intentions- und der Volitionsphase.

1. Die Rückkopplung zum Anfang der Motivierungsphase prüft, ob die Lernhandlung auf eine Reduktion der ursprünglich wahrgenommenen Soll-Ist-Diskrepanz wirken kann (vgl. Eraut 2000). Diese lange Rückkopplungsschleife beinhaltet auch eine Überprüfung der Lernmotivation. Dies kann und sollte bei ungünstig verlaufenden Lernprozessen zu einer qualitativen Modifikation der ursprünglichen Lernmotivation führen, etwa zu einer Fokussierung anderer Aspekte der Soll-Ist-Diskrepanz. Eine Motivdominanz, die sich von den eigentlichen Lernprozessen abkoppelt, kann zu nicht-funktionalen Lernhandlungen führen, die einen magischen Charakter annehmen können.
2. Die Rückkopplungsschleife zum Anfang der Intentionsphase überprüft, ob die durchgeführte Lernhandlung noch zur Lernmotivation passt oder ob eine andere Lernintention ausgewählt werden muss. Ohne diese Rückkopplung besteht die Gefahr, dass eine bestimmte Lernintention stereotyp immer wieder aktualisiert wird, obwohl die eigentliche Lernmotivation gar nicht befriedigt werden kann. Im Extremfall kann dies zu einem Abbruch des Lernprozesses führen oder zumindest zu einem ungewünschten Lernrückstand, weil die mangelnde Befriedigung der Lernmotivation zu spät erkannt worden ist. Eine Intentionsdominanz, die sich von den eigentlichen Lernprozessen abkoppelt, kann zu nicht-funktionalen Lernhandlungen führen, die den Charakter von stereotypen Wiederholungen haben.
3. Die Rückkopplungsschleife zum Anfang der Intentionsphase ermöglicht eine Überprüfung, ob die Handlungsausführung noch zur Lernintention passt oder ob die Lernhandlung anders ausgeführt werden muss, um etwa die selbstkongruente Zielverfolgung zu verbessern. Eine Missachtung solcher Informationen führt möglicherweise zu einer starren Handlungsausführung, die dann nur mit einem starken negativen Affekt aufrechterhalten werden kann (vgl. 3.3.5).

Die hier beschriebenen Rückkopplungsprozesse sind als idealtypische Sichtweisen zu verstehen. Weitere Regulationsschleifen innerhalb der drei Hautphasen sind denkbar und – wie für die Regulationsprozesse in der Volitionsphase bereits expliziert – auch wünschenswert.

4 Die empirische Prüfung des Integrierten Handlungsmodells – ein Beispiel

Empirische Prüfungen des Integrierten Handlungsmodells sind in verschiedenen Projekten erfolgt (Martens 2000, 2006, 2007; Martens et al. 2008; Martens et al. 2009; Martens & Rost 1998). Die Daten wurden in erster Linie mit Fragebögen in

Form von Selbstberichten erhoben und zum Teil durch qualitative Interviews und Verhaltensbeobachtungen ergänzt.

Berichtet wird im Folgenden über eine Studie, die die Anwendung des Integrierten Handlungsmodells auf das Lernen von Statistik im Bereich der universitären Hochschulbildung untersucht. Deren Ziel bestand darin, typische Muster der auf dem Integrierten Handlungsmodell beruhenden Lerngenese zu identifizieren. Die Stichprobe (n = 424) besteht aus Studierenden an vier Universitäten in Norddeutschland (Bremen, Greifswald, Hamburg und Rostock) im Zeitraum von 2002 bis 2004, die an Einführungskursen für Statistik teilnahmen. Sie wurden gebeten, innerhalb der ersten acht Kurswochen einen Online-Fragebogen auszufüllen.

4.1 Untersuchte Skalen

Die inhaltliche Konzeption der Skalen wurde von Martens (2000) und Schulmeister (1983) übernommen und an das spezifische Lernen von Statistik angepasst. Jede Skala (Bedrohungswahrnehmung, Handlungs-Ergebnis-Erwartung, Instrumentalität, Kompetenzerwartung, Lernstrategien und Durchhaltekompetenz) wird mit einer kurzen Erklärung eingeleitet, so dass deutlich wird, um welche Art von Lernen es geht. Die Items werden mit Hilfe einer vierstufigen Antwortskala beantwortet: "stimme voll und ganz zu", "stimme überwiegend zu", "stimme teilweise zu" und "stimme nicht zu".

Die sieben Items der *Bedrohungswahrnehmung* sollen die antizipierten Probleme bei den Statistikkursen erfassen, z. B.: "Ich befürchte, dass ich eine schlechte Note bekommen werde".

Die *Handlungs-Ergebnis-Erwartung* wurde mit acht Items erfasst, die sich auf die direkten Ergebnisse von Lernhandlungen beziehen, z. B.: "Durch das Lösen von Aufgaben kann ich meine statistischen Kompetenzen steigern".

Die Skala *Instrumentalität* umfasst acht Items, die sich auf die längerfristigen Folgen des Lernens von Statistik beziehen, etwa "Ich erwarte, dass gute Noten in Statistik meine Berufsaussichten steigern".

Die Skala *Kompetenzerwartung* besteht aus acht Items, die die persönlichen Fähigkeiten erfassen, Statistik zu lernen, etwa "Ich traue mir zu, auch schwierige Aufgaben zu lösen".

Die Skala *Lernstrategien* wird durch zehn Items abgefragt, die verschiedene volitionale Selbstregulationsstrategien erfassen, etwa "Wenn ich eigentlich keine Lust mehr habe, stelle ich mir Erfolgserlebnisse vor und mache weiter".

Die *Durchhaltekompetenz* bezieht sich auf die Selbsteinschätzung, ein Ziel persistent zu verfolgen, etwa "Wenn ich Statistik lerne, dann arbeite ich immer wieder an dem Problem, bis ich es begriffen habe".

4.2 Methoden und Ergebnisse

Für die empirische Prüfung wurden drei sukzessiv aufeinander folgende Auswertungsschritte durchgeführt (zur Auswertungsstrategie vgl. Martens 2000, 2007; Martens & Rost 1998). Zunächst wurde jede Skala mit dem ordinalen Rasch-Modell (Masters 1982) und mit dem allgemeineren Mixed-Rasch-Modell (Rost 1990) analysiert. Zur Beurteilung der jeweiligen Lösungsgüte wurde vor allem der CAIC (Bozdogan & Ramirez 1988) sowie der Q-Index (Rost 2004) herangezogen. Für die Skalen Bedrohungswahrnehmung, Instrumentalität, Handlungs-Ergebnis-Erwartung und Kompetenzerwartung passte das eindimensionale Rasch-Modell besser zu den empirischen Daten als die alternative 2-Klassen-Lösung des Mixed-Rasch-Modells[14]. Für die Lernstrategien (vgl. Abb. 5) und die Durchhaltekompetenz (vgl. Abb. 6) musste das einfache Rasch-Modell verworfen werden zugunsten der 2-Klassen-Lösung des Mixed-Rasch-Modells. Für diese beiden Skalen wurde deshalb der quantitative Personenparameter durch einen qualitativen ergänzt (Rost 1990). Der qualitative Personenparameter entspricht dabei der maximalen Zuordnungswahrscheinlichkeit zu einer von den durch das Mixed-Rasch-Modell identifizierten Subpopulationen.

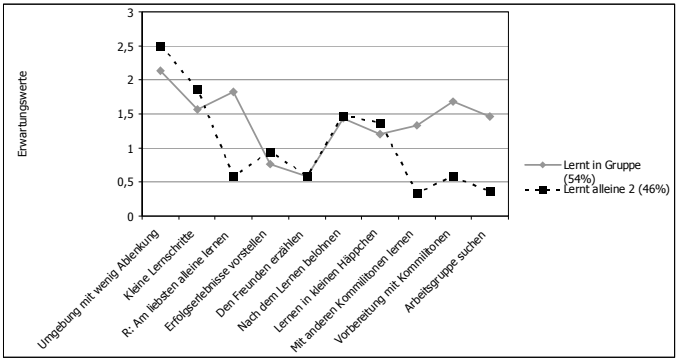


Abbildung 5: Die 2-Klassen-Lösung für die Skala Lernstrategien. Für die mit R gekennzeichneten Items wurden die Erwartungswerte umcodiert.

Bei der Skala Lernstrategie unterscheiden sich die beiden Subpopulationen vor allem in Bezug auf das Lernen in Gruppen. Die kleinere Subpopulation (46%) lernt lieber allein, die größere (54%) lieber in der Gruppe.

Die beiden Subpopulationen der Skala Durchhaltekompetenz unterscheiden sich im Spaßerleben. Die kleinere Subpopulation (43%) empfindet Spaß beim Lösen von Aufgaben, die größere hingegen (57%) nicht.

14 Es sei angemerkt, dass die kritische Prüfung eines eindimensionalen Rasch-Modells durch ein Mixed-Rasch-Modell mit zwei Klassen einer Prüfung aller nur möglichen DIF-Analysen entspricht.

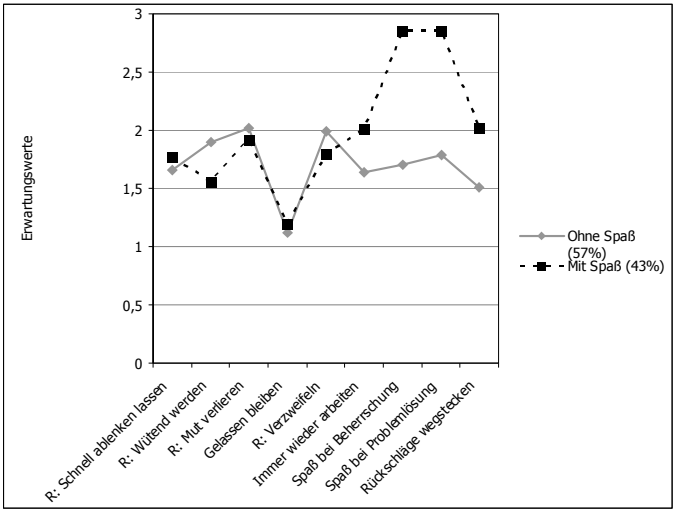


Abbildung 6: Die 2-Klassen-Lösung zur Skala Durchhaltekompetenz. Für die mit R gekennzeichneten Items wurden die Erwartungswerte umcodiert

In einem zweiten Schritt wurde eine Latent-Class-Analyse (Lazarsfeld & Henry 1968) angewendet, um die motivationalen, intentionalen und volitionalen Muster der Lern- und Handlungsgenese als entsprechende Subpopulationen zu identifizieren (vgl. etwa Martens 2007). Als Ausgangswerte dafür wurden die gerundeten Personenparameter aus den Rasch-Analysen verwendet. Bei den 2-Klassen-Lösungen wurde zusätzlich die wahrscheinlichere Klassenzugehörigkeit in die nachfolgenden Analysen einbezogen.

Mit diesen Ausgangswerten erfolgte eine Latent-Class-Analyse (LCA). Um die am besten passende Lösung zu identifizieren, wurden verschiedene statistische Kennwerte einbezogen, vor allem das Bootstrap-Verfahren (vgl. Efron & Tibshirani 1994, von Davier & Rost 1996a) und die mittleren Zuordnungswahrscheinlichkeiten zu den Subpopulationen. Um mögliche lokale Lösungen zu vermeiden, wurden die Berechnungen 5 Mal mit 15 Startwertesets wiederholt, die den jeweils besten Log-Likelihood-Wert durch mindestens zwei unterschiedliche Startwertesets bestätigten.

Nach Abwägung aller statistischen Kennwerte wurde die 7-Klassen-Lösung ausgewählt. Die mittleren Zuordnungswahrscheinlichkeiten liegen zwischen 0.84 und 0.88; dieses kann als zufriedenstellend eingestuft werden. Die Abweichung des identifizierten Modells von einer durch ein Bootstrap mit 200 generierten Vergleichsstichproben zeigt eine sehr gute Modellpassung (P(X>Z): 0.983 (Cressie-Read) 0.994 (Pearson χ^2)). Die empirische 7-Klassen-Lösung weicht kaum von den Vergleichsstichproben ab, die unter Annahme der Modellgültigkeit erzeugt worden sind.

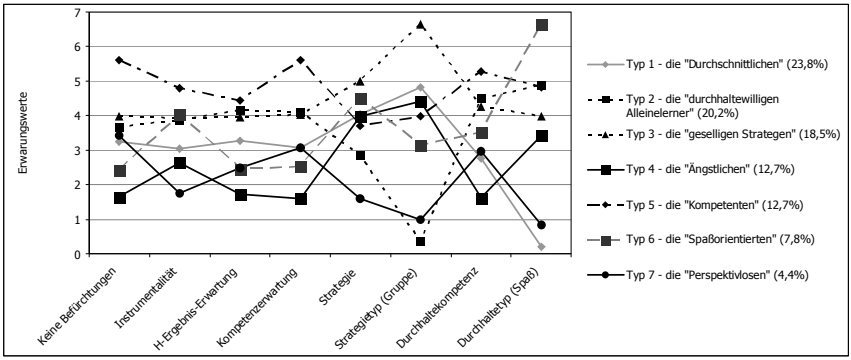


Abbildung 7: Die 7-Klassen-Lösung für die Lerntypen

In Abbildung 7 ist die 7-Klassen-Lösung der LCA zu sehen. Für eine bessere Lesbarkeit wurden alle Werte auf einen Wertebereich von 0 bis 7 angepasst (so wurden der Strategietyp und der Durchhaltetyp, die eigentlich nur zwischen 0 und 1 variieren, mit dem Faktor 7 multipliziert). Der Übersichtlichkeit wegen stellen die Abbildungen 8 und 9 noch einmal Teile der Ergebnisse dar.

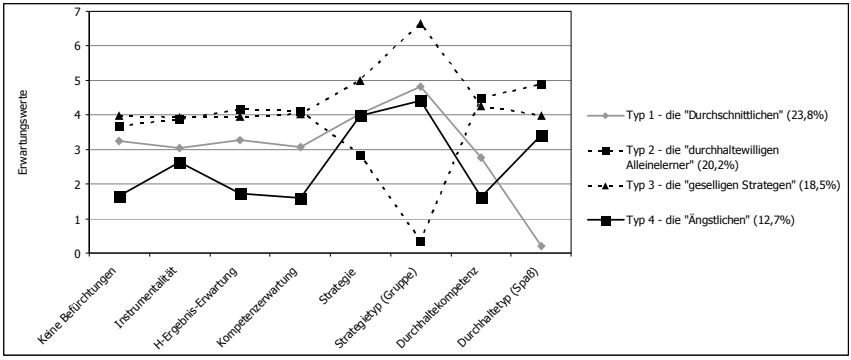


Abbildung 8: Die 7-Klassen-Lösung für die Lerntypen. Darstellung für Typ 1 bis 4

Die sieben durch die LCA identifizierten Lerntypen wurden wie folgt etikettiert:

Die *Durchschnittlichen* stellen mit etwa 24% die größte Subpopulation. Die Erwartungswerte der einzelnen Skalen (vgl. Abb. 8) bewegen sich jeweils im mittleren Bereich. Eine Ausnahme stellt die qualitative Ausprägung des Durchhaltetyps dar, die Gruppe der Durchschnittlichen empfindet eher keinen Spaß beim Lernen von Statistik.

