PHILOSOPHISCH-THEOLOGISCHE HOCHSCHULE VALLENDAR
Kirchlich und staatlich anerkannte Wissenschaftliche Hochschule in freier Trägerschaft
Pflegewissenschaftliche Fakultät

Albert Brühl (Hg.):

Pflegebedürftigkeit messen?

Herausforderungen bei der Entwicklung pflegerischer Messinstrumente am Beispiel des Neuen Begutachtungsassessments (NBA)

Wissenschaftlicher Bericht

des Lehrstuhls für Statistik und standardisierte Verfahren der Pflegeforschung

Juli 2012

Herausgeber:

Prof. Dr. Albert Brühl

Lehrstuhl für Statistik und standardisierte Verfahren der Pflegeforschung

Philosophisch-Theologische Hochschule Vallendar

Pflegewissenschaftliche Fakultät

Pallottistrasse 3

56179 Vallendar

0261-6402-257

abruehl@pthv.de

Dieser Bericht ist als Volltext über KiDokS (Kirchlicher DokumentenServer der AKThB und des VkwB) verfügbar: http://opus.bsz-bw.de/kidoks/suche_uebersicht.php?la=de

*„**Lücken.** - Die Aufforderung, man solle sich der intellektuellen Redlichkeit befleißigen, läuft meist auf die Sabotage der Gedanken heraus. Ihr Sinn ist, den Schriftsteller dazu anzuhalten, alle Schritte explizit darzustellen, die ihn zu seiner Aussage geführt haben, und so jeden Leser zu befähigen, den Prozess nachzuvollziehen und womöglich – im akademischen Betrieb – zu duplizieren. Das arbeitet nicht bloß mit der liberalen Fiktion der beliebigen, allgemeinen Kommunizierbarkeit eines jeden Gedankens und hemmt dessen sachlich angemessenen Ausdruck, sondern ist falsch auch als Prinzip der Darstellung selber. Denn der Wert eines Gedanken misst sich an seiner Distanz von der Kontinuität des Bekannten. Er nimmt objektiv mit der Herabsetzung dieser Distanz ab; je mehr er sich dem vorgegebenen Standard annähert, umso mehr schwindet seine antithetische Funktion, und nur in ihr, im offenbaren Verhältnis zu seinem Gegensatz, nicht in seinem isolierten Dasein liegt sein Anspruch begründet. Texte, die ängstlich jeden Schritt bruchlos nachzuzeichnen unternehmen, verfallen denn auch unweigerlich dem Banalen und einer Langeweile, die sich nicht nur auf die Spannung bei der Lektüre, sondern auch auf die eigene Substanz bezieht. […]*

Aber weit darüber hinaus ist die Forderung nach intellektueller Redlichkeit selber unredlich. Gäbe man ihr selbst einmal die fragwürdige Anweisung zu, die Darstellung solle den Denkprozess abbilden, so wäre dieser Prozess so wenig einer des diskursiven Fortschreitens von Stufe zu Stufe, wie umgekehrt dem Erkennenden seine Einsichten vom Himmel fallen. Erkannt wird vielmehr in einem Geflecht von Vorurteilen, Anschauungen, Innervationen, Selbstkorrekturen, Vorausnahmen und Übertreibungen, kurz in der dichten, fundierten, aber keineswegs an allen Stellen transparenten Erfahrung. Von ihr gibt die cartesianische Regel, man solle sich nur den Gegenständen zuwenden, „zu deren klarer und unzweifelhafter Erkenntnis unser Geist auszureichen scheine“, samt aller Ordnung und Disposition, worauf sie sich bezieht, einen so falschen Begriff wie die ihr entgegengesetzte und im innersten verwandte Lehre von der Wesensschau. Verleugnet diese das logische Recht, das trotz allem in jedem Gedanken sich geltend macht, so nimmt jene es in seiner Unmittelbarkeit, bezogen auf jeden einzelnen intellektuellen Akt nicht vermittelt durch den Strom des ganzen Bewusstseinslebens des Erkennenden. Darin aber liegt zugleich das Eingeständnis der tiefsten Unzulänglichkeit. Denn wenn die redlichen Gedanken unweigerlich auf bloße Wiederholung, sei´s des Vorfindlichen, sei´s der kategorialen Formen hinauslaufen, so bleibt der Gedanke, der der Beziehung zu seinem Gegenstand zuliebe auf die volle Durchsichtigkeit seiner logischen Genesis verzichtet, allemal etwas schuldig. Er bricht das Versprechen, das mit der Form des Urteils selber gesetzt ist. Diese Unzulänglichkeit gleicht der Linie des Lebens, die verbogen, abgelenkt, enttäuschend gegenüber ihren Prämissen verläuft und doch einzig in diesem Verlauf, indem sie stets weniger ist, als sie sein sollte, unter den gegebenen Bedingungen der Existenz eine

unreglementierte zu vertreten mag. Erfüllte Leben geraden Wegs seine Bestimmung, so würde es sie verfehlen. Wer alt und im Bewusstsein des gleichsam schuldenlosen Gelingens stürbe, wäre insgeheim der Musterknabe, der mit unsichtbarem Ranzen auf dem Rücken alle Stadien ohne Lücken absolviert. Jedem Gedanken jedoch, der nicht müßig ist, bleibt wie ein Mal die Unmöglichkeit der vollen Legitimation einbeschrieben, so wie wir im Traum davon wissen, dass es Mathematikstunden gibt, die wir um eines seligen Morgens im Bett willen versäumten, und die nie mehr sich einholen lassen. Der Gedanke wartet darauf, dass eines Tages die Erinnerung ans Versäumte ihn aufweckt und ihn in die Lehre verwandelt."[1]

Theodor W. Adorno

[1] Adorno, Theodor W. (1951): Lücken. In: Minima Moralia. Reflexionen aus dem beschädigten Leben. Frankfurt: Suhrkamp, S. 141-144

INHALTSVERZEICHNIS

EINLEITUNG

Das NBA ist der Einstieg in eine empirisch gehaltvolle Messung von Pflegebedürftigkeit. Das erste Mal werden Kriterien unabhängig von Zeitaufwand definiert, die Pflegebedürftigkeit erklären sollen. Diese Erklärungen können valide oder nicht–valide sein. Ob das NBA ein valides Abbild einer Theorie über Pflegebedürftigkeit ist, die erklären soll, was man differenzieren muss, um gut pflegen zu können, ist mit statistischen Methoden prüfbar. Eine solche Prüfung kann scheitern. Wir haben für die beiden NBA-Subskalen „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ anhand von 5131 Datensätzen überprüft, ob eine valide Differenzierung von Pflegebedürftigkeit mit diesem Teil des NBA möglich ist. Die Ergebnisse für beide Subskalen weisen auf einen fortbestehenden Entwicklungsbedarf hin.

Für eine weitere Entwicklung von Instrumenten sind sowohl eine Erweiterung des theoretischen Rahmens der eingesetzten Methoden sowie in der Folge ein passgenauerer Methodeneinsatz notwendig.

Das NBA ist ein Messinstrument, das Pflegebedürftigkeit quantifizieren will. Hierfür bedient es sich eines quantifizierenden Messmodells. Das Messmodell ist das Vehikel, mit dem die inhaltlichen Überlegungen zur Pflegebedürftigkeit in ein Messergebnis, also Zahlen, umgesetzt werden sollen, um stärker ausgeprägte Pflegebedürftigkeit von weniger stark ausgeprägter Pflegebedürftigkeit zu unterscheiden.

Ziel unserer Arbeit ist es, die Beziehung zwischen Gegenstand (=Pflegebedürftigkeit) und dem Ergebnis seiner Messung (=resultierenden Bedarfsgraden) zu prüfen.

Innerhalb des Entwicklungsprozesses des NBA sind solche empirisch gestützten Rückkopplungsschleifen zwischen Ergebnissen und zu überarbeitendem theoretischem Modell nur schwer zu realisieren. Begründet ist dies im Umfang der Aufgabendefinition und im zu engen Zeitrahmen, den die Ausschreibung des Projekts zur Entwicklung des NBA definiert hat. Dies wurde bereits im Vorfeld der Entwicklung des NBA kritisiert. (Becker et al, 2007). Da das Ergebnis des Projekts Jahre ungenutzt blieb, wirkt der enge Zeitrahmen aus heutiger Sicht künstlich.

In pflegerischen Messinstrumenten werden sehr häufig Quantifizierungen komplexer Konstrukte vorgenommen. Auch im NBA werden in einem ersten Schritt Summenwerte der einzelnen Subskalen gebildet, in einem zweiten Schritt werden dann vier Stufen pro Subskala unterschieden, die in einem dritten Schritt wiederum gewichtet und in einem vierten Schritt in einer Summe zusammengefasst und in einem fünften Schritt

einem Bedarfsgrad zugeordnet werden. Das ist ein Messmodell mit fünf Schritten, in denen mehrmals das Skalenniveau gewechselt wird, was die Gefahr von Informationsverlust in sich birgt. Wenn möglichst wenig Informationen auf dem Weg vom einzelnen Kriterium zum Bedarfsgrad verloren gehen soll, dann muss überprüft werden, ob bei den fünf Schritten die strukturerhaltende Abbildung des gemessenen Merkmals (Pflegebedürftigkeit) in den am Ende produzierten Zahlen (Bedarfsgrade) gewährleistet wird. Die Ergebnisse dieses Berichts verweisen darauf, dass ein valides Klassifikationsinstrument einfacher sein müsste als es das NBA aktuell ist.

Unabhängig vom Messmodell wird die Tatsache, dass die neuen Bedarfsgrade zu den alten Pflegestufen in Bezug gesetzt werden (Windeler 2008) dem inhaltlich innovativen Ansatz eines breiten Verständnisses von Pflegebedürftigkeit nicht gerecht. Das Ziel, eine empirisch valide Messung von Pflegebedürftigkeit zu entwickeln, ist methodisch auf diesem Weg nicht zu erreichen.

Das NBA ist das Ergebnis intensiver Arbeit, die wir ausdrücklich würdigen. Und dort, wo es Inhalte von Pflegebedürftigkeit mit bislang unberücksichtigten Kriterien differenziert, ist es ein Fortschritt im Vergleich zum bislang etablierten Verfahren. Probleme liegen in erster Linie in der Art begründet, in der aus einzelnen Kriterien in einem fünfstufigen Summierungs- und Gewichtungsprozess die Bedarfsgrade entstehen. Das NBA mischt in seiner aktuellen Form Erklärungs- und reine Abbildungsversuche von Pflegebedürftigkeit. Im Interesse einer besseren Erklärung von Pflegebedürftigkeit müsste die theoretische Entwicklung der Kriterien nochmals geprüft werden. Die im NBA-Ansatz enthaltene Hypothese, man müsse verschiedene Subskalen isoliert für sich summieren und deren Summen dann wiederum gewichtet erneut summieren, muss in der vorliegenden Struktur verworfen werden.

Uns ist bewusst, dass wir alleine mit der Prüfung des Messmodells des NBA keine Alternative zum NBA anbieten können. Wir können aber die Hauptansatzpunkte für Verbesserungen des NBA benennen:

- Die nicht valide Quantifizierung über Summenwerte der Items der beiden von uns geprüften Subskalen,
- die Quantifizierung über unterschiedliche Antwortskalen (Selbständigkeit, Beeinträchtigung, Häufigkeit) und
- die Setting-Abhängigkeit der Items[2].