Die *durchhaltewilligen Alleinelerner* machen etwa 20% der Stichprobe aus und weisen für die Handlungs-Ergebnis-Erwartung und die Kompetenzerwartung das zweithöchste Erwartungswertprofil auf. Charakteristisch sind die hohe Zuord-

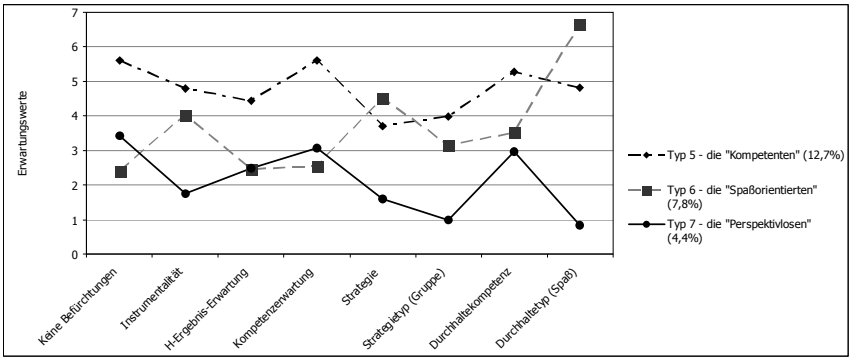


Abbildung 9: Die 7-Klassen-Lösung für die Lerntypen. Darstellung für Typ 5 bis 7

nungswahrscheinlichkeit zur Subgruppe der Studierenden, die lieber allein lernen, und eine relativ hohe Durchhaltekompetenz (vgl. Abb. 8).

Die *geselligen Strategen* haben eine Größe von etwa 19% und weisen ein ähnliches Niveau des Erwartungswertprofils auf wie die durchhaltewilligen Alleinelerner (vgl. Abb. 8). Sie unterscheiden sich aber von dieser und allen anderen Gruppen dadurch, dass sie am häufigsten Strategien einsetzen und am liebsten in der Gruppe lernen.

Die *Ängstlichen* machen 13% der Gesamtstichprobe aus und haben besonders große Befürchtungen, wenn sie an die Statistikveranstaltung denken. Sie zeigen die geringste Handlungs-Ergebnis-Erwartung, die geringste Kompetenzerwartung und die geringste Durchhaltekompetenz. Immerhin weisen sie beim Strategieeinsatz ein mittleres Niveau auf (vgl. Abb. 8).

Die *Kompetenten* stellen ebenfalls etwa 13% der Stichprobe, haben keine Befürchtungen und weisen auch bei Instrumentalität, Handlungs-Ergebnis-Erwartung und Durchhaltekompetenz die höchsten Erwartungswerte auf (vgl. Abb. 9).

Die *Spaßorientierten* haben einen Anteil von 8% an der Stichprobe und ein mittleres Erwartungswertprofil. Dabei zeigen sie einen relativ hohen Strategieeinsatz und sind mit großer Wahrscheinlichkeit der Subpopulation zuzuordnen, die auch Spaß beim Lernen von Statistik hat (vgl. Abb. 9).

Die *Perspektivlosen* sind mit etwa 4% die kleinste identifizierte Subpopulation. Sie haben die niedrigste Ausprägung bei der Instrumentalität und beim Strategieeinsatz. Sie gehören mit hoher Wahrscheinlichkeit zu denen, die lieber allein lernen und wenig Spaß beim Lernen von Statistik empfinden.

Insgesamt fällt auf, dass nur die Ängstlichen und die Perspektivlosen in ihren Lernprozessen ernsthaft gestört erscheinen, die zusammen gerade einmal 17% der Gesamtstichprobe ausmachen. Dies spricht für einen Selektionseffekt durch die Zugangsbeschränkungen zum Studium. Die Durchschnittlichen und die geselligen

Strategen haben vermutlich weniger Interesse an dem Fach Statistik als solches, wissen aber, welche Strategien sie einsetzen müssen, um das Studium erfolgreich abzuschließen. Die Kompetenten können sich auf ihre Kompetenz und ihre Durchhaltekompetenz verlassen und müssen vermutlich deshalb weniger kompensatorische Strategien einsetzen, sich z. B. Erfolgserlebnisse vorstellen. Die Spaßorientierten haben relativ große Befürchtungen und eine relativ niedrige Kompetenzerwartung. Sie setzen viele Strategien ein und haben Freude am Lernen. Wer Spaß an einer Sache hat, ist natürlich weniger auf die emotionale Unterstützung einer Lerngruppe angewiesen. Umgekehrt ist bei den Durchschnittlichen zu beobachten, dass ihnen das Lernen von Statistik nicht besonders viel Freude bereitet und dass sie lieber in der Gruppe lernen. Die durchhaltewilligen Alleinelerner wiederum verlassen sich – wie die Namengebung bereits verrät – lieber auf ihre Durchhaltekompetenz.

5 Methodische Konsequenzen

Mit dem Beispiel sollten einige wichtige methodische Konsequenzen für die empirische Prüfung des Integrierten Handlungsmodells illustriert werden, die ich im Folgenden noch einmal explizit erläutern möchte.

Bei der Entwicklung von Messinstrumenten, etwa Fragebögen, muss unbedingt beachtet werden, dass die Konstrukte für einen Lern- oder Handlungsbereich spezifiziert werden (vgl. Martens 2000). Aus forschungsökonomischen Gründen werden oft generalisierte Konstrukte verwendet, die nicht nach Lerndomänen unterscheiden und deshalb spezifische Regulationsdynamiken nicht erfassen. Dadurch werden spezifische Regulationsdynamiken in der Handlungsgenese oft gar nicht identifiziert. Generalisierte Konstrukte prägen sich ja erst dann aus, wenn sich Regulationsdynamiken über alle Lerndomänen hinweg gleichartig verfestigen. Die in dem Beispiel verwendete Spezifität, das Lernen von Statistik, könnte sicherlich zum Teil noch feiner in einzelne Lernhandlungen aufgelöst werden.

In der Natur eines Prozessmodells liegt es, dass jedes Individuum in einer unterschiedlichen Phase der Handlung sein kann. Darüber hinaus kann die Zeit, die für einzelne Prozesse benötigt wird, interindividuell *und* je nach Handlung sehr stark variieren. Aus den Funktionsprofilen der Makrosysteme (vgl. Kuhl 2001) lässt sich zumindest ableiten, dass die sequenziellen Verarbeitungsschritte des Intentionsgedächtnisses sehr lange Zeit beanspruchen können, während die parallel-holistische Verarbeitung des Extensionsgedächtnisses sehr schnell erfolgen kann. Wenn also viele Personen zu einem bestimmten Zeitpunkt querschnittlich zu einer bestimmten Motivation befragt werden, wird eine Momentaufnahme von Personen in sehr unterschiedlichen Prozessphasen erzeugt. Diese Mischung von qualitativ unterschiedlichen Prozesszuständen muss dann wieder entmischt werden. In Kapitel 4 wurde gezeigt, wie dies mit Hilfe von Mischverteilungsmodellen geschehen kann (Martens & Rost 1998), um damit bestimmte Lerntypen zu identifizieren.

Eine längsschnittliche Erhebung mit mehreren Messzeitpunkten über einen bestimmten Zeitraum kann entscheidend zur Erhöhung der Validität eines Prozessmodells beitragen. Dabei könnten Prozessdynamiken über die Zeit erfasst werden oder aber im Sinne der prognostischen Validität Maße für ein Lern- oder Handlungsergebnis (Martens 2000).

In der Regel können die zu untersuchenden Konstrukte nicht vollständig von Motivationen und Intentionen getrennt werden, die gar nicht im Fokus der jeweiligen Untersuchung stehen, so dass eine Konfundierung nicht ausgeschlossen werden kann. Dies kann die eigentlich interessierenden Motivationen und Intentionen abschwächen oder verstärken – etwa wenn besondere Lernleistungen erbracht werden, um den Lehrpersonen oder den Versuchsleitern zu gefallen. Im Grunde müssten also alle möglichen potenziell konkurrierenden Konstrukte mit erhoben werden, aber dies ist natürlich aus forschungspraktischen Überlegungen heraus – wie z. B. bei der referierten Studie – nicht immer möglich.

6 Ausblick

Das Integrierte Handlungsmodell in seiner Komplexität empirisch zu überprüfen, kann nicht von einzelnen Studien, sondern sicherlich nur durch ein breit gefächertes Forschungs*programm* geleistet werden. Ich hoffe deshalb, dass das Integrierte Handlungsmodell möglichst viele weitere Forschungsarbeiten anregen und parallel dazu als Reflektionsangebot in vielen Praxiszusammenhängen genutzt wird, wie es Jürgen Rost für das Umwelthandeln vorgemacht hat (Rost et al. 2001).

Literatur

Achtziger, Anja & Gollwitzer, Peter M. (2007). Motivation und Volition im Handlungsverlauf. In: Heckhausen, Jutta & Heckhausen, Heinz (Hrsg.), Motivation und Handeln. Heidelberg: Springer, S. 277-302.

Adams, Raymond J.; Wilson, Mark R. & Wang, Wen-Chung (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, pp. 1-23.

Adams, Raymond J.; Wu, Margaret L. & Macaskill, Greg (1997). Scaling methodology and procedures for the mathematics and science scales. In: Martin, Michael O. & Kelly, Dana L. (Eds.), Third International Mathematics and Science Study, Technical Report, Vol. II: Implementation and Analysis – Primary and Middle School Years. Chestnut Hill, MA: Boston College, pp. 111-145.

Agresti, Alan (1990). Categorical Data Analysis. New York: John Wiley & Sons, Inc.

Akaike, Hirotugu (1974). A new look at the statitical model identification. IEEE Transactions on Automatic Control, 19, pp. 716-723.

Akaike, Hirotugu (1987). Factor Analysis and AIC. Psychometrika, 52, pp. 317-332.

Andersen, Erling B. (1973). Conditional inference for multiple-choice questionnaires. British Journal of Mathematical and Statistical Psychology 26, pp. 31-44.

Andersen, Erling B. (1985). Estimating latent correlations between repeated testings. Psychometrika, 50, pp. 3-16.

Andrich, David (1978). A rating formulation for ordered response categories. Psychometrika, 43, pp. 561-573.

Andrich, David (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika, 47, pp. 105-113.

Andrich, David (1985). An elaboration of Guttman scaling with Rasch models for measurement. In: Tuma, N. B. (Ed.), Sociological methodology. San Francisco: Jossey-Bass, pp. 33-80.

Andrich, David (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. British Journal of Mathematical and Statistical Psychology, 49, pp. 347-365.

Andrich, David (2010). Understanding the response structure and process in the polytomous Rasch model. In: Nering, Michael L. & Ostini, Remo (Eds.), Handbook of polytomous item response theory models. New York: Routledge, pp. 123-152.

Andrich, David; de Jong, John H. A. L. & Sheridan, Barry E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In: Rost, Jürgen & Langeheine, Rolf (Eds.), Applications of latent trait and latent class models in the social sciences. Münster: Waxmann, pp. 59-70.

Aschaffenburg, Karen & Maas, Ineke (1997). Cultural and educational careers: the dynamics of social reproduction. American Sociological Review, 62, pp. 573-587.

Bamberg, Sebastian (1996). Allgemeine oder spezifische Einstellungen bei der Erklärung umweltschonenden Verhaltens? Zeitschrift für Sozialpsychologie, 27/1, S. 47-60.

Bandura, Albert (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84/2, pp. 191-215.

Barkmann, Claus; Erhart, M.; Schulte-Markwort, Michael & BELLA Study Group (2008). The German version of the Centre for Epidemiological Studies Depression Scale for Children: psychometric evaluation in a population-based survey of 7 to 17 years old children and adolescents – results of the Bella study. European Child and Adolescence Psychiatry, 17 (Suppl.), pp. 116-124.

Bauer, Petra (1991). Politische Orientierungen im Übergang: eine Analyse politischer Einstellungen der Bürger in West- und Ostdeutschland. 1990/91. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43, S. 433-453.

Baumert, Jürgen & Schümer, Gundel (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In: Deutsches PISA-Konsortium (Hrsg.), PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich, S. 323-407.

Becker, Rolf (2003). Educational expansion and persistent inequalities of education. Using subjective expected utility theory to explain increasing participation rates in upper secondary school in the Federal Republic of Germany. European Sociological Review, 19, pp. 1-24.

Belvedere, Shane L. & de Morton, Nathalie A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. Journal of Clinical Epidemiology, 63/12, pp. 1287-1297.

Billiet, Jaak B. & Davidov, Eldad (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. Sociological Methods & Research, 36, pp. 542-562.

Birnbaum, Allan (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, Frederic M. & Novick, Melvin R. (Eds.), Statistical theories of mental test scores. Reading, MA: Eddison-Wesley, pp. 395-479.

Bishop, Yvonne M. M.; Fienberg, Stephen E. & Holland, Paul W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Blossfeld, Hans-Peter; Schneider, Thorsten & Doll, Jörg (2009). Die Längsschnittstudie Nationales Bildungspanel: Notwendigkeit, Grundzüge und Analysepotential. Pädagogische Rundschau, 63/2, S. 249-259.

Blossfeld, Hans Peter & Shavit, Yossi (1993). Dauerhafte Ungleichheiten. Zur Veränderung des Einflusses der sozialen Herkunft auf die Bildungschancen in dreizehn industrialisierten Ländern. Zeitschrift für Pädagogik, 30, S. 25-52.

Bloom Howard S.; Hill, Carolyn J.; Black, Alison R. & Lipsey, Mark W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. Working paper. MDRC Working Papers on Research Methodology.

Bolt, Daniel. M.; Cohen, Allan S. & Wollack, James A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement, 39/4, pp. 331-348.

Bos, Wilfried; Bonsen, Martin; Baumert, Jürgen; Prenzel, Manfred; Selter, Christoph & Walther, Gerd (Hrsg.) (2008). TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Münster: Waxmann.

Bos, Wilfried; Lankes, Eva-Maria; Prenzel, Manfred; Schwippert, Knut; Walther, Gerd & Valtin, Renate (Hrsg.) (2003). Erste Ergebnisse aus IGLU. Münster: Waxmann.

Boudon, Raymond (1974). Education opportunity and social inequality. New York: Wiley.

Bourdieu, Pierre & Passeron, Jean Claude (1970). La reproduction. Eléments pour une théorie du système d'ensegnement. Paris: Edition de Minuit.

Bourdieu, Pierre (1982). Die feinen Unterschiede. Zur Kritik der gesellschaftlichen Urteilskraft. Frankfurt/M.: Suhrkamp.

Bourdieu, Pierre (1983). Ökonomisches Kapital, kulturelles Kapital und soziales Kapital. In: Kreckel, R. (Hrsg.), Soziale Ungleichheiten. Soziale Welt, Sonderband 2. Göttingen: Otto Schwarz, S. 183-198.

Bourdieu, Pierre (1993). Sozialer Sinn. Kritik der theoretischen Vernunft. Frankfurt/M.: Suhrkamp.

Bourdieu, Pierre (1993a). Soziologische Fragen. Frankfurt/M.: Suhrkamp.

Bozdogan, Hamparsum (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52, pp. 345-370.

Bozdogan, Hamparsum & Ramirez, Donald E. (1988). FACAIC: Model selection algorithm for the orthogonal factor model using AIC and CAIC. Psychometrika, 53/3, pp. 407-415.

Bradley, Ralph A. & Terry, Milton E. (1952). Rank analysis of incomplete block designs, I. the method of paired comparisons. Biometrika, 39, pp. 324-345.

Braun, Michael (1993). Ideologie oder objektive Lage? Anmerkungen zur Interpretation von Unterschieden und Ähnlichkeiten in den Einstellungen von Ost- und Westdeutschen. ZUMA-Nachrichten, 32, S. 7-21.

Breen, Richard & Goldthorpe, John (1997). Explaining educational differentials: towards a formal rational action theory. Rationality & Society, 9, pp. 275-305.

Breen, Richard; Luijkx, Ruud; Müller, Walter & Pollak, Reinhard (2009). Nonpersistent inequality in educational attainment. Evidence from eight European countries. American Journal of Sociology, 11, pp. 1475-1521.

Breen, Richard; Luijx, Ruud; Müller, Walter & Pollack, Reinhard (2010). Long-term trends in educational inequality in Europe: Class inequalities and gender differences. European Sociological Review, 28, pp. 31-48.

Brown, John Seely; Collins, Allan & Duguid, Paul (1989). Situated cognition and the culture of learning. Educational Researcher, 18/1, pp. 32-42.

Brunner, Wolfram & Walz, Dieter (1998). Selbstidentifikation der Ostdeutschen 1990-1997. Warum sich die Ostdeutschen zwar als Bürger zweiter Klasse fühlen, wir aber nicht auf die 'innere Mauer' treffen. In: Meulemann, Heiner (Hrsg.), Werte und nationale Identität im vereinten Deutschland. Opladen: Leske + Budrich, S. 229-250.

Burnham, Kenneth P. & Anderson, David R. (2002). Model selection and multimodel inference. New York: Springer.

Bybee, Rodger W. (1997). Toward an understanding of scientific literacy. In: Gräber, W. & Bolte, C. (Eds.), Scientific Literacy. Kiel: IPN, pp. 37-68.

Bybee, Rodger; McCrae, Barry & Laurie, Robert (2009). PISA 2006: An assessment of scientific literacy. Journal of research in Science Teaching 46/8, pp. 865-883.

Campbell, Donald & Fiske, Donald (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, pp. 81-105.

Carstensen, Claus H.; Frey, Andreas; Walter, Oliver & Knoll, Steffen (2007). Technische Grundlagen des dritten internationalen Vergleichs. In: Prenzel, Manfred; Artelt, Cordula; Baumert, Jürgen; Blum, Werner; Hammann, Marcus; Klieme, Eckhard & Pekrun, Reinhard (Hrsg.), PISA 2006 – Die Ergebnisse der dritten internationalen Vergleichsstudie. Münster: Waxmann, S. 367-390.

Carstensen, Claus H.; Lankes, Eva-Maria & Steffensky, Mirjam (2011). Naturwissenschaftliche Kompetenzen im Kindergartenalter – Instrument und Ergebnisse der Studie SnaKe. Zeitschrift für Erziehungswissenschaften, 14/4.

Cattell, Raymond B.; Weiß, Rudolf H. & Osterland, Jürgen (1997). Grundintelligenztest Skala 1 (CFT 1) (5., revid. Aufl.). Göttingen: Hogrefe.

Clogg, Clifford C. (1995). Latent class models. In: Arminger, Gerhard; Clogg, Clifford C. & Sobel, Michael E. (Hrsg.), Handbook of statistical modeling for the social and behavioral sciences. New York: Plenum Press, pp. 311-359.

Clogg, Clifford C.; Rudas, Tamas & Xi, Liwen (1995). A new index of structure for the analysis of models for mobility tables and other cross-classifications. In: Marsden, Peter (Ed.), Sociological methodology 1995. Oxford: Blackwell, pp. 197-222.

Cole, Jason C.; Rabin, Adele S.; Smith, Tom L. & Kaufman, Alan S. (2004). Development and validation of a Rasch-derived CES-D short form. Psychological Assessment, 16, pp. 360-372.

Coleman, James (1988). Social capital in the creation of human capital. American Journal of Sociology, 94, pp. 95-120.

Collins, Linda M. & Lanza, Stephanie T. (2010). Latent class and latent transition analysis: with applications in the Social, Behavioral, and Health Sciences. New York: Wiley.

Crayen, Claudia; Geiser, Christian; Scheithauer, Herbert & Eid, Michael (2011). Evaluating interventions with multimethod data: A structural equation modeling approach. Structural Equation Modeling, 18, pp. 497-524.

Crockett, L. J.; Randall, B. A.; Shen, Y.-L.; Russell, S. T. & Driscoll, A. K. (2005). Measurement equivalence of the Center for Epidemiological Studies Depression Scale for Latino and Anglo adolescents: A national study. Journal of Consulting and Clinical Psychology, 73, pp. 47-58.

Curran, Patrick (2003). Have multilevel models been structural equation models all along? Multivariate Behavioral Research, 38, pp. 529-569.

Dahrendorf, Ralf (1965). Bildung als Bürgerrecht. Hamburg: Nannen.

Davier, Matthias v. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. Methods of Psychological Research, 2/2, pp. 29-48. URL (Retrieved 14.01.2010): http://www.dgps.de/fachgruppen/methoden/mpr-online/issue3/art5/davier.pdf

Davier, Matthias v. (2001). WINMIRA 2001 user's guide. Kiel: IPN.

Davier, Matthias v. (2005). A general diagnostic model applied to language testing data. ETS Research Report RR-05-16. Princeton, NJ: Educational Testing Service.

Davier, Matthias v. (2007a). The mixture general diagnostic model. ETS Research Report RR-07-32. Princeton, NJ: Educational Testing Service. Also in: Hancock, G. & Samuelsen, K. M. (Eds.), Mixture models in latent variable research. Charlotte, NC: Information Age Publishing.

Davier, Matthias v. (2007b). Hierarchical general diagnostic models. Research Report, RR-07-19. ETS: Princeton, NJ.

Davier, Matthias v. (2009a). Mixture distribution item response theory, latent class analysis and diagnostic mixture models. In: Embretson, Susan (Ed.), Measuring Psychological Constructs: Advances in Model-Based Approaches. APA Press, pp. 11-34.

Davier, Matthias v. (2009b). Some notes on the reinvention of latent structure models as diagnostic classification models. Measurement – Interdisciplinary Research and Perspectives, 7/1, pp. 67-74.

Davier, Matthias v. (2010). Hierarchical mixtures of diagnostic models. Psychological Test and Assessment Modeling, 52/1, pp. 8-28. URL (Retrieved 10.05.2010): http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf

Davier, Matthias v. & Carstensen, Claus H. (Eds.) (2007). Multivariate and mixture distribution Rasch Models. New York: Springer.

Davier, Matthias v. & Rost, Jürgen (1995). Polytomous mixed Rasch models. In: Fischer, G. H. & Molenaar, I. W. (Eds.), Rasch models – Foundations, recent developments and applications. New York: Springer, pp. 371-379.

Davier, Matthias v. & Rost, Jürgen (1996). Die Erfassung transsituativer Copingstile durch Stimulus-Response Inventare. Diagnostica, 42/4, S. 313-331.

Davier, Matthias v. & Rost, Jürgen (1997). Self-Monitoring – A class variable? In: Rost, Jürgen & Langeheine, Rolf (Eds.), Applications of latent trait and latent class models in the socialsciences, pp. 296-304.

Davier, Matthias v.; Xu, Xueli & Carstensen, Claus H. (2011). Measuring learning and change in a longitudinal large-scale assessment with a general latent variable model. Psychometrika, 76/2, pp. 318-336.

Davier, Matthias v. & Yamamoto, Kentaro (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. Applied Psychological Measurement, 28/6, pp. 389-406.

Dayton, C. Mitchell (2003). Applications and computational strategies for the two-point mixture index of fit. British Journal of Mathematical and Statistical Psychology, 56, pp. 1-13.

De Boeck, Paul & Wilson, Mark R. (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer.

Deci, Edward & Ryan, Richard (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. Zeitschrift für Pädagogik, 39, S. 223-238.

De Graaf, Paul (1986). The impact of financial and cultural resources on educational attainment in the Netherlands. Sociology of Education, 59, pp. 237-246.

De Graaf, Nan Dirk; De Graaf, Paul & Kraaykamp, Gerbert (2000). Parental cultural capital and educational attainment in the Netherlands: A refinement of the cultural capital perspective. Sociology of Education, 73, pp. 92-111.

De Graaf, Nan Dirk & De Graaf, Paul (2002). Formal and popular dimensions of cultural capital: effects on children's educational attainment. The Netherland's Journal of Social Sciences, 38, pp. 167-186.

Di Maggio, Paul (1982). Cultural capital and school success: the impact of status culture participation on the grades of US highschool students. American Sociological Review, 47, pp. 189-201.

Di Maggio, Paul & Mohr, John (1985). Cultural capital, educational attainment, and marital selection. American Journal of Sociology, 90, pp. 1231-1261.

Di Stefano, Christine, & Motl, Robert W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. Structural Equation Modeling, 13, pp. 440-464.

Dittmar, Norbert & Schlobinski, Peter (Hrsg.) (1988). Wandlungen einer Stadtsprache. Berlinisch in Vergangenheit und Gegenwart. Berlin: Colloquium.

Dittmar, Norbert; Schlobinski, Peter & Wachs, Inge (Hrsg.) (1986). Berlinisch. Studien zum Lexikon, zur Spracheinstellung und zum Stilrepertoire. Berlin: Berlin Verlag Arno Spitz.

Ditton, Hartmut (1992). Ungleichheit und Mobilität durch Bildung: Theorie und empirische Untersuchung über sozialräumliche Aspekte der Bildungsbeteiligung. Weinheim: Juventa.

Dolan, Paul (1997). Modeling valuations for EuroQol health states. Medical Care, 35/11, pp. 1095-1108.

Dörner, Dietrich (1985a). Verhalten und Handeln. In: Dörner, Dietrich & Selg, Herbert (Hrsg.), Psychologie – Eine Einführung in ihre Grundlagen und Anwendungsfehler. Stuttgart, Berlin, Köln, Mainz: Kohlhammer, S. 73-86.

Dulisch, Frank (1986). Lernen als Form menschlichen Handelns. Eine handlungstheoretisch orientierte Analyse von Lernprozessen unter besonderer Berücksichtigung des Selbststeuerungsaspektes. Bergisch-Gladbach: Thomas Hobein.

Dumais, Susan A. (2002). Cultural capital, gender, and school success. The role of habitus. Sociology of Education, 75, pp. 44-68.

Dumenci, Levent (2000). Multitrait-multimethod analysis. In: Tinsley, Howard & Brown, Steven (Eds.), Handbook of applied multivariate statistics and mathematical modeling. San Diego, CA, US: Academic Press Inc., pp. 583-611.

Edwards, Michael C.; Cheavens, Jennifer S.; Heiy, Jane E. & Cukrowicz, Kelly C. (2010). A reexamination of the factor structure of the CES-D scale. Is a one-factor model plausible? Psychological Assessment, 22, pp. 711-715.

Efron, Bradley & Tibshirani, Robert John (1994). An introduction to the bootstrap. Boca Raton: Chapman & Hall/CRC.

Ehmke, Timo (2008). Welche Bedeutung haben lernförderliche Prozesse und naturwissenschaftsbezogene Einstellungen im Elternhaus für die Erklärung sozialer Disparitäten in der naturwissenschaftlichen Kompetenz? In: Prenzel, Manfred & Baumert, Jürgen (Hrsg.), Vertiefende Analysen zu PISA 2006. Zeitschrift für Erziehungswissenschaften, Sonderband 10. Wiesbaden: VS-Verlag, S. 129-148.

Eid, Michael (1995). Modelle der Messung von Personen in Situationen. Weinheim: Psychologie Verlags Union.

Eid, Michael (2000). A multitrait-multimethod model with minimal assumptions. Psychometrika, 65, pp. 241-261.

Eid, Michael (2010). Convergent and discriminant validation by the multitrait-multimethod matrix [Psychological Bulletin, 56/2, pp. 81-105.]. In: Salkind, Neil (Ed.) Encyclopedia of research design. Newbury Park: Sage, pp. 254-256.

Eid, Michael; Lischetzke, Tanja; Nussbeck, Fridtjof & Trierweiler, Lisa (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M–1) model. Psychological Methods, 8, pp. 38-60.

Eid, Michael; Lischetzke, Tanja & Nussbeck, Fridtjof (2006a). Structural equation models for multitrait-multimethod data. In: Eid, Michael & Diener, Ed (Eds.), Handbook of psychological measurement: A multimethod perspective. Washington, DC: American Psychological Association, pp. 283-299.

Eid, Michael; Nussbeck, Fridtjof; Geiser, Christian; Cole, David; Gollwitzer, Mario & Lischetzke, Tanja (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. Psychological Methods, 13, pp. 230-253.

Eid, Michael; Nussbeck, Fridtjof & Lischetzke, Tanja (2006b). Multitrait-Multimethod-Analyse. In: Petermann, Franz & Eid, Michael (Hrsg.), Handbuch der Psychologischen Diagnostik. Göttingen: Hogrefe, S. 332-345.

Eid, Michael & Rauber, Michael (2000). Detecting measurement invariance in organizational surveys. European Journal of Psychological Assessment, 16, pp. 20-30.

Eitle, Temela & Eitle, David J. (2002). Race, cultural capital, and the educational effects of participation in sports. Sociology of Education, 75, pp. 123-146.