Das NBA arbeitet in seiner Subskala „Selbstversorgung“ implizit wieder mit dem Faktor Zeit als Klassifikationsmerkmal. Letztlich wird ein Bezug zum Zeitaufwand zur

[2] Die Antwortskalen „funktionieren“ im stationären und ambulanten Pflegesetting unterschiedlich, obwohl sie Pflegebedürftigkeit unabhängig vom Versorgungssetting messen sollen (vgl. Bensch Kapitel 5 in diesem Band)

Kompensation von Unselbständigkeit hergestellt, um Pflegebedürftigkeit zu unterscheiden. Wenn bei der Unterscheidung von Klassifikationskriterien mit Zeitaufwand argumentiert wird, erklären die resultierenden Bedarfsgrade Pflegebedürftigkeit nicht, sondern bilden sie nur ab. Ein wesentliches Fazit des Berichts ist, dass wir uns bei einer Weiterentwicklung an den erklärenden Teilen der Skalen orientieren müssen.

Die den empirischen Analysen dieses Berichts zugrunde liegenden Daten wurden mit qualifizierten Pflegefachkräften und einem überarbeiteten Handbuch zum NBA erhoben. Die empirischen Analysen beziehen sich allesamt auf die beiden Subskalen „Mobilität“ sowie „Kognitive und kommunikative Fähigkeiten“. Beide Skalen sind zentral für eine Erklärung von Pflegebedürftigkeit und können als reflektive Messmodelle in ihrer Güte getestet werden. Bislang wurden unsere Analysen noch nicht mit denen durch das Institut für Pflegewissenschaft erhobenen Daten (Wingenfeld et al 2008) wiederholt und verglichen. Um ggf. Abweichungen der Ergebnisse aufgrund der unterschiedlichen Vorgehensweise bei der Datenerhebung auszuschließen, würden wir dies gerne nachholen, sofern uns die Daten zur Verfügung gestellt werden.

ÜBERBLICK

Im ersten Kapitel des vorliegenden Berichts entwickeln wir ein heuristisches Rahmenmodell für die Instrumentenentwicklung generell und prüfen die Nützlichkeit probabilistischer Testtheorie im Gegensatz zur Klassischen Testtheorie in der Pflegewissenschaft. Hier wird begründet, warum es zur Entwicklung pflegerischer Instrumente einer Erweiterung des methodologischen Rahmens bedarf. Im zweiten Kapitel stellen wir das bislang implizite Messmodell des NBA vor, explizieren und diskutieren es. Hier wird dargestellt, dass grundlegende messtheoretische Voraussetzungen im NBA nicht erfüllt werden. Im dritten Kapitel wird der Unterschied zwischen formativen und reflektiven Messmodellen erklärt. Dass dieser Unterschied nicht berücksichtigt wird, erschwert die Weiterentwicklung des Instruments. Im vierten Kapitel werden Methoden aus dem Bereich der Klassischen Testtheorie (Konfirmatorische Faktorenanalyse) zur Analyse einer Subskala des NBA eingesetzt und die dabei auftretenden Schätzprobleme thematisiert. Schätzprobleme entstehen, wenn die vorliegenden Daten nicht intervallskaliert und nicht normalverteilt sind. Ergebnisse von klassisch-testtheoretisch fundierten Verfahren sind daher einer kritischen Bewertung zu unterziehen. Im fünften Kapitel werden vier probabilistisch fundierte Methoden erprobt: dichotome und ordinale Rasch-Modelle sowie dichotome und ordinale latente Klassenanalysen. Probabilistische Methoden können die vorgenommenen Quantifizierungen prüfen und setzen sie nicht wie klassisch

testtheoretisch fundierte Methoden als valide voraus. Sie stellen ebenfalls qualitativ standardisierbare Alternativen zur Quantifizierung zur Verfügung.

Im sechsten und letzten Kapitel verweisen wir auf unsere Arbeiten im Bereich der stationären Langzeitpflege, in denen wir uns im Sinne eines Anschlusses an die Praxis und einer realistischen Unterscheidung zunächst um eine empirische Erklärung von Pflegeaufwand bemühen. Hierzu verweisen wir auf nützliche Verfahren, die aus Daten „lernen“ und Klassifikationsmodelle zu entwickeln vermögen: Nonparametrische Regression und Methoden des maschinellen Lernens. Eine empirisch valide Klassifikation von Pflegeaufwand wird auf diesem Weg im November 2012 mit dem Projektende von PiSaar[3] auf validem Niveau möglich sein. Allerdings reicht dieser empirische Zugang alleine nicht aus, weil in ihm nur erklärt wird, was mit der aktuellen personellen Ausstattung möglich ist, nicht aber was ein Pflegebedürftiger benötigt. In Methoden, die aus Daten lernen, spiegeln sich teilweise schon die Bedingungen, die in der Pflegepraxis herrschen, wider. Wir werden deshalb empirische Ergebnisse zur Klassifikation von Pflege mit normativen Konzepten zu „guter Pflege“ verbinden müssen. Hierzu sind im NBA wichtige Ansätze enthalten. Wenn man diese variierte, erweiterte und mit einem empirisch validen Messmodell verbände, wäre ein wichtiger Schritt auf dem Weg zu einem validen Klassifikationssystem von Pflegebedürftigkeit getan.

Unter Einsatz der Facettentheorie versuchen wir zudem, die Definition und Operationalisierung relevanter Konstrukte der Pflegebedürftigkeit weiter zu präzisieren. Auch hierauf gehen wir im letzten Kapitel ein.

Mit dem NBA sind die Bemühungen, ein theoretisch besser fundiertes Verständnis von Pflegebedürftigkeit über ein angemessenes Messinstrument in der Praxis zur Wirkung zu verhelfen nicht beendet. Es gilt, vieles weiter zu entwickeln: Die Theorie über Pflegebedürftigkeit, die Operationalisierung der Theorie in ein Messinstrument, das Struktur- und Messmodell und deren Bezug zur Pflegepraxis. Diese Art von Arbeiten mit einer verbindlichen Einführung des NBAs zu beenden, würde bedeuten, alle offenen Fragen für irrelevant zu erklären. Wir plädieren mit diesem Bericht nachdrücklich für eine Überarbeitung des NBA.

Ohne die dem Bericht zugrunde liegenden Qualifikationsarbeiten wäre die vorliegende Analyse nicht möglich gewesen. Und so danke ich vor allem Frau Bensch, die mit ihrer Dissertation das Fundament unserer Arbeiten gelegt und die im Kern fast alle Methoden, die in diesem Band diskutiert werden für die Pflege erschlossen hat.

[3] PiSaar (Pflegebedarf im Saarland): ein Projekt der Saarländischen Pflegegesellschaft (SPG) zum Personalbedarf und –einsatz.

Insbesondere gilt dies für die ordinalen latenten Klassenanalysen. Herrn Franken danke ich für die Detailgenauigkeit bei der Erarbeitung und Darstellung der Faktorenanalysen, die die begründete und differenzierte Kritik an ausschließlich klassisch-testtheoretischem Vorgehen ermöglicht sowie für seinen wichtigen Beitrag zur Unterscheidung von formativen und reflektiven Messmodellen. Frau Schröder danke ich für ihre wichtigen Erläuterungen zu den polychorischen Korrelationen, die wiederum eine zentrale Rolle bei der Behandlung ordinaler Daten in den konfirmatorischen Faktorenanalysen spielen, die Herr Franken nutzen konnte. Herrn Grebe danke ich für seine Einstiegs-Arbeit im Bereich des so genannten „Statistical Learning“, also den Ansätzen, die aus Daten lernen und uns helfen, Strukturen zu entdecken und Frau Planer für ihre fruchtbaren Anregungen für die notwendige Neukonstruktion von Subskalen mit Hilfe der Facettentheorie, ihren Esprit, ihre Begeisterung und ihre Ausdauer in der andauernden vor allem inhaltlichen aber auch der redaktionellen Bearbeitung des vorliegenden Textes.

Univ.-Prof. Dr. Albert Brühl, Vallendar im Juli 2012

1. Methodologische Orientierung der Pflegewissenschaft bei der Entwicklung standardisierter Messinstrumente

Albert Brühl

Einleitung

Die Entwicklung von standardisierten Messverfahren ist ein wichtiges Arbeitsfeld der Pflegewissenschaft. Politisch definierte Aufgaben wie z. B. die Entwicklung eines neuen Pflegebedürftigkeits-Einschätzungsinstruments oder der Transparenzkriterien zur Messung der Qualität nach § 115 SGB XI müssen im Rahmen von Auftragsforschungs-Projekten in relativ kurzer Zeit bewältigt werden.

Von den resultierenden Instrumenten (z. B. Neues Begutachtungsassessment zur Feststellung der Pflegebedürftigkeit [NBA]) wird erwartet, dass sie die gemessenen Konstrukte (z. B. Pflegebedürftigkeit) quantifizieren, also in ihrem Ausmaß bestimmen.

Die Dominanz politischer Handlungsorientierung in Projekten mit stark limitiertem Zeitbudget und der Widerstreit einer Vielzahl nicht gegenstandsbezogener Interessen erschwert aber die Entwicklung und Anwendung einer wissenschaftlich fundierten Methodologie und passender Methoden in der noch jungen Disziplin Pflegewissenschaft. Im Ergebnis überwiegen dann politisch definierte und fachlich plausible Anforderungen an die Instrumentenentwicklung[4]. Aufgrund eines meist sehr engen Zeitrahmens und politisch dominierten Handlungsdrucks findet eine wissenschaftliche Methodendiskussion kaum statt. Wir möchten trotzdem neue Methoden in die Instrumentenentwicklung der Pflegewissenschaft einführen, um zukünftig aus empirischen Ergebnissen Rückschlüsse auf Theorie- und Instrumentenentwicklung besser zu ermöglichen.

Wissenschaftlich diskutiert werden bislang in erster Linie die Inhalte solcher Instrumente, jedoch nicht die eingesetzten Methoden der Instrumentenentwicklung. Entwickelt werden meist ausschließlich reine Abbildungsmodelle[5] und nicht Theorien,

[4] Vgl. auch die Kriterien der Pflegetransparenzvereinbarungen nach § 115 SGB XI oder die Qualitätsindikatoren zur Beurteilung der Ergebnisqualität in der stationären Altenhilfe (Wingenfeld et al., 2011)

[5] Z. B. Monika Krohwinkels Strukturierungsmodell der ABEDL (Krohwinkel 2007)

also explizierte Erklärungsansätze, aus denen präzise Operationalisierungen für die Quantifizierung der Konstrukte abgeleitet werden könnten.

Bei der Instrumentenentwicklung bevorzugt die Pflege ein politisches, inhaltliches und induktives Vorgehen, ein empirisch-testendes bleibt bislang weitgehend außer Acht.

Die Beziehung zwischen Messergebnis und Merkmalsausprägung wird so nicht mehr wirklich geprüft, wenn ein Instrument erst einmal erstellt wurde. Das haben wir für Teile des NBA nachgeholt und wollen die Gelegenheit nutzen, neue Methoden in die Entwicklung von Instrumenten in die pflegewissenschaftliche Diskussion zu integrieren.

Für eine Erweiterung der eingesetzten Methoden in der Instrumentenentwicklung müssen folgende Fragen bearbeitet werden.

1. Wie könnte ein heuristisches Modell für die Instrumentenentwicklung der Pflege beschrieben werden?
2. Welcher testtheoretische Rahmen bietet welche Chancen und welche Erkenntnisgewinne für die Instrumentenentwicklung der Pflegewissenschaft?
3. Konkret: was leistet beispielsweise das Rasch-Modell als probabilistisches Verfahren im Vergleich zur Anwendung der internen Konsistenzanalyse mit Cronbachs alpha oder der Faktorenanalyse als Vertreter der klassischen testtheoretisch fundierten Verfahren?
4. Welche Konsequenzen ergeben sich aus einer Erweiterung des Methodenspektrums für die Entwicklung von Erhebungsinstrumenten in der Pflege?