Embretson, Susan E. (1991). A multidimensional latent trait model for measuring learning and change. Psychometrika, 56, pp. 495-515.

Epstein, Seymour (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. Journal of Personality, 51, pp. 360-392.

Eraut, Michael (2000). Non-formal learning and tacit knowledge in professional work. British Journal of Educational Psychology, 70, pp. 113-136.

Erikson, Robert & Rudolphi, Frida (2010). Change in social selection on upper secondary school – primary and secondary effects in Sweden. European Sociological Review, 26, pp. 291-305.

Farkas, George; Grobe, Robert; Sheehan, Daniel & Shaun, Yuan (1990). Cultural resources and school success: gender, ethnicity, and poverty groups within an urban district. American Sociological Review, 55, pp. 127-142.

Fischer, Gerhard H. (1974). Einführung in die Theorie psychologischer Tests. Bern: Huber.

Fischer, Gerhard H. & Formann, Anton K. (1972). An algorithm and a Fortran program for estimation of parameters for a linear logistic test model. Research Bulletin 9. Vienna, Austria: University of Vienna, Department of Psychology.

Fischer, Gerhard H. & Parzer, Peter (1991). An extension of the rating scale model with an application tot he measurement of change. Psychometrika, 56, pp. 637-651.

Fischer, Gerhard H. & Ponocny, Ivo (1994). An extension oft the partial credit model with an application to the measurement of change. Psychometrika, 59, pp. 177-192.

Fischer, Gerhard H. & Ponocny, Ivo (1995). Extended rating scale and partial credit models for assessing change. In: Fischer, Gerhard H. & Molenaar, Ivo W. (Eds.), Rasch models: Foundations, recent developments, and applications. New York: Springer, pp. 353-370.

Forgas, Joseph P.; Wyland, Carrie L. & Laham, Simon M. (2006). Hearts and minds: An introduction to the role of affect in social cognition and behavior. In: Forgas, Joseph P. (Ed.), Affect in social thinking and behavior. New York, Hove: Psychology Press, pp. 3-18.

Formann, Anton (2002). Identifying types, response errors, and unscalable respondents from personality questionnaires. Psychologische Beiträge, 44, pp. 78-93.

Formann, Anton (2006). Testing the Rasch model by means of the mixture fit index. British Journal of Mathematical and Statistical Psychology, 59, pp. 89-95.

Frese, Michael & Zapf, Dieter (1994). Action as the core of work psychology: A german approach. In: Triandis, Harry C.; Dunnette, Marvin D. & Hough, Leaetta M. (Eds.), Handbook of industrial and organizational psychology (2 ed., Vol. 4). Palo Alto, California: Consulting Psychologists Press, pp. 271-340.

Ganzeboom, Harry B. G.; De Graaf, Paul M. & Robert, Peter (1990). Cultural reproduction theory on socialist ground. Research in Social Stratification and Mobility, 9, pp. 79-104.

Ganzeboom, Harry B. G.; De Graaf, Paul M.; Treimann, Donald J. & De Leeuw, Jan (1992). A standard international socio-economic index of educational status. Social Science Research, 25, pp. 201-239.

GDSU – Gesellschaft für Didaktik des Sachunterrichts (2002). Perspektivrahmen Sachunterricht. Bad Heilbrunn: Klinkhardt.

Geiser, Christian; Eid, Michael & Nussbeck, Fridtjof (2008). On the meaning of the latent variables in the CT-C(M–1) model: A comment on Maydeu-Olivares & Coffman (2006). Psychological Methods, 13, pp. 49-57.

Georg, Werner (2004). Cultural capital and social inequality in the life course. European Sociological Review, 20, pp. 333-344.

Georg, Werner (2005). Die Reproduktion sozialer Ungleichheit im Lebenslauf. Zeitschrift für Soziologie der Erziehung und Sozialisation, 25, S. 178-197.

Georg, Werner (2006). Kulturelles Kapital und Statusvererbung. In: Georg, Werner (Hrsg.), Soziale Ungleichheit im Bildungssystem. Eine empirisch-theoretische Bestandsaufnahme. Konstanz: UVK, S. 123-146.

Georg, Werner (2012). Social inequality in German higher education during the 20th century. In: Salzborn, Samuel; Davidov, Eldad & Reinecke, Jost (Eds.), Methods, theories and empirical applications in the social sciences. Festschrift for Peter Schmidt. Wiesbaden: Springer VS, pp. 325-334.

Giles, Howard; Hewstone, Miles; Ryan, Ellen & Johnson, Patricia (1987). Research on language attitudes. In: Ammon, Ulrich; Dittmar, Norbert & Mattheier, Klaus (Hrsg.), Soziolinguistik. Erster Halbband. Berlin, New York: de Gruyter, S. 585-597.

Gollwitzer, Mario; Eid, Michael, & Jürgensen, Ralph (2005). Response styles in the assessment of anger expression. Psychological Assessment, 17, pp. 56-69.

Gollwitzer, Peter M. (1990). Action phases and mind-sets. In: Higgins, Tory E. & Sorrentino, Richard M. (Eds.), Handbook of motivation and cognition. Foundations of social behavior (Vol. 2). New York, London: The Guilford Press, pp. 53-92.

Goodman, Leo A. (1972). A modífied multiple regression approach to the analysis of dichotomous variables. American Sociological Review, 37, pp. 28-46.

Goodman, Leo A. & Kruskal, William H. (1954). Measures of association for cross-classifications. Journal of the American Statistical Association, 54, pp. 310-364.

Gruber, Hans & Renkl, Alexander (2000). Die Kluft zwischen Wissen und Handeln: Das Problem des trägen Wissens. In: Neuweg, Georg Hans (Hrsg.), Wissen – Können – Reflexion. Ausgewählte Verhältnisbestimmungen. Innsbruck, Wien, München: StudienVerlag, S. 155-174.

Gudjons, Herbert (2008). Handlungsorientiert lehren und lernen (7., aktualisierte Auflage). Bad Heilbrunn: Klinkhardt.

Haberman, Shelby J. (2007). The interaction model. In: Davier, Matthias v. & Carstensen, Claus H. (Eds.), Multivariate and mixture distribution Rasch models. Statistics for social and behavioral sciences, Part II. Springer: New York, pp. 201-216.

Hacker, Winfried (1978). Allgemeine Arbeits- und Ingenieurspsychologie. Bern, Stuttgart, Wien: Huber.

Hambleton, Ronald K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12/3, pp. 38-47.

Hancock, Gregory & Samuelsen, Karen M. (Eds.) (2007). Mixture models in latent variable research. Charlotte, NC: Information Age Publishing.

Handl, Johann (1985). Mehr Chancen im Bildungssystem. Erfolg der Bildungsreform oder statistisches Artefakt? Kölner Zeitschrift für Soziologie und Sozialpsychologie, 37, S. 698-729.

Hautzinger, Martin & Bailer, Maja (1993). Allgemeine Depressions-Skala. Weinheim: Beltz.

Hautzinger, Martin; Bailer, Maja; Hofmeister, Dirk & Keller, Ferdinand (2012). Allgemeine Depressions-Skala. (2. Aufl.). Göttingen: Hogrefe.

Hays, Ron D.; Morales, Leo S. & Reise, Steve P. (2000). Item response theory and health outcome measurement in the 21st century. Medical Care, 38 (9 suppl.), pp. II-28-II-42.

Heckhausen, Heinz (1989). Motivation und Handeln (2. Auflage). Berlin: Springer.

Herdman, Michael; Gudex, Claire; Lloyd, Andrew; Janssen, Bas; Kind, Paul; Parkin, David; Bonsel, Gouke & Badia, Xavier (2011). Development and preliminary testing of the new five-level version of the EQ-5D (EQ-5D-5L). Quality of Life Research, Epub ahead of print, April 9.

Hinz, Andreas; Brähler, Elmar; Geyer, Michael & Körner, Annett (2003). Urteilseffekte beim NEO-FFI. Diagnostica, 49, S. 157-163.

Hoeffler, David (2010). Erweiterte Item-Response-Modelle zur Analyse invertierter und nicht-invertierter polytomer Items. Universität Mannheim: Unveröffentlichte Diplomarbeit.

Hofer, Manfred (1981). Handlung und Handlungstheorien. In: Schiefele, Hans & Krapp, Andreas (Hrsg.), Handlexikon zur Pädagogischen Psychologie. München: Franz Ehrenwirth, S. 159-166.

Höfling, Volkmar; Moosbrugger, Helfried; Schermelleh-Engel, Karin & Heidenreich, Thomas (2011). Mindfulness or mindlessness? A modified version of the Mindful Attention and Awareness Scale (MAAS). European Journal of Psychological Assessment, 27, pp. 59-64.

Horan, Patrick M.; DiStefano, Christine & Motl, Robert W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? Structural Equation Modeling, 10, pp. 435-455.

Ilgen, Daniel R.; Fisher, Cynthia D. & Taylor, M. Susan (1979). Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology, 64/4, pp. 349-371.

Iwata, Noboru; Umesue, M.; Egashira, K.; Hiro, H.; Mizoue, T.; Mishima, N. & Nagata, S. (1998). Can positive affect items be used to assess depressive disorders in the Japanese population? Psychological Medicine, 28, pp. 153-158.

Jech, Thomas (1983). The ranking of incomplete tournaments: A mathematician's guide to popular sports. The American Mathematical Monthly, 90, pp. 246-266.

Jiao, Hong; Davier, Matthias v. & Wang, Shugong (2010). Polytomous mixture Rasch testlet model. Paper presented at the National Council for Measurement in Education. Annual meeting in Denver, CO.

Kaase, Max & Bauer-Kaase, Petra (1998). Deutsche Vereinigung und innere Einheit 1990-1997. In: Meulemann, Heiner (Hrsg.), Werte und nationale Identität im vereinten Deutschland. Opladen: Leske + Budrich, S. 251-267.

Kalmijn, Matthias & Kraaykaamp, Gerbert (1996). Race, cultural capital, and schooling: An analysis of trends in the United States. Sociology of Education, 69, pp. 22-34.

Kelderman, Henk (1984). Loglinear Rasch model tests. Psychometrika, 49, pp. 223-245.

Kelderman, Henk & Macready, George B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27, pp. 307-327.

Keller, Ferdinand (2010). Measurement equivalence of the CES-D in boys and girls during adolescence. Chinese Medical Journal, 123, Suppl. 2, p. 290.

Keller, Ferdinand; Meyer, Thomas D. & Pössel, Patrick (2008). Identifizierung von skalierbaren und unskalierbaren Personengruppen in der Allgemeinen Depressionsskala (ADS) anhand von Mischverteilungsmodellen. In: Reinecke, Jost & Tarnai, Christian (Hrsg.), Klassifikationsanalysen in Theorie und Praxis. Münster: Waxmann, S. 167-187.

Kempf, Wilhelm (2008). Forschungsmethoden der Psychologie. Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik. Band 2: Quantität und Qualität. Berlin: regener.

Kenny, David (1995). The multitrait–multimethod matrix: Design, analysis, and conceptual issues. In: Shrout, Patrick & Fiske, Susan (Eds.), Personality research, methods, and theory: A festschrift honoring Donald W. Fiske. Hillsdale, NJ: Erlbaum, pp. 111-124.

Klix, Friedhart (1980). Erwachendes Denken. Berlin: VEB Deutscher Verlag der Wissenschaften.

Klöckner, Christian A. & Matthies, Ellen (2004). How habits interfere with norm-directed behaviour: A normative decision-making model for travel mode choice. Journal of Environmental Psychology, 24/3, pp. 319-327.

Köhler, Helmut (1992). Bildungsbeteiligung und Sozialstruktur in der Bundesrepublik Deutschland. Zu Stabilität und Wandel der Ungleichheit von Bildungschancen. Berlin: Max Planck Institut für Bildungsforschung.

Kolen, Michael J. & Brennan, Robert L. (2004). Test equating, scaling, and linking: Methods and practices (2nd ed.). New York: Springer.

Kracauer, Siegfried (1952). The challenge of qualitative content analysis. Public Opinion Quaterly, 16, pp. 631-642.

Krapp, Andreas (1999). Intrinsische Lernmotivation und Interesse. Zeitschrift für Pädagogik, 45/3, S. 387-406.

Krauth, Joachim & Lienert, Gustav A. (1973). Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendung in Psychologie und Medizin. Freiburg: Alber.

Krone, Heinz Walter (1974). Untersuchung mit einer deutschen Form der Repression-Sensitization-Skala. Zeitschrift für Klinische Psychologie, 3, S. 238-260.

Kuhl, Julius (2000). The volitional basis of Personality Systems Interaction Theory: applications in learning and treatment contexts. International Journal of Educational Research, 33, pp. 665-703.

Kuhl, Julius (2001). Motivation und Persönlichkeit – Interaktionen psychischer Systeme. Göttingen, Bern, Toronto, Seattle: Hogrefe.

Kuhl, Julius & Stolzenbacher, Claudia (2010). nifbe Jahresbericht 2009. URL (Zugang 31.05.2011): http://nifbe.de/media/Jahresbericht_2009_Stand10%2012%2010.pdf

Kultusministerkonferenz (2004). Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss (Jahrgangsstufe 10), München: Luchterhand.

Kultusministerkonferenz (2009). Empfehlung der Kultusministerkonferenz zur Stärkung der mathematisch-naturwissenschaftlich-technischen Bildung (Beschluss der Kultusministerkonferenz vom 07.05.2009). URL (Zugriff 16.07.2010): http://www.kmk.org/fileadmin/veroeffentlichungen_ beschluesse/2009/2009_05_07-Empf-MINT.pdf

Landmann, Meike; Perels, Franziska; Otto, Barbara & Schmitz, Bernhard (2009). Selbstregulation. In: Pädagogische Psychologie. Berlin Heidelberg: Springer, S. 49-70.

Landmann, Meike & Schmitz, Bernhard (2007). Welche Rolle spielt Self-Monitoring bei der Selbstregulation und wie kann man mit Hilfe von Tagebüchern die Selbstregulation fördern? In: Gläser-Zikuda, M. & Hascher, T. (Hrsg.), Lernprozesse dokumentieren, reflektieren und beurteilen. Lerntagebuch & Portfolio in Forschung und Praxis. Bad Heilbrunn: Klinkhardt, S. 149-169.

Langeheine, Rolf (2006). Analyse kategorialer Daten und die Beurteilung der Modellgültigkeit mit dem Pi-Star Index. Zeitschrift für Sozialpsychologie, 37, S. 197-204.

Langeheine, Rolf & van de Pol, Frank (1990). Veränderungsmessung bei kategorialen Daten. Zeitschrift für Sozialpsychologie, 21, S. 88-100.

Langeheine, Rolf & Rost, Jürgen (1997). Latent trait and latent class models. New York: Plenum.

Lankes, Eva-Maria; Steffensky, Mirjam & Carstensen, Claus H. (2011). Naturwissenschaftliche Experimente im Kindergarten – Didaktische Analyse von Materialien zum Experimentieren. Zeitschrift für Grundschulforschung.

Lazarsfeld, Paul F. (1950). Logical and mathematical foundations of latent structure analysis. In: Stouffer, S. A.; Guttman, L.; Suchman, E. A.; Lazarsfeld, P. F.; Star, S. A. & Clausen, J. A. (Eds.), Studies in social psychology in world war II, Vol. IV. Measurement and Prediction. Princeton/N.J.: Princeton University Press, pp. 362-412.

Lazarsfeld, Paul F. & Henry, Neil W. (1968). Latent structure analysis. Boston: Houghton Mifflin.

Lazarus, Richard S. & Launier, Raymond (1981). Streßbezogene Transaktionen zwischen Personen und Umwelt. In: Nitsch, Jürgen R. (Hrsg.), Stress. Theorien, Untersuchungen, Maßnahmen. Bern, Stuttgart, Wien: Huber, S. 213-259.

Lewin, Kurt (1926a). Untersuchungen zur Handlungs- und Affekt-Psychologie. I. Vorbemerkungen über die psychischen Kräfte und Energien und über die Struktur der Seele. Psychologische Forschung, 7/1, S. 294-329.

Lewin, Kurt (1926b). Untersuchungen zur Handlungs- und Affekt-Psychologie. II. Vorsatz, Wille und Bedürfnis. Psychologische Forschung, 7/1, S. 330-385.

Lord, Frederic M. & Novick, Melvin R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Maaz, Kai & Nagy, Gabriel (2009). Der Übergang von der Grundschule in die weiterführenden Schulen des Sekundarschulsystems: Definition, Spezifikation und Quantifizierung primärer und sekundärer Herkunftseffekte. In: Baumert, Jürgen; Maaz, Kai & Trautwein, Ulrich

(Hrsg.), Bildungsentscheidungen. Zeitschrift für Erziehungswissenschaften, Sonderheft 12, S. 151-182.

Machunsky, Maya & Meiser, Thorsten (2006). Personal Need for Structure als differentialpsychologisches Konstrukt in der Sozialpsychologie: Psychometrische Analyse und Validierung einer deutschsprachigen PNS-Skala. Zeitschrift für Sozialpsychologie, 37, S. 87-97.

Magidson, Jay & Vermunt, Jeroen K. (2003). Comparing latent class factor analysis with the traditional approach in data mining. In: Bozdogan, Hamparsum (Ed.), Statistical data mining and knowledge discovery. Boca Raton: Chapman & Hall, pp. 373-383.

Mair, Patrick, & Hatzinger, Reinhold (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. Journal of Statistical Software, 20, pp. 1-20.

Mair, Patrick; Hatzinger, Reinhold & Maier, Marco (2010). Package 'eRm'. URL: http://r-forge.r-project.org/projects/erm/

Mandl, Heinz & Kopp, Birgitta (2006). Lehren in der Weiterbildung aus pädagogisch-psychologischer Sicht. Sechs Leitprinzipien didaktischen Handelns. In: Nuissl, Ekkehard (Hrsg.), Vom Lernen zum Lehren. Lern- und Lehrforschung für die Weiterbildung. Bielefeld: Bertelsmann, S. 117-128.

Marsh, Herbert W.; Scalas, L. Francesca & Nagengast, Benjamin (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. Psychological Assessment, 22, pp. 366-381.

Martens, Thomas (2000). Kognitive und affektive Bedingungen von Umwelthandeln. Berlin: dissertation.de.

Martens, Thomas (2006). Identification of motivational, intentional and volitional patterns: Implications for learning achievement and self-regulation strategies. In: Annual Meeting of the American Educational Research Association.

Martens, Thomas (2007). Handlungstypen als Grundlage für die Maßschneiderung von umweltpsychologischen Informationen. Umweltpsychologie, 11/1, S. 69-87.

Martens, Thomas (2009). Unterstützung der intrinsischen Motivation beim Lernen von Methodenlehre in Gruppen mit Hilfe von Online-Monitoring. In: Krämer, Michael; Preiser, Siegfried & Brusdeylins, Kerstin (Hrsg.), Psychologiedidaktik und Evaluation VII. Aachen: Shaker.

Martens, Thomas; Erdwien, Birgitt & Ramm, Karen (2008). Risikoverarbeitung und Risikoverhalten am Beispiel extremer Hochwasserereignisse. Schlussbericht für das Bundesministerium für Bildung und Forschung. Förderkennzeichen: 0330693C.: Universität Bremen.

Martens, Thomas; Garrelts, Heiko; Grunenberg, Heiko & Lange, Hellmuth (2009). Taking the heterogeneity of citizens into account: flood risk communication in coastal cities – a case study of Bremen. Nat. Hazards Earth Syst. Sci., 9, pp. 1931-1940.

Martens, Thomas & Rost, Jürgen (1998). Der Zusammenhang von wahrgenommener Bedrohung durch Umweltgefahren und der Ausbildung von Handlungsintentionen. Zeitschrift für Experimentelle Psychologie, 45/4, S. 345-364.

Martens, Thomas; Rost, Jürgen & Gresele, Christiane (1998). Identifikation von kognitiven, affektiven und sozialen Faktoren des Umwelthandelns – Skalensammlung. In: Glöckner-Rist, Angelika & Schmidt, Peter (Hrsg.), ZUMA Informationssystem. Version 2.0. Elektronisches Handbuch sozialwisenschaftlicher Erhebungssysteme. [CD-Rom, FTP] Mannheim: ZUMA.

Martin, Michael O.; Mullis, Ina V.S.; Gonzales, Eugenio J.; Gregory, Kelvin D.; Smith, Teresa A.; Chrostowski, Steven J.; Garden, Robert A. & O'Connor, Kathleen M. (2000). TIMSS 1999 International Science Report. Chestnut Hill, MA: The International Study Center, Boston College.

Masters, Geofferey N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, pp. 149-174.

Masters, Geofferey N. & Wright, Benjamin D. (1997). The partial credit model. In: van der Linden, Wim J. & Hambleton, Ronald K. (Eds.), Handbook of modern item response theory. New York: Springer, pp. 101-121.

Matschinger, Herbert; Schork, Astrid; Riedel-Heller, Steffi G. & Angermeyer, Matthias C. (2000). Zur Anwendung der CES-D bei älteren Menschen: Dimensionsstruktur und Messartefakte. Diagnostica, 46, S. 29-37.

Medrich, Elliot A. & Griffith, Jeanne E. (1992). International mathematics and science assessment: What have we learned. U.S. Department of Education. NCES 92-011.

Meiser, Thorsten (2007). Rasch models for longitudinal data. In: Davier, Matthias v. & Carstensen, Claus H. (Eds.), Multivariate and mixture distribution Rasch models. Extensions and applications. New York: Springer, pp. 191-199.

Meiser, Thorsten & Machunsky, Maya (2008). The personal structure of personal need for structure. A mixture-distribution Rasch analysis. European Journal of Psychological Assessment, 24, pp. 27-34.

Meulemann, Heiner (1992). Expansion ohne Folgen? Bildungschancen und sozialer Wandel in der Bundesrepublik. In: Glatzer, Wolfgang (Hrsg.), Entwicklungstendenzen in der Sozialstruktur. Frankfurt/M.: Campus, S. 123-157.

Meulemann, Heiner (1998). Wertunterschiede zwischen West- und Ostdeutschland – Fakten und Erklärungsmöglichkeiten. In: Meulemann, Heiner (Hrsg.), Werte und nationale Identität im vereinten Deutschland. Opladen: Leske + Budrich, S. 7-21.

Miller, Georg A.; Galanter, Eugene & Pribram, Karl H. (1960/1973). Strategien des Handelns – Pläne und Strukturen des Verhaltens. Stuttgart: Klett.

Millsap, Roger (1995). The statistical analysis of method effects in multitrait-multimethod data: A review. In: Shrout, Patrick & Fiske, Susan (Eds.), Personality research, methods, and theory. A festschrift honoring D. W. Fiske. Hillsdale, NJ: Erlbaum, pp. 93-109.

Mislevy, Robert J. & Verhelst, Norman (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrica, 55, pp. 195-215.

Mohr, John & DiMaggio, Paul (1995). The intergenerational transmission of cultural capital. Research in Social Stratification and Mobility, 14, pp. 167-199.

Mortimer, Duncan & Segal, Leonie (2008). Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights. Medical Decision Making, 28/1, pp. 66-89.

Müller, Walter (1998). Erwartete und unerwartete Folgen der Bildungsexpansion. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 50, Sonderheft 38, S. 81-112.

Muraki, Eiji (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16/2, pp. 159-176.

Muraki, Eiji & Carlson, James (1995). Full-information factor analysis for polytomous item responses. Applied Psychological Measurement, 19, pp. 73-90.

Muthén, Bengt (1998-2004). Mplus Technical Appendices. Los Angeles, CA: Muthén & Muthén.

Muthén, Bengt & Muthén, Linda (1998-2010). Mplus. User manual. Los Angeles, CA: Muthén & Muthén.

Muthén, Linda K. & Muthén, Bengt O. (1998-2010). Mplus user's guide. Los Angeles, CA: Muthen & Muthen.

National Science Education Standards (NSES) (1996). National Research Council, National Committee on Science Education Standards and Assessment. Washington, DC.

Neuberg, Steven L. & Newsom, Jason T. (1993). Personal need for structure: Individual differences in the desire for simple structure. Journal of Personality and Social Psychology, 65, pp. 113-131.

Nölke, Christina; Lankes, Eva-Maria; Steffensky, Mirjam & Carstensen, Claus H. (2011). Naturwissenschaftsbezogenes Interesse bei Vorschulkindern im Rahmen des SNaKE-Projekts. Vortrag auf der SGL-Tagung, PH Bern, August 2011.

Norris, Steven P. & Philipps, Linda M. (2003). How literacy in its fundamental sense is central to scientific literacy. Science Education, 87/2, pp. 224-240.

Novick, Melvin R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, pp. 1-18.

Nussbeck, Fridtjof, Eid, Michael, Geiser, Christian, Courvoisier, Delphine & Lischetzke, Tanja (2009). A CTC(M–1) model for different types of raters. Methodology, 5, pp. 88-98.

Nussbeck, Fridtjof; Eid, Michael & Lischetzke, Tanja (2006). Analyzing MTMM data with SEM for ordinal variables applying the WLSMV-estimator: What is the sample size needed for valid results? British Journal of Mathematical and Statistical Psychology, 59, pp. 195-213.

Nylund, Karen L.; Muthén, Bengt; Nishina, Adrienne; Bellmore, Amy & Graham, Sandra (2006). Stability and instability of peer victinisation during middle school: using latent transition analysis with covariates, distal outcomes, and modeling extensions. URL (Accessed 23.05.2009): http://www.statmodel.com/download/LTA_DP_FINAL.pdf

OECD (2005). PISA 2003 technical report. Paris: OECD.

Oliveri, Maria E. & Davier, Matthias v. (2010). Investigation of model fit and score scale comparability in international assessments. Paper presented at the 17th International Meeting of the Psychometric Society, Athens, Georgia.

Osborne, Jonathan & Dillon, Justin (2008). Science Education in Europe: Critical reflections. A report to the Nuffield Foundation. London: Kings College.

Peine, Margit & Schönfeld, Helmut (1981). Sprachliche Differenzierung und ihre Bewertung. In: Hartung, Wolfdietrich & Schönfeld, Helmut (Hrsg.), Kommunikation und Sprachvariation. Berlin: Akademie-Verlag, S. 214-258.

Peisert, Hansgert (1967). Soziale Lage und Bildungschancen in Deutschland. München: Pieper.

Pickard, Simon; Kohlmann, Thomas; Janssen, Mathieu F.; Bonsel, Gouke; Rosenbloom Sarah & Cella, David (2007). Evaluating equivalency between response systems: Application of the Rasch model to a 3-level and 5-level EQ -5D. Medical Care, 45/9, pp. 812-819.

Piontkowski, Ursula; Öhlschlegel-Haubrock, Sonja & Hölker, Paul (1997). Annäherung oder Abgrenzung? Ergebnisse einer Längsschnittstudie zur Wirksamkeit der Ost-West-Kategorie. Zeitschrift für Soziolgie, 26, S. 128-138.

Plener, Paul; Libal, Gerhard; Keller, Ferdinand; Fegert, Jörg M. & Muehlenkamp, Jennifer (2009). An international comparison of adolescent Non-Suicidal Self-Injury (NSSI) and suicide attempts: Germany and U.S. Psychological Medicine, 39, pp. 1549-1558.

Pollack, Detlef (1998). Ostdeutsche Identität – ein multidimensionales Phänomen. In: Meulemann, Heiner (Hrsg.), Werte und nationale Identität im vereinten Deutschland. Opladen: Leske + Budrich, S. 301-318.

Pössel, Patrick; Horn, Andrea B. & Hautzinger Martin (2006). Vergleich zweier schulbasierter Präventionsprogramme von depressiven Symptomen bei Jugendlichen. Zeitschrift für Klinische Psychologie und Psychotherapie, 35, S. 109-116.

Pötter, Ulrich & Rendtel, Ulrich (1993). Über Sinn und Unsinn von Repräsentativitätssstudien. Allgemeines Statistisches Archiv, 77, S. 260-280.

Prenzel, Manfred (2000). Lernen über die Lebensspanne aus einer domänenspezifischen Perspektive. In: Achtenhagen, F. & Lempert, W. (Hrsg.), Lebenslanges Lernen im Beruf – seine Grundlegung im Kindes- und Jugendalter. Bd. 4: Formen und Inhalte von Lernprozessen. Opladen: Leske + Budrich, S. 175-192.

Prucha, Jan (2001). Finnland.
URL (Zugriff 04.01.2011): www2.dipf.de/publikationen.tibi/sonder_gesamt_1.pdf

Qualifications and Curriculum Authority (QCA) (2000). Department for Education and Employment (DfEE). Science: The National Curriculum for England: Key Stages 1-4.

Rabin, Rosalind & de Charro, Frank (2001). EQ-5D: a measure of health status from the EuroQol Group. Annals of Medicine, 33/5, pp. 337-343.

Radloff, Lenore Sawyer (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1, pp. 385-401.

Rasch, Georg (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Rasch, Georg (1961). On general laws and the meaning of measurement in psychology. Berkeley: University of California Press.

Rasch, Georg (1980). Probabilistic models for some intelligence and attainment tests. Expanded edition with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rauch, Wolfgang A.; Schweizer, Karl & Moosbrugger, Helfried (2008). An IRT analysis of the Personal Optimism Scale. European Journal of Psychological Assessment, 24, pp. 49-56.