1. Ein heuristisches Modell für die Instrumentenentwicklung der Pflege

Ein Mess- oder Assessmentinstrument stellt im Idealfall eine in ein Erhebungsinstrument operationalisierte Anwendung einer Theorie dar. In der Pflege werden wie in anderen Disziplinen auch, sogenannte „latente Konstrukte“ gemessen. „Latent“ bedeutet, dass das Konstrukt als solches nicht direkt beobachtbar ist, sondern die Gesamtbetrachtung der Beobachtungen mehrerer manifester (beobachtbarer) Variablen, deren wechselseitige Abhängigkeiten oft nicht bekannt sind, eine Aussage über die Ausprägung des latenten Konstrukts machen soll. Pflegebedürftigkeit, Pflegequalität, Sturzrisiko, Lebenszufriedenheit, usw. sind latente Konstrukte, die sich

durch eine relativ hohe Komplexität ihrer inhaltlichen Dimensionen[6] auszeichnen.

Der Entwicklungsprozess eines Messinstruments lässt sich durch vier Aspekte, die in Abb. 1.1 dargestellt werden, strukturieren. Alle vier Aspekte stehen in wechselseitigen Beziehungen zueinander.

Um ein validierbares, standardisiertes Instrument zur Messung eines komplexen Konstrukts entwickeln zu können, muss die Aufgabe (Klassifikation oder Quantifizierung) klar sein, die Inhalte sollten theoretisch fundiert und das zugehörige Struktur- und Messmodell sollte expliziert und während der Instrumenten-konstruktion immer wieder auf seine Passung hin überprüft werden.

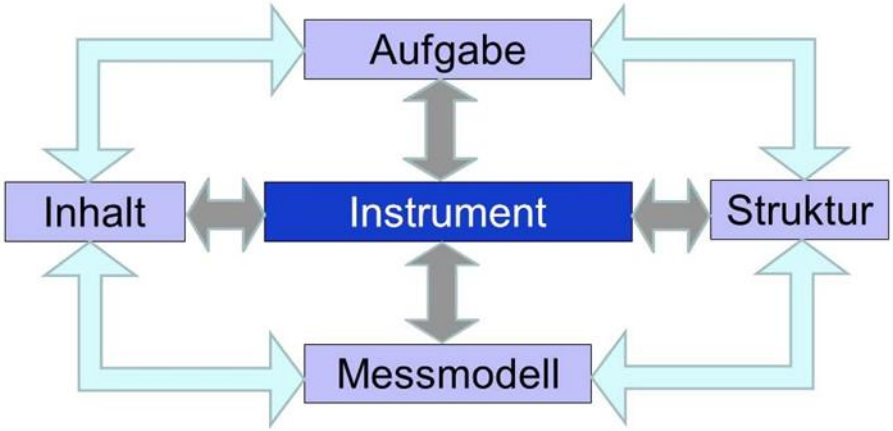


Abb. 1.1 Heuristischer Rahmen der Instrumentenentwicklung

Aufgabe

Was ist die Aufgabe, die das Instrument erfüllen soll? Ist diese Aufgabe so definiert, dass sie erfüllt werden kann und erfüllt das Instrument diese Aufgabe?

Von der im Rahmen der Aufgabe formulierten Zielsetzung ist das Messmodell eines Instruments abhängig: Ein Messmodell kann dazu dienen zu klassifizieren, also qualitative Aussagen zu machen oder es kann dazu dienen zu quantifizieren, also einen „größer als“ - bzw. „kleiner als“ - Zusammenhang herzustellen. Pflegebedürftigkeit könnte –hypothetisch angenommen - z. B. in vier Gruppen klassifiziert werden. Ein Klassifikationsmodell unterscheidet verschiedene Merkmalsprofile der „Mitglieder“ der einzelnen Klassen und produziert damit in jedem Fall eine qualitative Unterscheidung. Wird einem Instrument die Aufgabe gestellt,

[6] Dimensionen strukturieren Konstrukte auf unterschiedlichen Abstraktionsebenen. Ein Konstrukt (z. B. Pflegebedürftigkeit) kann als inhaltlich mehrdimensionales Konstrukt definiert werden, das sich aus mehreren Teil-Konstrukten zusammensetzt, z. B. Mobiltät und Kognition, die wiederum ebenfalls mehrere Dimensionen umfassen können.

Pflegebedürftigkeit zu quantifizieren, könnte ein kompensatorisches einfaches Summenmodell eine sinnvolle Zielsetzung im Rahmen der Aufgabe darstellen. Ein Instrument, das diese Aufgabe erfüllen kann, besteht aus Kriterien, die sich gegenseitig ersetzen (kompensieren) und sinnvoll summiert werden können. Menschen mit einem unterschiedlichen Summenwert würden sich in der „Schwere“ ihrer Pflegebedürftigkeit, bzw. im Umfang des Pflegeaufwands unterscheiden, Menschen mit gleichem Summenwert würden demnach bei einem kompensatorischen Summenmodell auch eine gleiche Pflegebedürftigkeit aufweisen. Im NBA wird mit einem Summenmodell ein quantitativer Index gebildet, der aufgrund einer normativen Setzung von Schwellenwerten in Pflegebedarfsgrade (Pflegestufen) mündet. Es wird davon ausgegangen, dass die Pflegebedürftigkeit von einem Bedarfsgrad zum nächsten ansteigt, Pflegebedürftigkeit wird also quantifiziert.

Inhalt

Bei dem Aspekt des Inhalts wird nach einem Erklärungsansatz gefragt, der begründen könnte, welche Kriterien im Instrument berücksichtigt werden müssen, damit die Aufgabe erfüllt werden kann. Für das Konstrukt der Pflegebedürftigkeit können verschiedene Aspekte als inhaltliche Dimensionen (z. B. Kognition und Mobilität) diskutiert werden, die berücksichtigt werden müssen, um verschiedene Grade von Pflegebedürftigkeit unterscheiden zu können. Aufgabe und Inhalt eines Instruments (siehe Abb. 1.1) sind in einer Instrumentenentwicklung die Bereiche, die methodisch in erster Linie eine klare theoretische Definition des Konstrukts erfordern. Aus einer Theorie zur Pflegebedürftigkeit müssen die konkreten Items abgeleitet und operationalisiert werden, mit denen diese Aufgabe bewältigt werden kann.

Messmodell

Der Aspekt des Messmodells hat die Frage zu beantworten, welches Messmodell angestrebt wird und welches Messmodell angepasst werden kann?

Lautet die Aufgabenstellung, Pflegebedürftigkeit zu quantifizieren, so muss das Instrument in der Lage sein, aus ursprünglich kategorialen Daten mit Hilfe von Quantifizierungsschritten [7] ein quantitatives Gesamtergebnis (Index [8]) über die

[7] Modelle der Item-Response-Theorie (IRT) identifizieren Items, die geeignet sind, die Selbständigkeit, bzw. im Umkehrschluss Pflegebedürftigkeit auf der Grundlage der Fähigkeiten der Person im Bezug zur Schwierigkeit der Anforderungen auf der Basis des Antwortverhaltens zu messen. Wesentliche Weiterentwicklungen hat die IRT im Rahmen der Kompetenzmessung der PISA-Tests (Programme for International Student Assessment) erfahren. Im Rahmen der PISA-Tests entspricht die hier mit „Selbständigkeit“ benannte Dimension der Dimension „Kompetenz“, die davon abhängig ist, welche Aufgaben unterschiedlicher Schwierigkeit eine Person aufgrund ihrer Fähigkeiten erfolgreich bewältigen

Ausprägung der Pflegebedürftigkeit zu produzieren. Die Bedingungen dafür, dass ein Instrument korrekt quantifizieren kann (valider Summenwert) sind dabei sehr streng: Um auf der abstrakten Dimension „Selbständigkeit“ als Konstrukt eindimensional messbar zu sein, müssen die Merkmale oder Kriterien (Items) inhaltlich geeignet und relevant sein, und sie müssen die gleichen Trennschärfen[9] auf unterschiedlichen Schwierigkeitsniveaus aufweisen.

Wird eine Auswahl von fünf Items mit einer ordinalen Likert-Antwortskala versehen, deren einzelnen Antworten Punktwerte zugeordnet werden, die summiert werden, so werden mit einem solchen kompensatorischen Summenmodell (ohne Auswahlmöglichkeit unter den wenigen Items) hohe Hürden für die Validierung einer solchen Skala aufgestellt. Einfacher wäre es, aus einem größeren Itempool die Items selektieren zu können, die dann diese Voraussetzungen erfüllen.

Struktur

Um das latente Konstrukt in Form eines Instruments messbar machen zu können, müssen seine inhaltlichen Aspekte in eine Struktur gebracht werden, die die Beziehungen und Verhältnisse seiner inhaltlichen Elemente zueinander definieren. Latente und beobachtbare Variablen stellen die Operationalisierung der Inhalte in Bezug auf die definierte Struktur des Konstrukts dar. Es gibt die Möglichkeit, Annahmen über das Verhältnis von latenten und beobachtbaren Variablen in einem „nomologischen Netz“[10] abzubilden, das einer Theorie der Messung des Konstrukts entspricht. Ziel dabei ist es, die Relevanz und Struktur der Prädiktoren des nomologischen Netzes schrittweise prüfen zu können (vgl. Franken 2012, in diesem Band, S. 77). Aufgrund theoretischer Überlegungen kann die Hypothese formuliert werden, dass es sich bei Pflegebedürftigkeit um ein mehrdimensionales Konstrukt handelt. Fragen nach der Anzahl der Dimensionen und ihrer Beziehung zueinander schließen sich an.

---

kann. Maße der Aufgabenschwierigkeit und der Personenfähigkeit werden mit Hilfe der empirischen Daten ermittelt.

[8] Ein Index ist eine Kennzahl, die sich für Vergleichszwecke aus der Verrechnung verschiedener Einzelwerte ergibt.

[9] Die Trennschärfe eines Items drückt aus, wie geeignet ein Item, bzw. seine Antwortkategorien sind, zwischen verschiedenen Eigenschaftsausprägungen der Testperson zu unterscheiden.

[10] Ein nomologisches Netz stellt die Beziehungen, Wechselwirkungen und Abhängigkeiten der beobachtbaren Variablen, die auch Prädiktoren genannt werden, in Beziehung zum latenten Konstrukt dar. Die Beziehungen der beobachtbaren Variablen zum latenten Konstrukt lassen sich durch die Prüfung von Korrespondenzhypothesen konkretisieren und dienen damit sowohl der Entwicklung valider Tests oder Instrumente als auch der Präzisierung der Theorie.

Für Mobilität und Kognition stellt sich bei der Messung von Pflegebedürftigkeit die Frage, ob Einschränkungen in der Kognition durch bessere Mobilität ausgeglichen werden können und es sich dadurch bei der Struktur des Konstrukts also tatsächlich um ein kompensatorisches Modell handelt oder nicht.

Ob das Instrument seine Aufgabe erfüllt, weil sich das postulierte Mess- und Strukturmodell anhand empirischer Daten bestätigen lässt, kann nur dann geprüft werden, wenn bei der Entwicklung des Instrument die Komplementarität der genannten Aspekte berücksichtigt wurde. Über die Passung von Struktur- und Messmodell auf die Daten und über die Erfüllung der operational definierten Aufgabe kann mit Hilfe statistischer Verfahren entschieden werden.

Denn schon bei der Operationalisierung der Inhalte in verschiedene Variablen (Items) gilt es zu überlegen, welche inhaltlichen Hypothesen sich mit welchen statistischen Verfahren prüfen lassen. Ferner gilt es zu bedenken, welcher Methodeneinsatz für die Validierung des Instruments aus den empirischen Ergebnissen für die Anpassung von Mess- und Strukturmodell Erklärungsansätze für die theoretische Weiterentwicklung ermöglichen.

Bei der Entwicklung eines Instruments ist davon auszugehen, dass die empirischen Daten, die sich mithilfe eines ersten Entwurfs eines Instruments erheben lassen, einen Erkenntnisgewinn über die Theorie des latenten Konstrukts, die verwendeten Items und die genutzte Skala liefern.