Raudenbusch, Stephen W. & Bryk, Anthony S. (2002). Hierarchical linear models. Thousand Oaks, CA: Sage.

Rausch, Andreas (2012). Erleben und Lernen am Arbeitsplatz in der betrieblichen Ausbildung. Wiesbaden: VS-Verlag.

Rausch, Andreas; Scheja, Susanne; Dreyer, Kristina; Warwas, Julia & Egloffstein, Marc (2010). Emotionale Befindlichkeit in Lehr-Lern-Prozessen – Konstruktverständnis und empirische Zugänge. In: Seifried, Jürgen; Wuttke, Eveline; Nickolaus, Rainer & Sloane, Peter F. E. (Hrsg.), Lehr-Lern-Forschung in der kaufmännischen Berufsbildung – Ergebnisse und Gestaltungsaufgaben. Zeitschrift für Berufs- und Wirtschaftspädagogik, Beiheft 23, S. 193-215.

Regener, Irena (2000). Kontinuität, Anpassung, Selbstbehauptung. Zum Verhältnis zwischen tradierten Unterschieden und neuen Abgrenzungsidentitäten in der berliner Sprachge-

meinschaft der 90er Jahre. Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 48. URL: http://kops.ub.uni-konstanz.de/volltexte/2000/454/

Revuelta, Javier (2008). Estimating the π^* goodnes of fit index for finite mixtures of item response models. British Journal of Mathematical and Statistical Psychology, 61, pp. 93-113.

Reynolds, Henry T. (1977). The analysis of cross-classifications. London: The Free Press.

Rheinberg, Falko (2008). Motivation (7., aktualisierte Auflage). Stuttgart: Kohlhammer.

Rijmen, Frank & De Boeck, Paul (2003). A latent class model for individual differences in the interpretation of conditionals. Psychological Research, 67/3, pp. 219–231.

Rijmen, Frank; Tuerlinckx, Francis; De Boeck, Paul & Kuppens, Peter (2003). A nonlinear mixed model framework for item response theory. Psychological Methods, 8, pp. 185-205.

Rippetoe, Patricia A. & Rogers, Ronald W. (1987). Effects of components of protection-motivation theory on adaptive and maladaptive coping with a health threat. Journal of Personality & Social Psychology, 52/3, pp. 596-604.

Rizopoulos, Dimitris (2006). ltm: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software, 17/5, pp. 1-25.

Rizopoulos, Dimitris (2010). Latent trait models under IRT.
URL : http://rwiki.sciviews.org/doku.php?id=packages:cran:ltm

Roberts, James S. (2001). GGUM2000: Estimation of parameters in the generalized graded unfolding model. Applied Psychological Measurement, 25, p. 38.

Roberts, James S.; Rost, Jürgen & Macready, George B. (2010). MIXUM: An unfolding mixture model to explore the latitude of acceptance concept in attitude measurement. In: Embretson, Susan E. (Ed.), Measuring psychological constructs: Advances in model-based approaches. Washington, DC: American Psychological Association, pp. 175-197.

Robinson, Robert V. & Garnier, Maurice (1985). Class reproduction among men and women in France. American Journal of Sociology, 91, pp. 250-280.

Roeder, Kathryn & Wassermann, Larry (1997). Practical Bayesian density estimation using mixtures of normal. Journal of the American Statistical Association, 92, pp. 894-902.

Rogers, Ronald W. (1983). Protection motivation theory. Journal of Social Psychology, 211, pp. 223-244.

Ros, Maria (1984). Speech attitudes to speakers of language varieties in a bilingual situation. International Journal of the Sociology of Language, 47, pp. 73-90.

Roscigno, Vincent J. & Ainsworth-Darnell, James (1999). Race, cultural capital, and educational resources: persistent inequalities and achievement returns. Sociology of Education, 72, pp. 158-178.

Rössel, Jörg & Beckert-Zieglschmid, Claudia (2002). Die Reproduktion kulturellen Kapitals. Zeitschrift für Soziologie, 31, S. 497-613.

Rost, Jürgen (1988a). Measuring attitudes with a threshold model drawing on a traditional scaling concept. Applied Psychological Measurement, 12, pp. 397-409.

Rost, Jürgen (1988b). Quantitative und qualitative probabilistische Testtheorie. Bern: Huber.

Rost, Jürgen (1988c). Test theory with qualitative and quantitative latent variables. In: Langeheine, Rolf & Rost, Jürgen (Eds.), Latent trait and latent class models. New York: Plenum Press, pp.147-171.

Rost, Jürgen (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, pp. 271-282.

Rost, Jürgen (1991). A logistic mixture distribution model for polytomous item responses. British Journal of Mathematical and Statistical Psychology, 44, pp. 75-92.

Rost, Jürgen (1996, 2004). Lehrbuch Testtheorie – Testkonstruktion. Bern: Huber.

Rost, Jürgen (1996a). Theorien menschlichen Umwelthandelns. In: Michelsen, Gerd (Hrsg.), Handbuch zur Umweltberatung. Bonn: Economica, S. 55-62.

Rost, Jürgen (1999). Was ist aus dem Rasch-Modell geworden? Psychologische Rundschau, 50, S. 140-156.

Rost, Jürgen (2002). When personality questionnaires fail to be unidimensional. Psychologische Beiträge, 44, pp. 108-125.

Rost, Jürgen; Carstensen, Claus H. & Davier, Matthias v. (1999). Sind die Big Five Raschskalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten. Diagnostica, 45, S. 119-127.

Rost, Jürgen & Davier, Matthias v. (1993). Measuring different traits in different populations with the same items. In: Steyer, R.; Wender, K. F. & Widaman, K. F. (Eds.), Psychometric Methodology. Proceedings of the 7th European Meeting of the Psychometric Society in Trier. Stuttgart: Gustav Fischer, pp. 446-450.

Rost, Jürgen & Davier, Matthias v. (1995). Mixture distribution Rasch models. In: Fischer, Gerhard H. & Molenaar, I. W. (Eds.), Rasch models – Foundations, recent developments and applications. New York: Springer, pp. 257-268.

Rost, Jürgen; Gresele, Christiane & Martens, Thomas (2001). Handeln für die Umwelt – Anwendung einer Theorie. Münster: Waxmann.

Rost, Jürgen & Langeheine, Rolf (1997). Applications of latent trait and latent class models in the social sciences. Münster: Waxmann.

Rost Jürgen & Luo, Guanzhong (1997). An application of a Rasch-based unfolding model to a questionnaire on adolescent centrism. In: Rost, Jürgen & Langeheine, Rolf (Eds.), Applications of latent trait and latent class models in the social sciences. Münster: Waxmann.

Rost, Jürgen & Walter, Oliver (2006). Multimethod item response theory. In: Eid, Michael & Diener, Ed (Eds.), Handbook of multimethod measurement in psychology. Washington, DC: American Psychological Association, pp. 249-268.

Rubinstein, Sergej L. (1973, 1981). Probleme der allgemeinen Psychologie. Darmstadt: Dr. Dietrich Steinkopff.

Rudas, Tamas (2002). A latent class approach to measuring the fit of a statistical model. In: Hagenaars, Jacques A. & McCutcheon, Allan L. (Eds.), Applied latent class analysis. Cambridge, MA: Cambridge University Press, pp. 345-365.

Rudas, Tamas; Clogg, Clifford C. & Lindsey, Bruce G. (1994). A new index of fit based on mixture methods for the anlysis of contingency tables. Journal of the Royal Statitical Society, Series B, 56, pp. 623-639.

Russell, J. A. & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. Psychological Bulletin, 125, pp. 3-30.

Samejima, Fumiko (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement, 17.

Sander, Kerstin; Roth, Patricia & Scheich, Henning (2003). Left-lateralized fMRI activation in the temporal lobe of high repressive women during the identification of sad prosodies. Cognitive Brain Research, 16/3, pp. 441-456.

Santor, Darcy A. & Coyne, James C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. Psychological Assessment, 9, pp. 233-243.

Schiel, Stefan; Hess, Doris & Gilberg, Reiner (2003). Alterssurvey – Die zweite Lebenshälfte – Methodenbericht zur Erhebung der 2. Welle 2002. Bonn: infas GmbH.

Schimpl-Neimanns, Bernhard (2000). Soziale Herkunft und Bildungsbeteiligung zwischen 1950 und 1989. Kölner Zeitschrift für Psychologie und Sozialpsychologie, 52, S. 636-669.

Schmidt-Regener, Irena (1998a). Von der Akzeptanz des Berlinischen, von Liberalisierungstendenzen und Berührungsängsten. Spracheinstellungen in der berliner Sprachgemeinschaft nach 1989. In: Reiher, Ruth & Kramer, Undine (Hrsg.), Sprache als Mittel von Identifikation und Distanzierung. Frankfurt/M., Berlin, Bern, New York, Paris, Wien: Peter Lang, S. 153-185.

Schmidt-Regener, Irena (1998b). Distanz und Nähe. Soziolinguistische Aspekte deutscher Identität(en) nach 1989. In: Kempf, Wilhelm & Schmidt-Regener, Irena (Hrsg.), Krieg, Nationalismus, Rassismus und die Medien. Münster: LIT, S. 181-194.

Schöffski, Oliver & von der Schulenburg, J.-Matthias Graf (2007). Gesundheitsökonomische Evaluationen. Berlin: Springer.

Schulmeister, Rolf (1983). Angst vor Statistik: Empirische Untersuchungen zum Problem des Statistik-Lehrens und -Lernens. Hamburg: Arbeitsgemeinschaft für Hochschuldidaktik.

Schulze, Gerhard (1992). Die Erlebnisgesellschaft. Frankfurt/M.: Campus.

Schumacher, Jörg; Klaiberg, Antje & Brähler, Elmar (2003). Diagnostische Verfahren zu Lebensqualität und Wohlbefinden. Göttingen: Hogrefe.

Schurer, Bruno (1984). Gegenstand und Struktur der Lernhandlung. Ein Beitrag zu einer lernerzentrierten Didaktik unter besonderer Berücksichtigung des arbeitsmotorischen Lernens. Bergisch-Gladbach: Thomas Hobein.

Schwartz, Shalom H. (1973). Normative explanations of helping behavior: A critique, proposal, and empirical test. Journal of Experimental Social Psychology, 9/4, pp. 349-364.

Schwartz, Shalom H. (1977). Normative influences on altruism. In: Berkowitz, Leonard (Ed.), Advances in experimental social psychology. Vol. 10. New York: Academic Press, pp. 221-279.

Schwartz, Shalom H. & Howard, Judith A. (1981). A normative decision-making model of altruism. In: Rushton, J. P. & Sorrentino, R. M. (Eds.), Altruism and helping behavior. Hillsdale: Lawrence Erlbaum, pp. 189-211.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. Annals of Statistics, 6, pp. 461-464.

Schwarzer, Ralf (1994). Volitionstheorie der Gesundheitserziehung. Zeitschrift für Pädagogik, 40/6, S. 907-922.

Shafer, Alan B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. Journal of Clinical Psychology, 62, pp. 123-146.

Shockey, James W. (1988). Latent class analysis: an introduction to discrete data models with unobserved variables. In: Long, J. Scott (Ed.), Common problems/proper solutions: Avoiding error in quantitative research. Berverly Hills, CA: Sage, pp. 288-315.

Smit, J. A., Kelderman, Henk & Flier, Henk van der (2000). The mixed Birnbaum model: Estimation using collateral information. Methods of Psychological Research Online, 5, pp. 1-13.

Smit, J. A., Kelderman, Henk & Flier, Henk van der (1999). Collateral information and mixed Rasch models. Methods of Psychological Research Online, 4/3, pp. 1-13.

Sonntag, Karlheinz (2007). Theorien der Arbeitstätigkeit. In: Schuler, Heinz & Sonntag, Karlheinz (Hrsg.), Handbuch der Arbeits- und Organisationspsychologie. Göttingen: Hogrefe, S. 35-42.

Stansbury, James P.; Ried, L. Douglas & Velozo, Craig A. (2006). Unidimensionality and bandwidth in the Center for Epidemiological Studies Depression (CES-D) Scale. Journal of Personality Assessment, 86, pp. 10-22.

Steffensky, Mirjam; Lankes, Eva-Maria & Carstensen, Claus H. (2012a). Was bedeutet naturwissenschaftliche Kompetenz bei Fünfjährigen und wie kann man sie erfassen? In: Gläser-Zikuda, Michaele; Seidel, Tina; Rohlfs, Carsten; Gröschner, Alexander & Ziegelbauer, Sascha (Hrsg.), Mixed Methods in der empirischen Bildungsforschung. Münster et al.: Waxmann.

Steffensky, Mirjam; Lankes, Eva-Maria; Carstensen, Claus H. & Nölke, Christina (2012b). Alltagssituationen und Experimente: Was sind bessere Lerngelegenheiten für Kindergartenkinder? Ergebnisse aus dem SNaKE-Projekt. Zeitschrift für Erziehungswissenschaft, DOI 10.1007/s11618-012-0262-3.

Steffensky, Mirjam; Lankes, Eva-Maria; Carstensen, Claus H.; Nölke, Christina & Kuhn, Nicole (2010). SNaKE – Untersuchung naturwissenschaftlicher Kompetenz im Kindergarten, Vortrag vor der Gesellschaft für die Didaktik der Chemie und Physik, 13.-16.09.2010, Potsdam.

Stern, Paul C.; Dietz, Thomas & Kalof, Linda (1993). Value orientations, gender, and environmental concern. Environment and Behavior, 25/3, pp. 322-348.

Stob, Michael (1984). A supplement to A Mathematician's Guide to Popular Sports. The American Mathematical Monthly, 91/5, pp. 277-282.

Streiner, David L. & Norman, Geoffrey R. (1989). Health measurement scales. Oxford: Oxford University Press.

Sullivan, Alice (2001). Cultural capital and educational attainment. Sociology, 35, pp. 893-912.

Suppes, Patrick & Zanotti, Mario (1981). When are probabilistic explanations possible? Synthese, 48, pp. 191-199. Reprinted in: Suppes, P. & Zanotti, M. (1995). Foundations of probability with applications: Selected Papers, 1974-1995. Cambridge: Cambridge University Press.

Szende, Agota; Oppe, Mark & Devlin, Nancy (2007). EQ-5D value sets. Dordrecht: Springer.

Takane, Yoshio & de Leeuw, Jan (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52, pp. 393-408.

Thomas, Michael L. (2010). The value of item response theory in clinical assessment: A review. Assessment, published online July 19, pp. 1-17.

Thompson, Megan M.; Naccarato, Michael E.; Parker, Kevin C. H. & Moskowitz, Gordon B. (2001). The personal need for structure and personal fear of invalidity measures: Historical perspectives, current applications, and future directions. In: Moskowitz, Gordon

B. (Ed.), Cognitive social psychology: The Princeton symposium on the legacy and future of social cognition. Mahwah, NJ: Erlbaum, pp. 19-39.

Van de Pol, Frank; Langeheine, Rolf & de Jong, Wil (2000). PANMARK 3 user's manual: PANel analysis using MARKov chains, a latent class analysis program. St. Paul, MN: Assessment Systems Corporation.

Van de Vijver, Fons & Tanzer, Norbert K. (2004). Bias and equivalence in cross-cultural assessment: an overview. Revue euopéenne de psychologie appliquée, 54, pp. 119-135.

Van Hout, Ben; Janssen, Bas; Kohlmann, Thomas; Feng, You-Shan; Golicki, Dominik; Scalone, Luciana; van Busschbach, Jan; Lloyd, Andrew; Kind, Paul & Pickard, Simon (2011). Value sets for the EQ-5D-5L: A mapping approach. [submitted]

Vautier, Stéphane & Pohl, Steffi (2009). Do balanced scales assess bipolar constructs? The case of the STAI scales. Psychological Assessment, 21, pp. 187-193.

Velmans, Max (1991). Is human information processing conscious?
URL (Zugriff 13.05.2011): http://cogprints.org/593

Vermunt, Jeroen K. (1997). LEM: A general program for the analysis of categorical data. User's manual. Tilburg, The Netherlands: Tilburg University.

Vermunt, Jeroen K. (2003). Multilevel latent class models. Sociological Methodology, 33, pp. 213-239.

Vermunt, Jeroen K. & Magidson, Jay (2003, 2005). Latent GOLD user's guide. Belmont, MA: Statistical Innovations Inc.

Vermunt, Jeroen K. & Magidson, Jay (2006). LatentGOLD 4.0 and IRT modeling.
URL (Zugriff: 07.10.2011): http://www.statisticalinnovations.com/products/LGIRT.pdf

Vester, Michael (2004). Die Illusion der Bildungsexpansion. Unveröffentlichtes Manuskript.

Walter, Oliver; Senkbeil, Martin; Rost, Jürgen; Carstensen, Claus H. & Prenzel, Manfred (2006). Die Entwicklung der naturwissenschaftlichen Kompetenz von der neunten zur zehnten Klassenstufe: Deskriptive Befunde. In: Prenzel, M.; Baumert, J.; Blum, W.; Lehmann, R.; Leutner, D.; Neubrand, M.; Pekrun, R.; Rost, J. & Schiefele, U. (Hrsg.), PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres. Münster: Waxmann, S. 309-323.

Warm, Thomas A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54, pp. 427-445.

Watermann, Rainer & Baumert, Jürgen (2006). Entwicklung eines Strukturmodells zum Zusammenhang zwischen sozialer Herkunft und fachlichen und überfachlichen Kompetenzen. Befunde national und international vergleichender Analysen. In: Baumert, J.; Stanat, P. & Watermann, J. (Hrsg.), Herkunftsbedingte Disparitäten im Bildungswesen. Vertiefende Analysen im Rahmen von PISA 2000. Wiesbaden: VS-Verlag.

Watson, David; Hubbard, Brock & Wiese, David (2000). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. Journal of Personality and Social Psychology, 78, pp. 546-558.

Watson, David & Tellegen, A. (1985). Toward a consensual structure of mood. Psychological Bulletin, 98, pp. 219-235.

Weber, Max (1972). Wirtschaft und Gesellschaft. Tübingen: Mohr.

Weijters, Bert; Geuens, Maggie & Schillewaert, Niels (2010). The stability of individual response styles. Psychological Methods, 15, pp. 96-110.

Weinert, Sabine; Artelt, Cordula; Prenzel, Manfred; Senkbeil, Martin; Ehmke, Timo & Carstensen Claus H. (2011). Development of Competencies Across the Life Span. In: Blossfeld, H. P.; Roßbach, H. G. & Maurice, J. v. (Eds.), Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14.

Weinert, Sabine; Doil, Hildegard & Frevert, Sabine (2008). Kompetenzmessungen im Vorschulalter: eine Analyse vorliegender Verfahren. In: Rossbach, H.-G. & Weinert, S. (Hrsg.), Kindliche Kompetenzen im Elementarbereich: Förderbarkeit, Bedeutung, Messung. Berlin: Bundesministerium für Bildung und Forschung, S. 89-209.

Whitely, Susan. E. & Nieh, K. (1981). Program LINLOC (Tech. Rep. NIE-81-3 for National Institute of Education). Lawrence, KS: University of Kansas.

Wilson, Mark (1992). The ordered partition model: An extension of the partial credit model. Applied Psychological Measurement, 16, pp. 309-325.

Wright, Benjamin D. & Masters, Geoffrey N. (1982). Rating scale analysis: Rasch Measurment. Chicago: MESA Press.

Wu, Margaret L.; Adams, Raymond J. & Wilson, Mark R. (1997). ConQuest. Generalised item response modelling software. Melbourne: Australian Council for Educational Research.

Wu, Margaret L.; Adams, Raymond J.; Wilson, Mark R. & Haldane, Sam (2007). ACER ConQuest 2.0 – Generalised item response modelling software. Camberwell: ACER Press.

Zermelo, Ernst (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift, 29, S. 436-460.

Zickar, Michael J. & Drasgow, Fritz (1996). Detecting faking on a personality instrument using appropriateness measurement. Applied Psychological Measurement, 20, pp. 71-87.

Zimmerman, Barry J. (1998). Academic studying and the development of personal skill: A selfregulatory perspective. Educational Psychologist, 33, pp. 78-86.

Über die Autoren

Carstensen, Claus H., Prof. Dr., geb. 1967. Professor für Psychologie mit Schwerpunkt Methoden empirischer Bildungsforschung, Otto-Friedrich-Universität Bamberg. Arbeitsschwerpunkte: Methoden empirischer Bildungsforschung, insbesondere Item-Response-Theorie, Methoden von Large Scale Assessments & Längsschnittmodellierung, Kompetenzdiagnostik über die Lebensspanne und international vergleichende Bildungsforschung.

Kontakt: eMail: claus.carstensen@uni-bamberg.de
Website: http://www.uni-bamberg.de/psymethodenbf/

Davier, Matthias von, Dr. rer.nat., geb. 1965. Research Director at ETS, Princeton, NJ, USA & Honorary Senior Research Fellow, Oxford University, England. Arbeitsschwerpunkte: Research in Statistics and Psychometrics, Item Response Theory, Diagnostic Classification Models, Latent Structure Models, Mixture Models, Large Scale Assessment Methodology.

Kontakt: ETS, MS 12E, Princeton, NJ 08541, USA
eMail: mvondavier@ets.org
Website: http://www.von-davier.com

Eid, Michael, Prof. Dr., geb. 1963. Professor für Methoden und Evaluation im Fachbereich Erziehungswissenschaften und Psychologie an der FU Berlin. Arbeitsschwerpunkte: Veränderungsmessung, Multitrait-Multimethod-Analysen, Strukturgleichungsmodelle, Item-Response-Theorie, multimethodale Erfassung von Persönlichkeit, Wohlbefinden, Persönlichkeit und Emotionen, Präventionsmaßnahmen gegen Hautkrebs.

Kontakt: FU Berlin, FB Erziehungswissenschaften und Psychologie
Habelschwerdter Allee 45, 14195 Berlin
eMail: eid@zedat.fu-berlin.de
Website: http://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/psymeth/mitarbeiter/meid/index.html

Feng, You-Shan, PhD, geb. 1981. Wissenschaftliche Mitarbeiterin in der Abteilung Methoden der Community Medicine, Universität Greifswald. Arbeitsschwerpunkte: Messung der gesundheitsbezogenen Lebensqualität, epidemiologische Methoden,

patient-reported outcomes (PROs) für Patienten mit Erkrankungen des Bewegungsapparates.

Kontakt: Institut für Community Medicine, Walther-Rathenau-Str. 48, 17475 Greifswald
eMail: You-shan.feng@unigreifswald.de
Website: www.medizin.uni-greifswald.de/icm/

Georg, Werner, Prof. Dr., geb. 1953. Professor für Empirische Sozialforschung mit Schwerpunkt Bildungssoziologie am Fachbereich Geschichte und Soziologie der Universität Konstanz. Seine Arbeitsschwerpunkte sind neben den genannten Bereichen Sozialstrukturanalyse und soziale Ungleichheit sowie Lebensstilforschung.

Kontakt: eMail: werner.georg@uni-konstanz.de

Hoeffler, David, Dipl.-Psych., geb. 1984. IT-Berater für statistische Analysen im Personalbereich. Arbeitsschwerpunkte: Beratung von Unternehmen bez. Einführung und Konfiguration statistischer Analysesysteme im Personalbereich sowie Konzeption und Einführung von Prozessen zur strategischen Personalplanung.

Kontakt: SAP Deutschland, Hasso-Plattner-Ring 7, 69190 Walldorf
eMail: david.hoeffler@sap.com

M. F. (Bas) Janssen, PhD, geb. 1974. Principal Researcher, EuroQol Group. Arbeitsschwerpunkte: Qualitätsadjustierte Lebensjahre (QALYs), gesundheitsbezogene Lebensqualität, psychometrische Methodik.

Kontakt: Rotterdam EuroQol Group Executive Office, PO Box 4443, 3006 AK Rotterdam
eMail: janssen@euroqol.org
Website: www.euroqol.org

Jasper, Fabian, Dipl.-Psych. Dr., 1982. Wissenschaftlicher Angestellter. Arbeitsschwerpunkte: Somatoforme Störungen, Krankheitsangst, Mathematikfähigkeiten.

Kontakt: Universität Mainz, Wallstr. 3, 55122 Mainz
eMail: jasper@uni-mainz.de
Website: http://www.klinische-psychologie-mainz.de

Keller, Ferdinand, Prof. Dr., geb. 1956. Leitender Psychologe in der Klinik für Kinder- und Jugendpsychiatrie am Universitätsklinikum Ulm. Arbeitsschwerpunkte: Statistische Methoden der Veränderungsmessung in Psychotherapie, Psychiatrie und Präventionsstudien, insbesondere hierarchische lineare Modelle und growth mixture models, Fragebogenentwicklung und psychometrische Analysen, insbesondere IRT-Verfahren (Rasch- bzw. Mixed Rasch-Modelle, Latent-Class-Analyse), Depression, Behandlungszufriedenheit in der Kinder- und Jugendpsychiatrie, Evaluation von Therapieverfahren.

Kontakt: Klinik für Kinder- und Jugendpsychiatrie/Psychotherapie, Universitätsklinik Ulm
Steinhövelstr. 5, 89075 Ulm
eMail: Ferdinand.Keller@uniklinik-ulm.de
Homepage Klinik: http://www.uniklinik-ulm.de/index.php?id=1420

Kempf, Wilhelm, Prof. Dr. phil. habil., geb. 1947. Seit 1977 Professor für psychologische Methodenlehre und Leiter der Projektgruppe Friedensforschung an der Universität Konstanz. Seit 2002 Herausgeber der transdisziplinären Fachzeitschrift conflict & communication online (www.cco.regener-online.de). Arbeitsschwerpunkte: Methodologie und Methoden der Psychologie; psychologische Friedens- und Konfliktforschung.

Kontakt: Universität Konstanz, FB Psychologie, 78457 Konstanz
eMail: Wilhelm.Kempf@uni-konstanz.de
Website: http://www.pfkn.regener-online.de/

Kohlmann, Thomas, Prof. Dr., geb. 1953. Leiter der Abteilung Methoden der Community Medicine, Universität Greifswald. Arbeitsschwerpunkte: Epidemiologie von muskulo-skelettalen Erkrankungen, Entwicklung und Validierung von Erhebungsinstrumenten zur Messung der subjektiven Gesundheit, Konzepte und Methoden der gesundheitsökonomischen Evaluation.

Kontakt: Institut für Community Medicine, Walther-Rathenau-Str. 48, 17475 Greifswald
eMail: Thomas.Kohlmann@uni-greifswald.de
Website: http://www.medizin.uni-greifswald.de/icm/

Langeheine, Rolf, Dr., geb. 1937. Seit 2002 im Ruhestand. Von 1978 bis 2002 wissenschaftlicher Mitarbeiter der Abteilung Pädagogisch-Psychologische Methodenlehre des Leibniz-Instituts für die Pädagogik der Naturwissenschaften an der Universität Kiel. Arbeitsschwerpunkte: Analyse von kategorialen Daten mit loglinearen, latenten Klassen- und latenten Markov-Modellen.

Kontakt: Wildhofstr. 38, 24582 Bordesholm
eMail: abellangeheine@aol.com

Lankes, Eva-Maria, Prof. Dr., geb. 1953. Lehrstuhl für Schulpädagogik, TUM School of Education, TU München. Arbeitsschwerpunkte: Kompetenzentwicklung, Kompetenzmodellierung: naturwissenschaftliche Kompetenz im Elementarbereich, Lern- und Unterrichtsbedingungen im internationalen Vergleich, evidenzbasierte Schul- und Unterrichtsentwicklung.

Kontakt: eMail: lankes@tum.de
Website: http://www.schulpaed.edu.tum.de

Martens, Thomas, Dr., geb. 1967. Senior Researcher am Deutschen Institut für Internationale Pädagogische Forschung. Forschungsschwerpunkte: Motivationale Lern- und Handlungsregulation, E-Learning und Formatives Assessment.

Kontakt: Schloßstr. 29, 60486 Frankfurt am Main,
eMail: martens@dipf.de
Website: http://www.thomas-martens.de

Meiser, Thorsten, Prof. Dr., Dipl.-Psych., geb. 1967. Professor für Psychologische Methodenlehre und Diagnostik an der Universität Mannheim. Arbeitsschwerpunk-

te: Statistische Modellierung kognitiver Prozesse, experimentelle Gedächtnispsychologie, Item-Response-Theorie, Modellierung von Antwortstilen.

Kontakt: Universität Mannheim, Arbeitsgruppe Psychologische Methodenlehre und Diagnostik,
Schloss Ehrenhof, 68131 Mannheim
eMail: t.meiser@uni-mannheim.de
Website: http://methods.uni-mannheim.de/startseite/index.html

Nussbeck, Fridtjof W., Dr., geb. 1976. Oberassistent am Psychologischen Institut der Universität Zürich, Abteilung für Klinische Psychologie mit Schwerpunkt Kinder/ Jugendliche und Paare/Familien. Arbeitsschwerpunkte: Veränderungsmessung, Multitrait-Multimethod-Analysen, Strukturgleichungsmodelle, log-lineare Modelle, Beurteilerübereinstimmung, Analyse von dyadischen und Familiendaten.