Der Prozess einer empirischen Prüfung von Theorien ist ggf. mehrmals zu wiederholen, um aus den empirischen Daten lernen zu können. Ziel ist es, das Instrument zu verbessern und im Ergebnis die Validität des Instruments dadurch zu steigern.

2. Was unterscheidet die klassische Testtheorie (KTT) von der probabilistischen Testtheorie (PTT) und welche von beiden ist in der Instrumentenentwicklung wofür geeignet?

KTT und PTT beziehen sich auf unterschiedliche Datenarten.

Die Methoden, die im Rahmen der klassischen Testtheorie eingesetzt werden, setzen mindestens intervallskalierte Daten voraus, die PTT kann mit kategorialen und mit metrischen Daten arbeiten.

Die Fixierung auf metrische Daten ist durch das erste Axiom der klassischen Testtheorie bedingt, das besagt, dass ein beobachteter Wert X aus einem wahren Wert, auch „True score“ = T und einem Messfehler(= Error) E zusammengesetzt ist. Die bekannte Gleichung lautet demnach X = T + E.

Eine Unterscheidung von wahrem Wert und einem Fehleranteil in einem beobachteten Wert ist nur dann sinnvoll, wenn die Werte metrischer Natur sind. Hierauf haben bereits Steyer und Eid (in Rost 1999, 141) hingewiesen.

Für Daten, die eine eindeutige kategoriale Zuordnung, jedoch keine metrischen Informationen enthalten, ist diese Annahme sinnfrei. Die Frage, ob eine Person selbständig einen Positionswechsel im Bett vornehmen kann, ist sicher mit „ja“ oder „nein“ beantwortbar. Allerdings kann in diesem dichotomen Antwortformat nicht sinnvoll über einen Fehleranteil und einen wahren Wert (als Grundannahme der Klassischen Testtheorie) spekuliert werden.

Wird die gleiche Frage, ob eine Person selbständig einen Positionswechsel im Bett vornehmen kann, mit einer Likert-skalierten Antwort von z. B. „selbständig“ (= 0), „überwiegend selbständig“ (= 1), „überwiegend unselbständig“ (= 2) und „unselbständig“ (= 3) verbunden, kann hieraus nicht einfach abgeleitet werden, dass es sich um eine sinnvolle Quantifizierung eines qualitativen Merkmals handele. Eine valide Quantifizierung setzt voraus, dass die Kategorien sich tatsächlich in einem Größer-kleiner-Verhältnis abstufen lassen. Die zugewiesenen numerischen Werte machen deutlich, dass im Messmodell davon ausgegangen wird, dass eine Person, die als „überwiegend unselbständig“ mit dem Wert 2 eingeschätzt wird als doppelt so unselbständig angesehen wird, wie eine Person, die als „überwiegend selbständig“ den Wert 1 zugewiesen bekommt. Jedoch würde nur eine solche valide Quantifizierung des gemessenen Merkmals eine Unterscheidung konzeptionell „wahrer“ Varianzanteile von „Fehlervarianzen“ ermöglichen.

Eine lineare Beziehung zwischen latenter Variable und der Ausprägung der beobachtbaren Antwort auf obengenanntes Item ist Voraussetzung für den Einsatz klassisch testtheoretischer Methoden, kann aber mit ihrer Hilfe nicht geprüft oder hergestellt werden. Es müssen probabilistisch testtheoretisch fundierte Methoden eingesetzt werden, um aus qualitativen Daten valide quantitative Daten zu machen.

Daher muss mit probabilistischen Verfahren geprüft werden, ob eine solche Antwortskala tatsächlich valide metrische Daten liefert. Eine solche Prüfung ist mit Methoden, die bereits metrische Daten voraussetzen (KTT), nicht möglich.

Da die Grundannahmen der KTT ebenso wie viele Abbildungsmodelle der Pflege die Eigenschaft miteinander teilen, nicht prüfbar zu sein und auch keine prüfbaren Modell-Annahmen für kategoriale Daten aus ihnen ableitbar sind (Abb. 1.2), passen die Methoden in die induktiven, nicht empirisch-deduktiven Entwicklungsstrategien der Instrumentenentwicklung innerhalb der Pflegewissenschaft.

Deshalb ist es bislang in der Pflegewissenschaft auch nicht aufgefallen, dass die empirisch mögliche Prüfung von Struktur- und Messmodell innerhalb des Rahmens der KTT auch mit Reliabilitäts- oder Faktorenanalysen dann nicht aussagekräftig stattfindet, wenn keine metrischen Daten vorliegen oder durch Skalen eben nur scheinbar produziert werden.

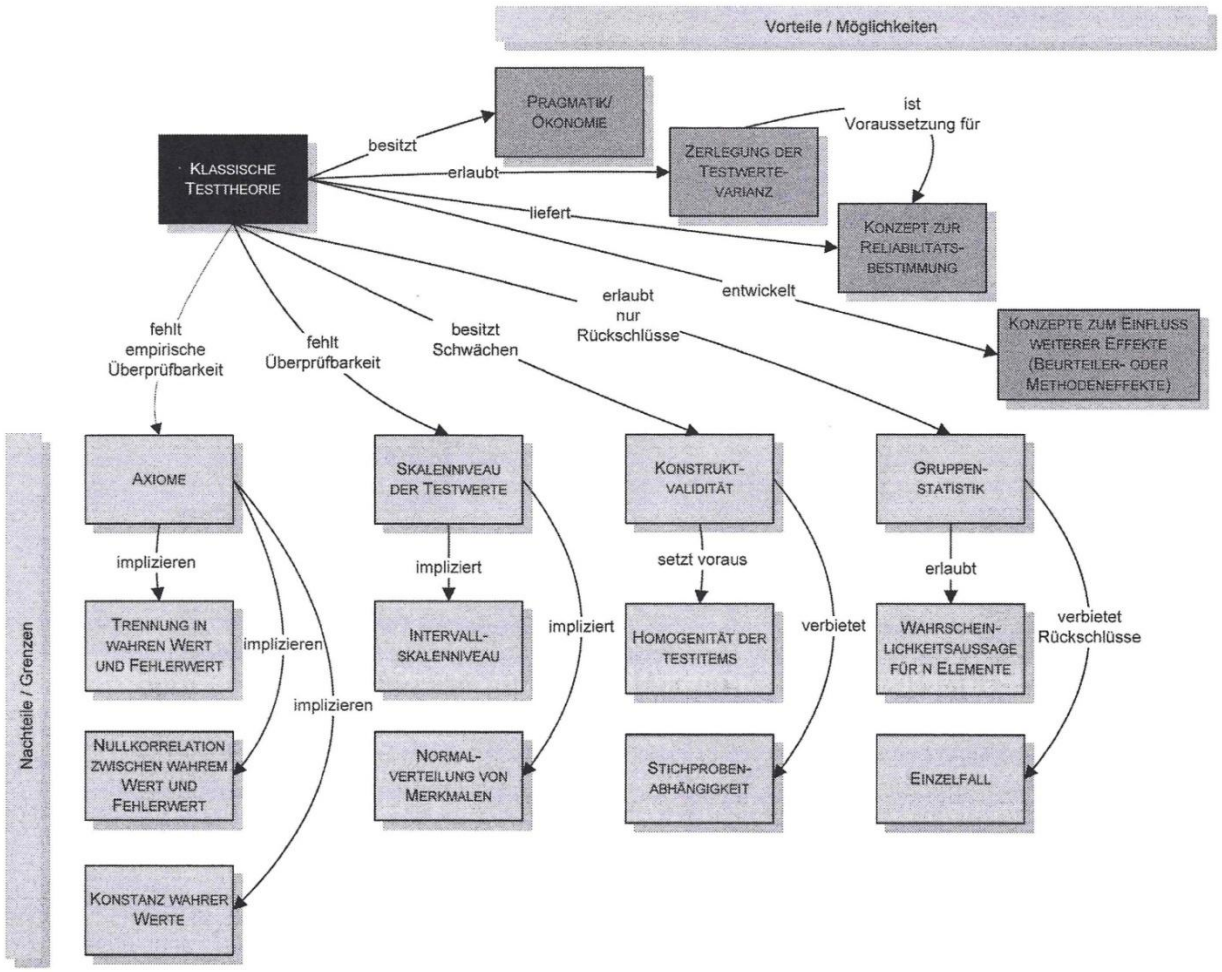


Abb. 1.2 Kritik an der Klassischen Testtheorie im Überblick (Pospeschill, 2010, S. 112)

Klassisch testtheoretisches Vorgehen setzt zusätzlich die Konstanz des gemessenen Merkmals über die Zeit voraus. Das ist für Konstrukte wie Pflegebedürftigkeit aber keine sinnvolle Annahme. Jede Varianz bei z. B. einer Messwiederholung wäre in der KTT eine Fehlervarianz und senkte die Reliabilität[11]. Eine Unterscheidung von Fehler- und wahrer Varianz bei Messwiederholungen zur Veränderungsmessung bei variablen Merkmalen wie z. B. Pflegebedürftigkeit ist innerhalb der KTT nicht sinnvoll möglich.

PTT und KTT unterscheiden sich auch in der Art, in der sie Strukturannahmen testen. In der klassischen Testtheorie werden Strukturhypothesen in Faktorenanalysen mittels bivariater Kovarianzmatrizen getestet. Kern dieser Strukturprüfungen sind Variablenpaare in eben diesen Kovarianz- und Korrelationsmatrizen.

Innerhalb der probabilistischen Testtheorie wird z. B. im Rasch-Modell von vornherein von Eindimensionalität auf einer abstrakteren Dimension, wie z. B. „Schwierigkeit“, bzw. „Selbständigkeit“ ausgegangen. Will man innerhalb der PTT mehrere Subdimensionen voneinander unterscheiden, so müssen diese Dimensionen vorab mit eigens dazu einzusetzenden anderen Verfahren (z. B. der Multidimensionalen Skalierung) identifiziert werden. Das Verhältnis der Variablen spielt dabei im Rasch-Modell nicht nur in der Form von Variablenpaaren eine Rolle, sondern hier werden Beziehungen höherer Ordnung, also in einer Beziehung aller Variablen zueinander gleichzeitig erfasst. Die Reaktion auf ein bestimmtes Kriterium innerhalb eines Instrumentes z. B. in einem einparametrischen (1-PL-Modell)[12] Rasch-Modell ist allein abhängig von der Differenz zwischen Itemschwierigkeit und Personenfähigkeit und wird in der probabilistischen Testtheorie als Wahrscheinlichkeit ausgedrückt (Abb. 1.3). Damit kann innerhalb dieses Ansatzes auch die Messung nicht konstanter Merkmale gut modelliert werden, denn hier ist nicht länger die Varianz der Ergebnisse entscheidend, sondern die Wahrscheinlichkeit der empirischen Daten aufgrund des theoretischen Modells. Ändert sich das Merkmal, so ändert sich auch die Wahrscheinlichkeit von Reaktionen auf ein Kriterium. Insofern ist Merkmalskonstanz keine notwendige Voraussetzung für den Einsatz der PTT.

[11] Als Reliabilität wird die Messgenauigkeit eines Instruments oder eines Items bezeichnet. Die Reliabilität erhöht sich, wenn der Fehleranteil des gemessenen Wertes niedrig ist, d.h. der Messwert den wahren Wert möglichst gut abbildet und kaum einen Fehleranteil enthält. Perfekte Reliabilität würde ein Instrument/Item dann aufweisen, wenn der Messwert keinen Fehleranteil enthält, der Messwert also den wahren Wert darstellt. Aufgrund fehlender Merkmalskonstanz erscheinen in der Pflege gewöhnlicherweise auftretende Merkmalsschwankungen als Fehleranteile in Messwerten obwohl es sich tatsächlich nicht um Fehler handelt, sondern um wahre Schwankungen des Merkmals.