Kontakt: Universität Zürich, Psychologisches Institut
Binzmühlestrasse 14 / Box 23, 8050 Zürich, Schweiz
eMail: fridtjof.nussbeck@psychologie.uzh.ch
Website: http://www.psychologie.uzh.ch/fachrichtungen/kjpsych/team1/nussbeck.html

Pickard, A. Simon, Prof. Ph.D. Associate Professor, Departments of Pharmacy Practice and Pharmacy Administration and Assistant Director, Center for Pharmacoepidemiology and Pharmacoeconomic Research, University of Illinois at Chicago. Arbeitsschwerpunkte: Methoden und Instrumente zur Messung von patient-reported outcomes (PROs), medizinische Entscheidungsfindung, Gesundheitsökonomie, Pharmakoepidemiologie, Arzneimittelpolitik.

Kontakt: Department of Pharmacy Practice (MC 886), College of Pharmacy
833 South Wood Street, RM164 MC886, Chicago, IL 60612, USA
eMail: pickard1@uic.edu
Website: www.uic.edu/pharmacy/depts/Pharmacy_Practice

Regener, Irena, Dr. phil., geb. 1956. Verlegerin. Arbeitsschwerpunkte: Soziolinguistik, Geschichte der sprachwissenschaftlichen Germanistik.

Kontakt: verlag irena regener berlin, Ostseestr. 109, 10409 Berlin
eMail: regener@regener-online.de
Website: http://www.regener-online.de

Steffensky, Mirjam, Prof. Dr., geb. 1970. Didaktik der Chemie am Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik. Arbeitschwerpunkte: Modellierung und Förderung von naturwissenschaftlicher Kompetenz im Elementar- und Primarbereich sowie im naturwissenschaftlichen Anfangsunterricht, fachbezogenes Professionswissen von Erzieher/-innen und Grundschullehrpersonen; Entwicklung und Evaluation von Materialien für den Elementar- und Primarbereich.

Kontakt: eMail: steffensky@ipn.uni-kiel.de
Website: http://www.ipn.uni-kiel.de/persons/steffensky.html

Sachregister

Akaike's Information Criterion (AIC) 43, 45, 47, 49ff., 57, 60-64, 89, 92, 97f., 100, 103, 177, 179, 181, 184, 199, 201
Allgemeine Depressionsskala (ADS) 12, 171, 173-178, 180-187
Antisemitismus 10, 43-47
Antwortmuster 35f., 39ff., 51f., 56-59, 64, 117, 172f., 183, 191, 193f., 197, 200
Antworttendenz 150, 168

Bayesian Information Criterion (BIC) 49ff., 57, 60-64, 76ff., 117ff., 165-168, 177ff., 181, 184ff., 199, 201
Bildungssoziologie 68
CTC(M-1)-Modell 142f., 145-148

Daten, kategoriale 18, 129f.
Depression 171, 174, 179, 181f., 185, 188
Diskriminationsparameter 126, 195f., 199f., 202-205, 208
Dispersions-Modell (DM) 154f.
Disutilities 193f., 196ff., 203, 205f.

Eindimensionalität 27f.
Einstellung 19, 32, 35, 37, 46, 83-88, 91-96, 100, 106ff.
Einstellungsmessung 10, 32, 35, 39, 50
Entropie 41, 51f.
EQ-5D-Fragebogen 12, 191-195, 197f., 200ff., 204f., 208
Evaluation, gesundheitsökonomische 189, 208
Explanatory-Power-Index (EP) 41, 46

Faktorenanalyse, konfirmatorische 129f., 148, 171
Fragebogen 12, 33ff., 37, 43f., 47, 51, 86f., 91,97, 150f., 154, 156ff., 159, 161, 164f., 168f., 175, 177, 186, 190
Fremdbericht 130f., 134-146, 148

Gesundheitsforschung 12, 209
Gesundheitszustand 189f., 192f., 198f., 203, 206ff.
Goodness-of-Fit-Statistik 43, 45, 47, 56, 61, 63f., 89, 92, 98, 101, 103, 134, 166, 174, 199
Graded-Response-Modell 11, 133f.

Habitus 70f., 73
Hauptkomponentenanalyse 64
Homogenität 32-35, 40, 43ff., 51, 56, 59, 75, 88, 113, 115, 117, 125, 167f., 177, 181
Hybrid-Modell (HM) 39, 173f.

Idealtypus 69
Identität 11, 83, 85, 91, 96, 106ff.
Indikator 12, 16, 18, 32f., 76f., 81, 88, 131f., 139, 143, 146, 149, 175, 189, 193
Information, statistische 35, 38, 41-45, 50f.
Intentionsphase 211f., 214, 216ff., 221f.
International Socio Economic Index (ISEI) 10, 68, 75, 79
Itemanalyse 19, 129
Itemfunktion 24, 29, 59, 65
Items, invertierte 12, 150f., 155f., 158-169

Item-Response-Modell 9, 31f., 35ff., 40, 133
Item-Response-Theorie (IRT) 9, 11, 13, 129, 133, 169, 190
Itemschwierigkeit 28, 30, 111f., 117f., 120, 126, 150, 153f., 177
Itemscore 32f., 47f.
Intentionsgedächtnis 212, 217-221, 228

Kapital, kulturelles 10, 68, 70-75, 77, 79ff., 84
Klassengröße 39f., 76, 78, 106, 179, 181, 184f.
Koeffizient α 34f., 43f., 178
Konsistenz, interne 34f., 43f., 171

Längsschnittuntersuchung 11, 82, 109f., 113f., 118, 124f., 175, 187, 229
Latent-Class-Analyse (LCA) 11, 30f., 35, 37, 40-51, 73-81, 88-89, 101, 103, 173f., 177, 186, 225f.
Latent-Class-Modell 10, 38, 56, 60, 65f., 83, 156, 179
Latent-Trait-Modell 38
Lebensqualität 12, 189ff., 209
Lebensqualitätsmessung, präferenzbasierte 190
Lesekompetenz 68
Likelihood-Ratio 43ff., 47, 56ff., 60-63
Likert-Skala 32, 43f., 158
Loss-of-Information-Index (LI) 10, 42ff., 46f., 50f.
Lügenkriterium 172, 175-182, 185f.

Maximum-Likelihood-Methode 134, 165, 196
Messäquivalenz 134, 136, 138f., 141, 144
Messwiederholung 110, 112
Methoden
- austauschbare 129f., 135, 142f., 145, 147f.
- strukturell unterschiedliche 11, 129f., 142ff., 147ff.
Methodeneffekt 128f., 139, 147, 149, 175
Methodenspezifität 141f., 144f., 147
Mischverteilung 88
Mischverteilungsmodell 13, 31, 66, 156, 174, 228
Mischverteilungs-Rasch-Modell siehe Mixed-Rasch-Modell
Missing Data 44, 46ff., 51
Mixed-Rasch-Modell (MRM) 9, 15, 19f., 28-31, 39, 58, 60, 157f., 171, 174, 177f., 224
Mobilitätshypothese 72, 81
Modalwert 76
Modell
- gemeinsamer Faktoren 131-135
- loglineares 56, 60
- saturiertes 39f., 42f., 45-48, 57, 60, 117
Modellgütekoeffizient siehe Goodness-of-Fit-Statistik
Modelltest 55, 59
Modelltestindizes siehe Goodness-of-Fit Statistik
Monotonie 24f., 28, 153
Motivation 13, 109, 213-218, 220, 222, 225, 228f.
Motivierungsphase 211f., 214f., 222
Multitrait-Multimethod-Analyse 11, 127

Nutzwert 190, 192-195, 205-208

Objekterkennung 212, 214f., 217
Ordnungsrelation 33

Partial-Credit-Modell (PCM) 12, 116, 154ff., 161, 164-169, 177, 179, 181, 184, 195f., 199, 208
Persistenzhypothese 68f.
Personenfähigkeit 25f., 28, 30, 56, 111f., 116, 119, 126
PISA-Studie 10, 55, 65, 68, 72-76, 80, 125
pi*-Index 10, 55, 57-67
Polungseffekt 158, 161, 167-170
Populationsunabhängigkeit 9, 15, 23, 27ff.
Profillinie 33ff., 44f., 47f.
Pure-Random-Modell (PR) 37, 42f., 45ff.

Quasi-Experiment 109f., 113

Rasch-Modell (RM) 9-12, 15, 26ff., 30, 35, 38f., 55-66, 110, 112f., 115, 124f., 129, 150-158, 162, 165, 168ff, 173f., 177f., 190f., 195, 199-205, 208f., 224
Ratingskala 155, 157, 163f., 198
Rating-Scale-Dispersion-Modell (RSDM) 155ff.
Rating-Scale-Modell (RSM) 12, 150, 154ff., 161-166
- erweitertes (eRSM) 155, 161f., 166f., 169
- modifiziertes (mRSM) 155, 161-169
- symmetrisches (sRSM) 155, 163-168
Rational-choice-Paradigma 69f.
Redundanz 41, 52
Rekodierung 12, 150f., 156, 158-163, 165, 167ff.
Relevanzmaß 35, 40f., 51
Reliabilität 19, 34, 121-124, 128, 175, 178f., 181, 184f.
Reproduktion, soziale 70f.
Reproduktionshypothese 72, 81
response mapping 194f.
Rost-Modell siehe Mixed-Rasch-Modell

Schwellenabstand 12, 154-169
Schwellenwahrscheinlichkeit 151f., 158
Screening 177, 187
Selbstbericht 130-148
Selbstwirksamkeit 211, 217
Sozialprestige 84, 86, 88
Strukturgleichungsmodelle 73, 76, 81, 175, 210
Suffizienz 35
Summenscore 9f., 15-20, 22f., 32, 34f., 38, 40f., 44ff., 50f.

τ-Äquivalenz, essentielle 33f.
Testtheorie
- klassische 10, 32, 34, 36, 150, 168, 178, 185, 190, 193
- probabilistische 110, 150, 190, 193
- stochastische 35f.
TIMSS 10, 55, 60, 65
True-Score 20, 36

Unabhängigkeit
- bedingte, siehe U., lokale
- lokale 9, 20-23, 28f., 76, 133
Ungleichheit, soziale 68f., 71, 73f., 79ff.

Validität 175, 229
- diskriminante 11, 127ff., 147
- konvergente 11, 127ff., 134f., 140
Verantwortungsübernahme 211, 215f.
Verhaltenssteuerung 211, 219, 221
Volitionsphase 211-214, 216-222

WLSMV-Verfahren 134ff.

Zielverfolgung
- persistente 218
- selbstkongruente 211, 219f., 222

verlag irena regener berlin Wir setzen Qualitätsstandards jenseits des Mainstreams

Wilhelm Kempf

Forschungsmethoden der Psychologie

Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik

Band II: Quantität und Qualität

(Hochschullehrbücher, Bd. 3). 2008. 396 S., brosch., mit 47 Abbildungen und 59 Tabellen, € 44,90. ISBN 978-3-936014-15-0.

Gegenstand des zweiten Bandes der Forschungsmethoden der Psychologie sind zwei Teilgebiete der Methodenlehre, die auf geradezu entgegengesetzten Polen des Methodenspektrums angesiedelt sind: (1) Textinterpretation und Inhaltsanalyse, die dem Wissensideal der qualitativen Sozialforschung verpflichtet sind, und (2) Psychometrie, die von manifesten Beobachtungsdaten auf zugrunde liegende Eigenschaften der Probanden schließt und die wohl konsequenteste Umsetzung eines quantifizierenden Wissensideals in der Psychologie darstellt.
In didaktischer Hinsicht dem Prinzip des Lernens am konkreten Objekt verpflichtet, will das Buch den Leser in die Lage versetzen, die einschlägige Fachliteratur kritisch zu rezipieren, die Methoden in der eigenen Forschungspraxis anzuwenden und sie in gegenstandsangemessener Weise miteinander zu verbinden. Nicht nur die Anwendungsnähe der Methoden, sondern auch ihre theoretische und (im Falle der Psychometrie) mathematische Grundlegung stehen daher im Zentrum des Buches.

Aus dem Inhalt

1. Textinterpretation und Inhaltsanalyse: Qualitative Sozialforschung / Grundlagen der sozialpsychologischen Rekonstruktion / Textinterpretation / Qualitative und quantitative Inhaltsanalyse / Integration quantitativer und qualitativer Analysemethoden
2. Psychometrie: Grundlagen / Testen und Messen / Klassische und stochastische Testtheorie / Parameterschätzung, Modellkontrolle und Beurteilung der Testleistung der Probanden / Rasch-Modell / Verallgemeinerungen des Rasch-Modells / Latent-Class-Analyse / Modellvergleiche / Validität

verlag irena regener berlin
ostseestr. 109
10409 berlin

fax: #49.30.42 85 67 62
www.regener-online.de
eMail: verlag@regener-online.de

regener
publishers off mainstream

verlag irena regener berlin Wir setzen Qualitätsstandards jenseits des Mainstreams

Forschungsmethoden der Psychologie

Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik
(Hochschullehrbücher, Bd. 2-4)

In konsequenter Auffassung der Psychologie als Wissenschaft von der subjektiven Welt des Menschen, deren Gegenstandsbereich von den physiologischen Ursachen bis hin zu den Prozessen der sozialen Konstruktion von Wirklichkeit reicht, verfolgen die Autoren eine wissenschaftstheoretische und didaktische Konzeption, welche die Methoden der Psychologie im Spannungsfeld zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik verortet, eine gegenstandsangemessene Umsetzung des galileischen Denkens in der Psychologie zum Ziel hat und die Kluft zwischen natur- und kulturwissenschaftlicher Herangehensweise an den Forschungsgegenstand der Psychologie ebenso zu überwinden trachtet wie jene zwischen quantitativen und qualitativen Forschungsmethoden.

Band I:
Wilhelm Kempf, Theorie und Empirie
3. Auflage. 2009. 377 S., 377 S., brosch., mit 71 Abb. und 34 Tab., 42,90 €.
ISBN 978-3-936014-21-1.

Band II:
Wilhelm Kempf, Quantität und Qualität
1. Auflage. 2008. 396 S., 396 S., brosch., mit 47 Abb. und 59 Tab., 44,90 €.
ISBN 978-3-936014-15-0.

Band III:
Wilhelm Kempf & Markus Kiefer (Hrsg.), Natur und Kultur
1. Auflage. 2009. 366 S., 366 S., brosch., mit 63 Abb. und 8 Tab., 44,50 €.
ISBN 978-3-936014-17-4.

Alle drei Bände im Paket zum Sonderpreis: Statt 132,30 € nur 120,- €!
ISBN 978-3-936014-20-4.

verlag irena regener berlin
ostseestr. 109
10409 berlin

fax: #49.30.42 85 67 62
www.regener-online.de
eMail: verlag@regener-online.de