[12] Ein-Parameter-Modell (1-PL-Modell) wird auch als Bezeichnung für das dichotome Rasch-Modell verwendet, das auf einem dichotomen Antwortformat aufbaut. Beim 2-Parameter- oder auch Birnbaum-Modell (ordinales Antwortformat) wird neben den Item- und Personenparametern des einparametrischen Modells auch noch ein Trennschärfeparameter geschätzt. Das 3-Parameter-Modell berücksichtigt darüber hinaus einen Parameter für die Ratewahrscheinlichkeit.

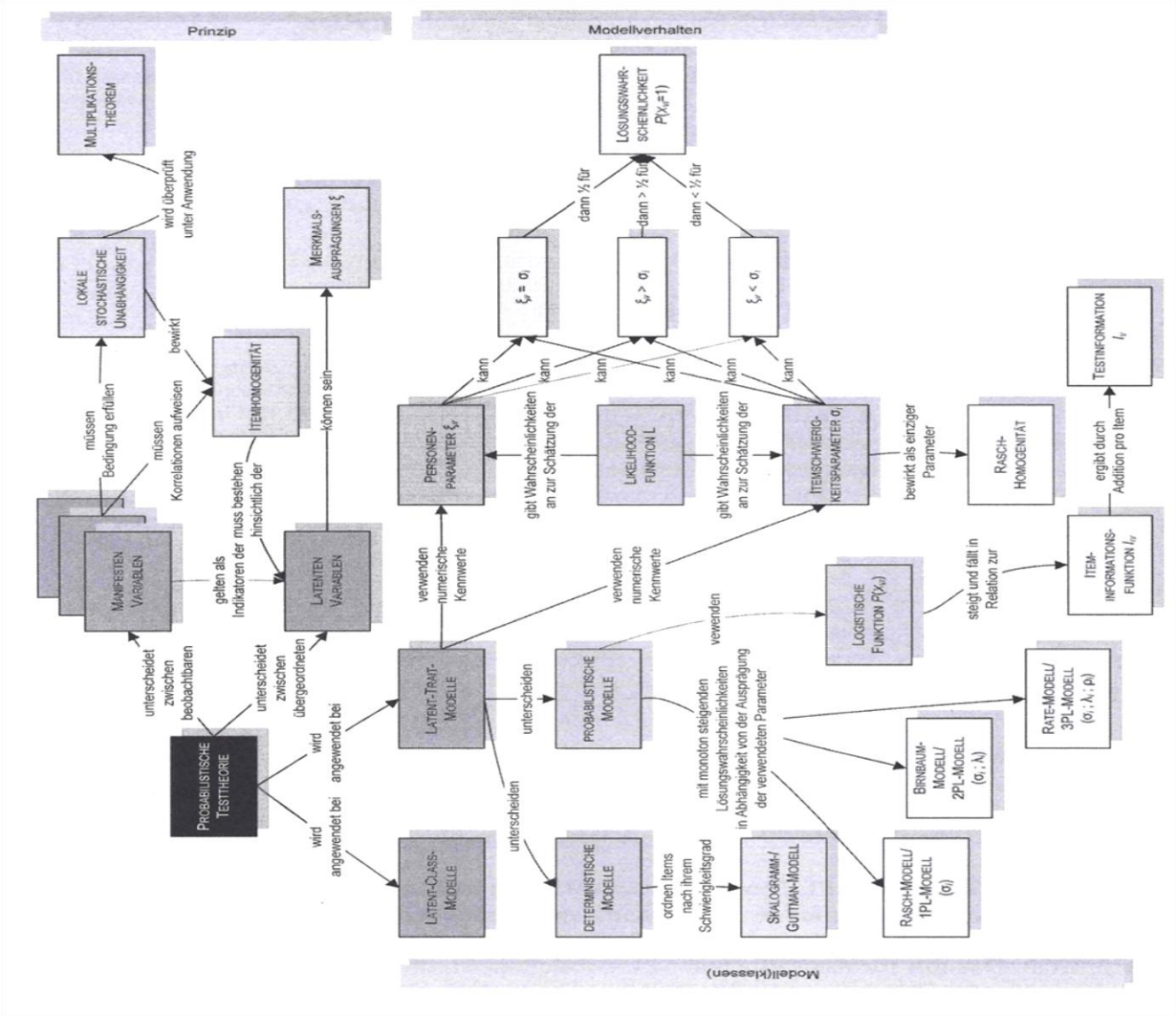


Abb. 1.3 „Latent Trait Modelle“ in der Probabilistischen Testtheorie im Überblick (Pospeschill 2010, 134)

Mit probabilistischen Verfahren lassen sich aus dichotomen Daten dann quantitative Daten machen, wenn z.B. ein Rasch-Modell angepasst werden kann. Mit der PTT lassen sich zudem Veränderungsmessungen modellieren, die im Rahmen der KTT nicht möglich sind. Tab. 1.1 zeigt Verfahren, die sich der KTT und PTT zuordnen lassen im Vergleich ihrer Leistungsfähigkeit.

Test-theorie	Orien-tierung	Statistische Verfahren	Kate-goriale Daten	Metrische Daten	Was wird messtheoretisch geprüft?	Was wird strukturtheoretisch geprüft?
PTT	Wahrschein lichkeit von Modellen	1-PL-Rasch-Modell	Beginnt hier	Endet hier	Ob aus kategorialen Daten metrische werden	Ob ein Summenwert aller Items alle Verhältnisse aller Items untereinander empirisch erklärt
PTT		Latente Klassen-analyse	Beginnt und endet hier		Ob es für kategoriale Daten ein passendes Klassenmodell gibt	Welche Kriterien/ Kriterienkombinationen, die Gruppen unterscheiden
PTT		Ordinales Rasch-Modell (Partial-Credit)		Beginnt bei ordinalen und endet bei intervallskalierten Daten	Ob aus ordinalen Daten intervallskalierte werden	Ob ein Summenwert aus ordinalen Daten aller Items die Verhältnisse aller Items empirisch erklärt
KTT	Korrelatio-nen	Cronbach's alpha		Setzt metrische Daten voraus	Verhältnis der Varianz der Einzelitems zur Varianz der Summen	Nichts
KTT		Faktorenanalyse (CFA)		Kann ordinale Daten verarbeiten	Ordinale Daten können als ordinal behandelt werden	Ob es mehrere Paare von Items gibt, die mehr zusammenhängen als andere Paare von Items

Tab. 1.1 Was leisten KTT und PTT im Überblick anhand ausgewählter Verfahren?

3. Was leistet das Rasch-Modell konkret als probabilistisches Verfahren im Vergleich zur Anwendung der internen Konsistenzanalyse mit Cronbachs alpha oder der Faktorenanalyse als Vertreter der klassischen Testtheorie?

Für einen Likert-skalierten Datensatz zum Modul 1 „Mobilität" des NBA (N=5131), den Bensch im Rahmen ihrer Promotion am Lehrstuhl für Statistik und standardisierte Verfahren der PTHV erhoben hat (vgl. Bensch 2012, in diesem Band, S. 118), wird die Berechnung von Cronbachs Alpha, eine Faktorenanalyse (beides KTT) sowie die Berechnung eines ordinalen (Partial Credit Modell) Rasch-Modells (PTT) durchgeführt und verglichen. In diesem Vergleich der varianzbasierten Verfahren (Reliabilitäts- und konfirmatorische Faktorenanalysen) der klassischen Testtheorie mit dem wahrscheinlichkeitsbasierten Rasch-Modell wird gezeigt, welchen Nutzen die Verfahren haben.

Reliabilität: Cronbachs alpha

Eine zentrale Aufgabe des Wissenschaftlers bei der Entwicklung standardisierter Instrumente ist die Auswahl passender Items aus einer theoretisch begründbaren größeren Anzahl von bislang ungeprüften Items. Die Itemauswahl kann sich an unterschiedlichen Kriterien orientieren.

Bei der Analyse einer Skala in der Phase der Instrumentenentwicklung gibt es z.B. die Möglichkeit, Items nach ihrem Beitrag zur Reliabilität einer Skala auszuwählen. Eine Methode der Item-Selektion ist es, den Wert von Cronbachs alpha zu maximieren. Cronbachs alpha wird dann nach der folgenden Formel berechnet:

$$\alpha = \frac{c}{c-1} \bullet \left(1 - \frac{\sum_{i=1}^{j} S_i^2}{S_x^2} \right)$$

wobei die Kürzel Folgendes bedeuten:

$S_i^2 = \text{Varianz der einzelnen Testitems}$
$S_x^2 = \text{Varianz des Gesamttests}$
$c = \text{Anzahl der Testitems}$
$j = \text{Testitem1 bis c}$

Zur Selektion der Items wird die „part-whole“ korrigierte Trennschärfe berechnet, die aus einer Korrelation eines jeden einzelnen Items mit dem Gesamttestwert berechnet werden kann. Liegt dieser Wert für ein Item unter einem Wert von 0.4, dann sollte das Item aus der Skala entfernt werden.

Cronbachs alpha sagt dabei über die Struktur der Skala nichts aus, sondern setzt die Varianz zwischen den Itemwerten ins Verhältnis zur Varianz zwischen den Gesamtergebnissen der Testpersonen.

Cronbachs Alpha testet nicht die Eindimensionalität von Skalen und kann auch bei mehrdimensionalen Skalen hoch sein.

Trotzdem wird diese „part-whole“ Trennschärfe oft allein als Methode der Item-Selektion eingesetzt (Erhart et al. 2009, S. 476).

Bei einer Item-Zahl unter sechs kann nach einer Reliabilitätsanalyse eine konfirmatorische Faktorenanalyse weiterhelfen, um zu überprüfen, ob wir es bei einer Skala wirklich mit einer eindimensionalen Skala zu tun haben (Bühner 2006, S.134). In

der Pflegewissenschaft wird das oft nicht berücksichtigt (Huber 2008; Panfil 2004, S. 43; Müller-Staub 2010).

Konfirmatorische Faktorenanalyse

In der konfirmatorischen Faktorenanalyse (CFA) wird versucht, die beobachteten Variablenausprägungen aufgrund von latenten Variablen und Messfehlern zu erklären. Im Fall der Mobilitätsskala des NBA bedeutet dies, dass sich die Varianz der fünf Mobilitätsitems aus einem gemeinsamen Faktor „Mobilität" heraus erklären lassen müsste.

Die konfirmatorische Faktorenanalyse prüft „hypothetisch angenommene Strukturen in Form eines Modells auf ihre Übereinstimmung mit den beobachteten Zusammenhängen. Hierzu werden die Beziehungen zwischen den manifesten und latenten Variablen eines Modells spezifiziert und auf Grundlage dieser Informationen und der beobachteten Zusammenhänge die theoretisch noch nicht bestimmten Parameter des Modells geschätzt. Für die Durchführung einer CFA ist es daher erforderlich, das zugrunde gelegte Modell zu spezifizieren und die Methode der Parameterschätzung sowie die Verfahren zur Evaluation des Modells zu bestimmen (Bühner 2006; Moosbrugger, Schermelleh-Engel 2008)." (Franken 2010, S. 100, f.)

Die Voraussetzungen, ein Modell spezifizieren zu können liegt in der Definition der Beziehungen der Variablen zueinander. Möglich ist dies in einem Pfaddiagramm oder in einem Gleichungssystem. In Abb. 1.4 werden manifeste Variablen durch ein Rechteck, latente Variablen durch eine Ellipse, gerichtete Beziehungen (partielle Regressionsgewichte) durch einen einfachen, ungerichtete Beziehungen (Korrelationen, Kovarianzen) durch einen Doppelpfeil dargestellt. Wendet man diese Grundgedanken auf die Subskala „Mobilität" des NBA an, so ergibt sich aus der Annahme der Eindimensionalität der Subskala das Modell der Abb. 1.4 für eine

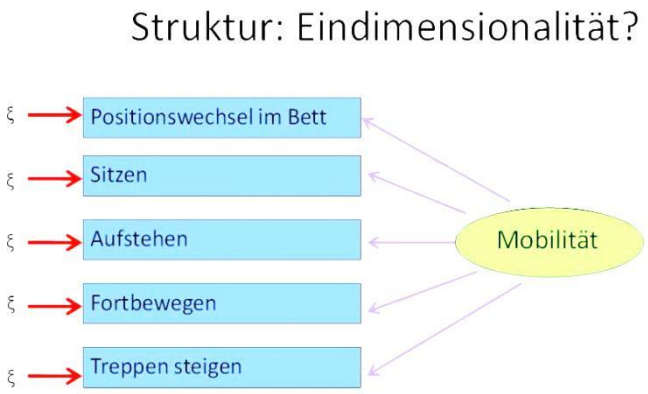


Abb. 1.4 Modell einer CFA mit einer latenten Variablen und fünf manifesten Variablen (die fünf Pfeile links symbolisieren individuelle Messfehler)

konfirmatorische Faktorenanalyse der Mobilität:

Als Gütekriterium der konfirmatorischen Faktorenanalyse wird der Chi²-Wert der Modellanpassung[13] genutzt. Je größer der Chi²-Wert, desto weniger stimmen das geprüfte (hier das eindimensionale) theoretische Modell und die empirischen Daten überein. Als Schätzverfahren für die Annäherung von empirischer und theoretischer Kovarianzmatrix werden die Weighted Least Squares[14] (WLS) eingesetzt, da eine große Stichprobe und keine normalverteilten[15] Daten vorliegen. Wegen der Größe der Stichprobe muss der berechnete Chi²-Wert (χ^2) als Ablehnung der H_0-Hypothese[16] interpretiert werden, wenn er doppelt so groß ist wie die Freiheitsgrade des Modelltests, um dem Umstand Rechnung zu tragen, dass bei großen Stichproben auch kleinste Abweichungen zwischen theoretischer und empirischer Verteilung im Signifikanztest bedeutsame Abweichungen von der H_0-Hypothese produzieren (Moosbrugger, Schermelleh-Engel 2008, S. 319).

Bei unzureichender Modellgüte ist es aufschlussreich, sich die standardisierten Residuen anzusehen. Sie können Hinweise darauf geben, wo die Probleme einer Skala liegen könnten, weil sie für einzelne Variablenpaare und damit inhaltlich eine Richtung der Abweichungen zwischen theoretischem Modell und empirischen Daten angeben.

Die Strukturprüfung mittels der CFA basiert auf Variablenpaaren und konstituiert aufgrund deren paarweiser wechselseitiger Einflüsse das Strukturgleichungsmodell. Hierbei werden keine Beziehungen höherer Ordnung (über paarweise Beziehungen hinaus) auf der Basis von Interaktionseffekten aller Variablenkombinationen miteinander berücksichtigt.

Residuen ergeben sich aus der Differenz zwischen der beobachteten und geschätzten Kovarianz. Standardisierte Residuen ergeben sich aus dem Verhältnis der Residuen zum geschätzten Standardfehler. Sie liefern ein statistisches Maß, um die Größe der Residuen zu beurteilen (Jöreskog, Sörbom 1993, S. 126 in Franken 2010, S. 130).

[13] Die Modellanpassung sagt etwas darüber aus, wie gut die empirischen Daten und die theoretisch durch das getestete Modell berechneten Daten zusammenpassen.

[14] Weighted Least Squares sind als gewichtete kleinste Quadrate eine Methode, um die Abweichung zwischen theoretischem und empirischen Modell zu berechnen.

[15] Normalverteilte Daten haben in ihrer idealen Ausprägung die Eigenschaft, dass sie symmetrisch sind und der häufigste Wert, der mittlere Wert und der Mittelwert in einem Wert zusammenfallen.

[16] Die H_o-Hypothese (Nullhypothese) geht entgegengesetzt der Alternativhypothese davon aus, dass sich zwei zu vergleichende Datensätze oder Stichproben nicht unterscheiden, bzw. dass keine Zusammenhänge oder Unterschiede signifikant wahrscheinlich sind.

Nr.	Item	selbständig	überwiegend selbständig	überwiegend unselbständig	unselbständig
1	Positionswechsel im Bett	☐0	☐1	☐ 2	☐ 3
2	Stabile Sitzposition halten	☐0	☐1	☐ 2	☐ 3
3	Aufstehen aus sitzender Position/Umsetzen	☐0	☐1	☐ 2	☐ 3
4	Fortbewegen innerhalb des Wohnbereichs	☐0	☐1	☐ 2	☐ 3
5	Treppensteigen	☐0	☐1	☐ 2	☐ 3

Tab. 1.2 Modul 1 „Mobilität“ des NBA (Wingenfeld et al 2008, S.35)

Rasch-Modell (PTT)

Im quantitativen NBA-Messmodell sollen u. a. die fünf Items einer Mobilitätsskala (siehe Tab. 1.2) summiert werden, so dass aus den jeweiligen Summenwerten eine „Größer-Kleiner“-Ausprägung des Konstrukts „Mobilität“ ableitbar ist.

Der Summenwert aus den empirischen Werten der Einzelitems kann jedoch ausschließlich als Aussage für eine quantitative Merkmalsausprägung verwendet werden, wenn notwendige messtheoretische Voraussetzungen erfüllt sind. Diese lassen sich mit dem Rasch-Modell überprüfen.

Das Rasch Modell ist ein Verfahren der probabilistischen Testtheorie. Es eignet sich besonders zur Entwicklung von Messinstrumenten, die latente Konstrukte in ihrer Ausprägung messen sollen. Es wurde in der Pflege bereits mehrfach eingesetzt (Hagquist 2009; Brühl, Berger 2011).

Die Grundidee des Rasch-Modells, übertragen auf die o.g. Mobilitätsskala, lautet: Die Ausprägung des Merkmals „Selbständigkeit in der Mobilität“ ist abhängig von den Schwierigkeiten der fünf Items (vgl. Tab. 1.3) und dem Fähigkeitsgrad der zu testenden Person. Je schwieriger ein Kriterium (Item) ist, desto geringer ist die Wahrscheinlichkeit, dass eine potentiell pflegebedürftige Person dies erfüllt. Je mobiler eine Person ist, desto höher ist die Wahrscheinlichkeit, viele oder alle der Mobilitätskriterien zu erfüllen.

Bei der Verwendung des Rasch-Modells werden die tatsächlich produzierten Verteilungen der erhobenen Daten vieler Personen mit den theoretisch aus dem Rasch-Modell berechneten Verteilungen verglichen. Stimmen die Verteilungen der erhobenen Daten mit den theoretischen Verteilungen annähernd überein, wird von einer Gültigkeit des Rasch-Modells gesprochen.

Gilt das Rasch-Modell nicht, dürfen die Items nicht summiert werden, da die Summe keinen korrekten Index für das Ausmaß der Mobilität (und damit einem wesentlichen Teil der Pflegebedürftigkeit) einer Person darstellt.

Zur Erläuterung des Rasch-Modells ist ein Blick auf eine sogenannte Datenmatrix hilfreich, die aus Spalten und Zeilen besteht. Für die Erklärung wurde die ordinalskalierte Mobilitätsskala (1-4) in eine dichotome Skala (0/1) überführt (zur Vorgehensweise siehe Bensch Kapitel 5 in diesem Band). Tab. 1.3 zeigt eine Datenmatrix, in der die fünf NBA-Mobilitätskriterien für fünf Personen (A – E) dargestellt sind. Mit 1 werden die Items bewertet, in denen die Person „selbständig“ ist (das bedeutet, das Item wurde „gelöst“ oder erfüllt). Personen, die mit „überwiegend selbständig“, „überwiegend unselbständig“ oder „unselbständig“ bewertet wurden, werden in dieser Matrix mit einer 0 für „unselbständig“ dargestellt (das heißt, das Item konnte nicht „gelöst“ werden).

Modul 1 "Mobilität" des NBA		Personen					Summe der Personen, für die das Item mit "unselbständig" bewertet wurde
		A	B	C	D	E	
1	Stabile Sitzposition halten	1	1	1	1	1	0
2	Positionswechsel im Bett	1	1	1	1	0	1
3	Aufstehen aus sitzender Position/Umsetzen	1	1	1	0	0	2
4	Fortbewegen innerhalb des Wohnbereichs	1	1	0	0	0	3
5	Treppensteigen	1	0	0	0	0	4
	Summenwert der Person	5	4	3	2	1	

Tab. 1.3 Guttmann-Skala zum Modul 1 „Mobilität“ des NBA

In der letzten Spalte steht die Anzahl der Personen, bei denen das Kriterium als nicht gelöst eingestuft wurde, die demnach einen Pflegebedarf haben. In der untersten Zeile ist die Anzahl der Kriterien aufsummiert, die die Person lösen konnte, bzw. in der sie selbständig und nicht auf Hilfe angewiesen ist.

Für Person A kann die höchste Mobilität bescheinigt werden, da sie in allen Kriterien selbständig ist. Mit nur einem erfüllten Kriterium ist Person E die Person, bei der die Mobilität am schlechtesten bewertet wurde. Die in Abb. 7 gezeigte Matrix folgt einer Guttman-Skala[17], in der die personenbezogenen Summenwerte entsprechend der impliziten Hypothese des NBA kontinuierlich und geordnet vom leichtesten über das nächstschwierigere Kriterium ansteigen. Die am wenigsten mobile Person E unterscheidet sich von Person D dadurch, dass bei Person D neben dem mutmaßlich leichtesten Kriterium („Positionswechsel im Bett“) auch das mutmaßlich zweitleichteste Kriterium („Stabile Sitzposition halten“) erfüllt ist. Person D erzielt dadurch einen doppelt so hohen Summenwert wie Person E. Genau so nimmt die Mobilität der Personen von Person E über die Personen D, C, B und A immer über das nächstschwierigere Kriterium zu.

Das gleiche gilt auch für die Kriterien, deren Erfüllung von der mobilsten zur am wenigsten mobilen Person kontinuierlich abnimmt.

Eine Guttman-Skala kommt bei Messungen komplexer Konstrukte allerdings kaum vor. Daher ist es erforderlich, das Ausmaß der Abweichung von dieser Annahme zu überprüfen.

Die Funktionsweise des Rasch-Modells lässt sich anhand dieses Beispiels einfach erläutern: Aus den jeweiligen Randsummen der Personen und der Kriterien lässt sich die Wahrscheinlichkeit berechnen, mit der bei einer konkreten Person ein bestimmtes Kriterium erfüllt ist.

So liegt die errechnete Differenz (und damit die Schwierigkeit des Kriteriums für die Person B) zwischen der Randsumme der Person B und der Randsumme des Kriteriums 2 mit dem Wert 3 (4 – 1 = 3) relativ hoch. Somit ist die Wahrscheinlichkeit ebenfalls hoch, dass Person B das Kriterium 2 erfüllt. Im Rasch-Modell wird also auf der Basis von Summenwerten für Personen und Items berechnet, mit welcher theoretischen Wahrscheinlichkeit für eine Person mit einem bestimmten Summenwert ein konkretes Kriterium vorliegt. Genau diese Wahrscheinlichkeiten berechnet das

[17] Die Guttman-Skala (GS) entspricht zwar einem deterministischen Modell, bei dem jeder Fähigkeitsausprägung bei jeder Aufgabe nur die Lösungswahrscheinlichkeit „1“ oder „0“ zugeordnet werden kann. Die GS unterscheidet sich vom Rasch-Modell, da beim Rasch-Modell die Lösungswahrscheinlichkeiten einer Fähigkeitsausprägung pro Kriterium variieren kann. Hier ziehen wir die GS heran, um in die Funktionsweise des Rasch-Modells einführen zu können.

Rasch-Modell für alle Personen und alle Kriterien. Im Ergebnis wird dann verglichen, wie sehr sich tatsächliche empirische Lösungswahrscheinlichkeiten von den theoretischen unterscheiden.

Diese Darstellung dient der Einführung[18]. Tatsächlich ist das Verfahren komplexer. Es werden Parameter für Personen und Kriterien in iterativen Verfahren[19] mit Algorithmen geschätzt. Jedem Summenwert einer Person und jedem Summenwert eines Items wird ein transformierter Wert auf der gleichen sogenannten Logit-Skala zugeordnet. Zur Einführung bleiben wir bei einfachen Summenwerten.

Eine Person, die in der oben dargestellten Datenmatrix einen Summenwert von „vier" erreicht kann vier Kriterien selbständig durchführen und ist damit recht mobil. Die Wahrscheinlichkeit, dass eine solche Person auch das relativ leichte Kriterium 2 selbständig bewältigen kann, ist deshalb ebenfalls hoch. Damit ist auch die Wahrscheinlichkeit hoch, dass grundsätzlich bei vier erfüllten Kriterien das Kriterium 2 beinhaltet ist.

Unter der Voraussetzung, dass jedes Item die gleiche Trennschärfe[20] besitzt und jedes Item das gleiche Konstrukt misst (hier: Mobilität), bedeuten höhere Summenwerte eine höhere Ausprägung des Konstrukts, also eine bessere Mobilität als niedrigere Summenwerte. Damit enthalten die Summenwerte alle Informationen und werden als „erschöpfende Statistiken" bezeichnet.

Mit dem ein-parametrisch-logistischen Rasch-Modell kann also statistisch geprüft werden, ob eine Summenbildung messtheoretisch zulässig ist.

Partial-Credit-Modell (PTT)

Das einfache ein-parametrisch-logistische Rasch-Modell kann nur angewandt werden, wenn als Antworten dichotome Werte vorliegen. Für ordinale Daten, wie sie die Antwortskala beim NBA produziert, muss auf eine Variante des Rasch-Modells ausgewichen werden, die ordinale Daten verarbeiten kann: Das so genannte Partial-Credit-Modell**.**

[18] Für eine detaillierte Einführung zum Rasch-Modell sei auf Rost (2004) oder Strobl (2010) verwiesen.

[19] Eine schrittweise und zielgerichtete Annäherung an eine möglichst exakte Lösung durch wiederholte Anwendung des gleichen Rechenverfahrens.

[20] Ein trennscharfes Item differenziert selbständige von unselbständigen Probanden entsprechend der Schwierigkeit des Items: Items, die schwierig selbständig zu erfüllen sind werden nur von selbständigen Probanden erfüllt, leichte Items können auch noch von eingeschränkten Probanden erfüllt werden. Items die zu leicht (alle sind selbständig) oder zu schwierig sind (alle sind unselbständig) sind ungeeignet, die Probanden in selbständige und unselbständige zu differenzieren; sie sind nicht trennscharf.

Das ein-parametrische Rasch-Modell für dichotome Daten wurde nach Rost von Andrich und von Masters (Rost 2004, S. 214) in getrennten Arbeiten für die Analyse ordinaler Daten erweitert und unter dem Namen Partial Credit Modell veröffentlicht (Rost 2004, S. 201 ff). Das Modell fand bislang in der Kompetenzmessung Anwendung. Mit dem Begriff des „Partial Credit“ wird angezeigt, dass Schüler, die nicht vollkommen korrekte Leistungen zeigen, von der Gesamtpunktzahl (full credit) einen Anteil (partial credit) erhalten (Walter 2005, S. 36).

Probabilistische Modelle und damit auch das Rasch-Modell für Items mit ordinalen Antwortkategorien bestimmen die Wahrscheinlichkeit p, dass eine bestimmte Itemantwort z.B. auf die Frage „Können Sie Treppensteigen?“ gegeben wird. Gilt das Rasch-Modell für ordinale Antworten, ergibt sich die Wahrscheinlichkeit der Antwortmöglichkeit „selbständig“ aus der Differenz der Fähigkeit θ (theta) der Person „v“, also θ_v und der Schwierigkeit σ (sigma) der Itemantwortstufe „x“ beim Item i also σ_{ix}. Beides, Personenfähigkeiten und Itemschwierigkeiten werden auf derselben Skala abgetragen, die die Ausprägung der latenten Variablen „Mobilität“ wiedergibt. Ein Item, das mit einer vierstufigen Antwortskala beantwortet werden kann, hat drei „Schwellen“, die beim Wechsel von einer zur nächsten Antwortkategorie überwunden werden müssen. Das Verhältnis von den drei Antwortschwellen eines vierstufigen Items im ordinalen Rasch-Modell zueinander sollte geordnet sein, d. h. die Fähigkeit einer Person nimmt von der Antwortmöglichkeit „selbständig“(= 0) über „überwiegend selbständig“(= 1), „überwiegend unselbständig“(= 2) bis hin zu „unselbständig“(= 3) kontinuierlich ab. Eine solche geordnete Abfolge von Antwortwahrscheinlichkeiten für die vier Antwortmöglichkeiten abgetragen auf der Logit-Skala enthält als Beispiel die Abb. 1.5 für das Item „Positionswechsel im Bett“ der Mobilitätsskala.

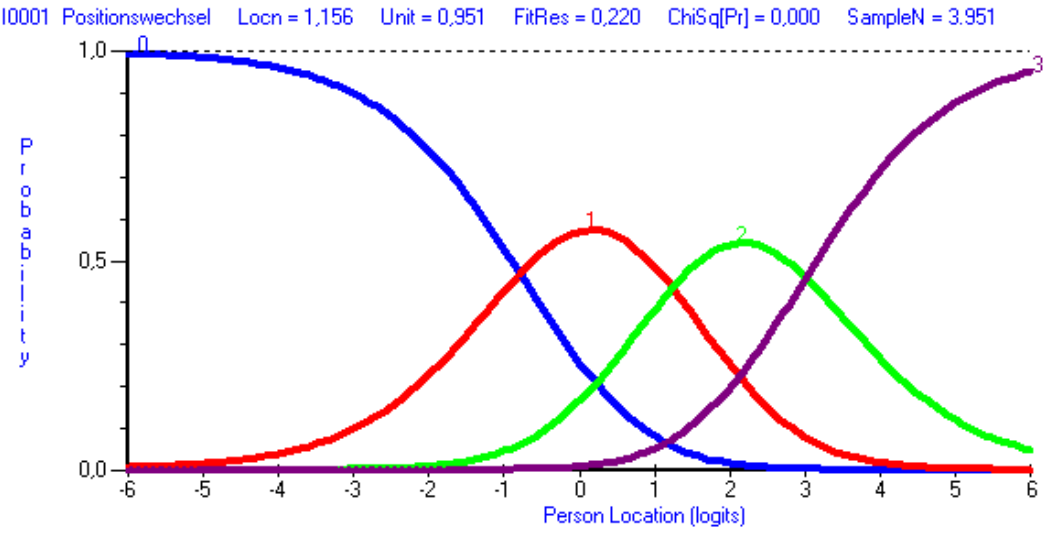


Abb. 1.5 Antwortwahrscheinlichkeiten für das Item „Positionswechsel im Bett“

Aus den Wahrscheinlichkeiten der vier Antwortmöglichkeiten werden drei Schwellenwertfunktionen abgeleitet, die jeweils die Wahrscheinlichkeiten für den Übergang von Antwortmöglichkeit 0 (= selbständig) zur Antwortmöglichkeit 1 (= überwiegend selbständig), von 1 zu 2 (= überwiegend unselbständig) und von 2 zu 3 (= unselbständig) kennzeichnen.

Auf der x-Achse wird Unselbständigkeit abgetragen. Bei -6 liegt ein sehr niedriger Wert von „Unselbständigkeit“ vor. Personen mit einem Wert von -6 sind also sehr selbständig, weil die Antwortkategorie 0 = selbständig für den „Positionswechsel im Bett“ bei ihnen die höchste Wahrscheinlichkeit hat. Es existierten drei Schwellenwerte, das sind die Schnittpunkte der jeweiligen Antwortkategorien 0 mit 1, 1 mit 2 usw. An diesen Schwellen wechselt die jeweils wahrscheinlichste Antwortkategorie, z. B. bei ca. -0.8 die häufigste Antwortkategorie von „selbständig“ zu „überwiegend selbständig“.

Die Wahrscheinlichkeit p, dass einer Person v mit der Fähigkeit θ_v und der Schwierigkeit σ_{ix} (einer Itemantwortstufe [x] eines Items [i]) die Antwortmöglichkeit X_{vi}= x zugeordnet wird, kann mit der folgenden logistischen Funktion (Rost 2004, S. 209) errechnet werden:

$$p(X_{vi} = x) = \frac{\exp(x\theta_v - \sigma_{ix})}{\sum_{s=0}^{m} \exp(s\theta_v - \sigma_{is})}$$

Dabei zeigt m die Anzahl der Antwortkategorien minus 1, also die zu überwindenden Schwellen des items i an. θ_v ist der Fähigkeitsparameter theta der Person v und σ_{ix} ist der kumulierte Schwellenparameter x des Items i[21].

An dieser Stelle wird bereits deutlich, dass schon das Partial-Credit-Modell im Gegensatz zum einfachen ein-parametrisch-logistischen Rasch-Modell (1-PL) eine Quantifizierbarkeit des Personenparameters annimmt. Die Testung der Quantifizierbarkeit eines Merkmals ist nicht mehr so von Grund auf möglich wie im 1-PL-Rasch-Modell, da im Partial-Credit-Modell aus dem qualitativ-mehrdimensionalen

[21] Wobei der kumulierte Schwellenparameter des Items i σ_{is} für ein s = 0, also für null zu überwindende Schwellen ebenfalls 0, und auch θ_v, also die Fähigkeit theta der Person v mal 0 und damit der ganze erste Summand aus dem Summenzeichen des Nenners wegen exp(0 x θ_v +0) = 1 immer 1 wird (Rost 2004, S. 208f), vereinfacht ausgedrückt ermöglicht es die oben stehende Formel für jede Person v die Wahrscheinlichkeit auszurechnen, mit der ihr die Antwort x im Item i zugeordnet wird. Im Zähler steht die konkrete Differenz zwischen der jeweiligen Fähigkeit, die Antwort x bedeutet, nämlich x mal die Fähigkeit Theta der Person v = x σ_v und der kumulierten Schwierigkeit der Antwort x bei item i. Im Nenner steht die Summe aller dieser Differenzen über die m Antwortschwellen hinweg.

Personenparameter des einfachen 1-PL-Rasch-Modells eine quantitativ-eindimensionale Variable wird. Die ordinale Skalierung der Antworten erhöht somit die Anforderungen an die Exaktheit, mit der hier die Personenvariable Mobilität gemessen werden muss, wenn man sie in der vorliegenden Form quantifizieren will (vgl. hierzu Rost 2004 S. 209). Da bei den fünf Mobilitätsitems von vornherein eine ordinale und damit eine quantitative Antwortskala vorliegen soll, kann der Vorteil des einfachen dichotomen Rasch-Modells, eine Quantifizierung von kategorialen Daten her beginnend modellieren zu können, nicht mehr voll zur Anwendung gebracht werden. Eine nachträgliche Dichotomisierung der ordinalen NBA-Antworten zur Mobilität in zwei Antworten „selbständig“ und „unselbständig“ ist möglich, wirft aber immer die Frage auf, ab welchem Punkt die Antworten der vierer-Skala der Kategorie „selbständig“ und ab welchem Punkt sie der Kategorie „unselbständig“ zugeordnet werden sollen.

Aus dem ordinalen Rasch-Modell kann man für jede Person und für jede Antwortstufe eines jeden Items einen theoretischen Erwartungswert berechnen und ihn mit den beobachteten Werten vergleichen. Passt ein solches Partial Credit Modell auf einen empirischen Datensatz, so sind die Antwortschwellen der Antwortskala geordnet und ein Summenwert kann sinnvoll berechnet werden.

Die Prüfung, ob das in der Auswertung berechnete Modell überhaupt zu den empirischen Daten passt, erfolgt mit der Hilfe von Modellgeltungstests. Die Gültigkeit eines Modells kann sowohl in Bezug auf das Gesamtmodell als auch für die Gültigkeit einzelner Items evaluiert werden.

Die verfügbaren unterschiedlichen Testverfahren zur Prüfung der Modellgüte liefern alleine meist keine eindeutige Antwort auf die Frage, ob das berechnete Modell hinreichend zu den empirischen Daten passt, weil sich in der Berechnung von Wahrscheinlichkeiten lediglich Tendenzen abbilden lassen. Die Ergebnisse der Modellgeltungstests sind damit stets in einem inhaltlich-fachlichen Kontext des untersuchten Konstrukts zu interpretieren.

Wir vergleichen die Modellschätzungen mit verschiedenen Algorithmen und wenden hierzu verschiedene Programme an. Mit der software eRm in R, RUMM 2020 und 2030 wurden die Daten für N=3951 verarbeitet, da das Programm die extremen Antwortmuster (die, die bei allen Items [un]selbständig angegeben haben) ausschließt. Mit Winmira wurden die Ergebnisse für N= 5080 berechnet, da 51 Antwortmuster mit fehlenden Werten ausgeschlossen wurden[22].

[22] Als Gesamtmodellgeltungstests wurde der Gesamt-Modell chi²-Test und der Andersen Likelihood-Ratio-Test durchgeführt und als Tests für die Passung der einzelnen Items wurden die Item-chi²-Werte, die absolute Größe der Residuen und die Z-Werte des Wald-Tests eingesetzt. Es wurden die Schätz-Algorithmen des Programmpakets eRm aus „R“ (Unconditioned Maximum Likelihood) verwendet und die

ERGEBNISSE DES METHODENVERGLEICHS

Ergebnis der Prüfung der Mobilitäts-Skala mit Cronbachs alpha: Hohe Reliabilität!

Für die Items der Mobilitätsskala ergibt sich der Cronbachs alpha Wert 0.93, sowohl für die einfachen als auch die standardisierten[23] Item-Werte. Der Wert ist also in beiden Varianten vergleichbar hoch und kann in seiner standardisierten Variante auch mit alpha-Werten fünf- oder siebenstufiger Skalen verglichen werden. Ein solcher Wert wird in der Pflegewissenschaft oft alleine als Hinweis auf die „guten psychometrischen Eigenschaften“ (Panfil 2004, S. 43) einer Skala interpretiert.

In der Trennschärfebestimmung liegen alle alpha-Werte für die fünf Items der Mobilitäts-Skala oberhalb des Wertes von 0.4, so dass alle Items in der Skala verbleiben sollten. Auch hier würde keine Änderung der Mobilitätsskala vorgenommen werden müssen.

Die mittlere Item-Interkorrelation ist mit 0.72 ebenfalls hoch.

Bis hierher ergäbe sich also eine sehr gute Reliabilität der Mobilitätsskala. Nach diesem einfachen Einsatz von Cronbachs alpha wäre die Mobilitätsskala als sehr gut zu bewerten. Verfahren wie z.B. Cronbachs alpha, die alleine auf bivariaten Korrelationen basieren, würden kein einziges Item aus der Skala ausschließen und der Skala mit 0.93 eine sehr gute Reliabilität bescheinigen.

Eine Prüfung der Quantifizierbarkeit des Konstrukts liefert die Reliabilität nicht, sie setzt quantitative Daten bereits voraus. Über die Struktur der Skala und deren Validität sagt diese Prüfung nichts aus.

Konfirmatorische Faktorenanalyse (CFA): Keine eindeutige Passung eines eindimensionalen Modells

Mit der Wurzel aus der Varianz der Item-Interkorrelationen kann die Exaktheit des berechneten Cronbachs alpha bestimmt werden. Liegt der Wert über 0.01, so kann eine konfirmatorische Faktorenanalyse durchgeführt werden, um zu überprüfen ob hier tatsächlich eine eindimensionale Skala vorliegt. Der Wert für die Exaktheit von

---

Schätz-Algorithmen aus dem Programm RUMM2020 (=Pairwise). Zusätzlich wurde das Programm „Winmira“ eingesetzt, weil es den WLE-Algorithmus nutzt und mit dem so genannten „Q-Index“ (Rost 2004, 373; Rost et al 1994, 171f.) einen weiteren Index zur Testung der Übereinstimmung zwischen modellierten und empirischen Verteilungen anbietet.

[23] Ein standardisierter Wert berücksichtigt unterschiedliche Standardabweichungen

Cronbachs alpha liegt mit 0.04 deutlich über 0.01, so dass zur Absicherung, ob es sich hier um eine eindimensionale Skala handelt, eine konfirmatorische Faktorenanalyse durchgeführt wird.

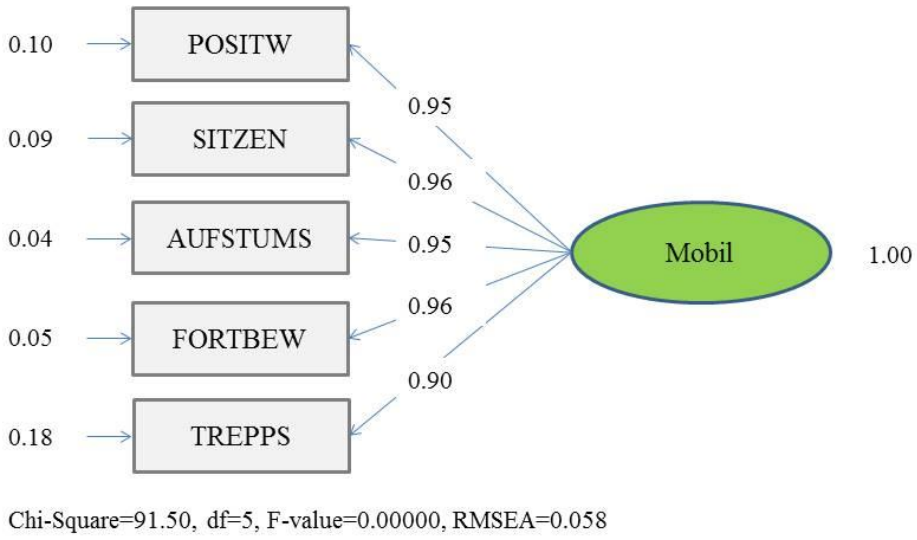


Abb. 1.6 Pfadanalyse zur konfirmatorischen Faktorenanalyse

Die CFA weist daraufhin, dass die Subskala „Mobilität“ trotz der sehr hohen Korrelationen zwischen den Variablenpaaren nicht eindeutig eine eindimensionale Skala ist. Ein Verfahren, um ordinale Daten in intervallskalierte zu überführen und diese Überführung auch zu testen, liefert die CFA nicht. Sie kann die Unterschiede, die ordinale Daten im Vergleich zu intervallskalierten produzieren jedoch mit Hilfe der polychorischen Korrelationen berücksichtigen (vgl. Schröder, 2010; Franken 2012, in diesem Band, S. 92). Eine Hilfe bei der Quantifizierung des Merkmals bietet die CFA nicht.

Das Pfaddiagramm in Abb. 1.6 zeigt die sehr hohen Ladungen der fünf Items auf der latenten Variablen „Mobilität“ in einer CFA, die auf der Basis polychorischer Korrelationen[24] mit der Methode der Weighted Least Squares (WLS) berechnet wurde.Der chi² (χ^2) -Test weist das einfaktorielle Modell für einen exakten Modell-Fit eindeutig zurück (χ^2 [df: 5]= 91.50; p = 0.00). Auch die Ergebnisse weiterer Modellgeltungstest[25] lassen Zweifel aufkommen, dass das theoretische Modell der

[24] Die polychorische Korrelation berechnet mit Hilfe von geschätzten Schwellenwerten (durch Maximum-Liklihood-Berechnungen), den Zusammenhang beobachteter, nominal- oder ordinalskalierter Variablen, von denen angenommen wird, dass sie latente Merkmale metrisch abbilden und einen linearen Zusammenhang aufweisen. Polychorische Korrelationen dienen dazu, ordinale Daten auch wie ordinale Daten zu behandeln.

[25] Auch das für einen approximativen Modell-Fit akzeptable Verhältnis des χ2 Werts zur Anzahl der Freiheitsgrade (χ2 ≤ 3df) wird nicht erreicht (Schermelleh-Engel et al. 2003, S. 52). Dies ließe sich mit der Verletzung der Normalverteilung und der Stichprobengröße erklären (Schermelleh-Engel et al. 2003, S. 32; Jöreskog 2002, S. 22). Der Root Mean Square Error of Approximation (RMSEA) beträgt 0.058 (p-Wert für RMSEA < 0.05 = 0.086). Der RMSEA entspricht so nur knapp einem akzeptablen Fit, da er 0.05 nicht

Eindimensionalität mit den empirischen Daten übereinstimmt. Tab. 1.4 zeigt die standardisierten Residuen zum untersuchten Modell.

Nr	Item	Positionswechsel im Bett	Stabile Sitzposition halten	Aufstehen aus sitzender Position/ Umsetzen	Fortbewegen innerhalb des Wohnbereichs	Treppensteigen
1	Positionswechsel im Bett					
2	Stabile Sitzposition halten	2.33				
3	Aufstehen aus sitzender Position/Umsetzen	-7.23	-8.00			
4	Fortbewegen innerhalb des Wohnbereichs	-7.99	-4.90	-3.86		
5	Treppensteigen	-6.78	-7.24	-3.13	0.71	

Tab. 1.4 Standardisierte Residuen des Moduls 1 „Mobilität“ des NBA

In den standardisierten Residuen zeigen sich hohe negative Abweichungen insbesondere bei den Kovarianzen aller Indikatoren außer zwischen „Treppensteigen“ und „Fortbewegen“. Danach überschätzt das angenommene Modell mit einer latenten Variablen die Kovarianzen zwischen den Variablen.

Dies verweist auf eine Fehlspezifikation (Jöreskog, Sörbom in Franken 2010, S. 130) Gerade die Zusammenhänge zwischen den Kriterien, die Mobilität „auf den eigenen Füßen“ erfragen und den Kriterien, die dies nicht tun, werden überschätzt. Die Korrelation der Items ist von vornherein so hoch, dass Schätzprobleme beim Einsatz der konfirmatorischen Faktorenanalyse auftreten.

Fasst man die Ergebnisse aus der Reliabilitätsanalyse zusammen, so wäre keine Änderung der Mobilitätsskala nötig. Die CFA liefert Hinweise, dass ein eindimensionales Modell nicht gut passt und vermittelt bei einer genaueren Analyse der paarweisen Residuen auch Hinweise, dass sich die Items mindestens in zwei

überschreiten sollte (Schermelleh-Engel et al., 2003, S. 36). Der SRMR beträgt 0.02 und wäre nach Schermelleh-Engel et al (2003, S. 38) auch noch akzeptabel. Die Fit-Indizes NFI, NNFI, CFI, GFI und AGFI betragen 1.00.

Gruppen unterteilen lassen: Ein Teil der Items setzt voraus, dass man „auf seinen Füßen stehen kann“ und ein anderer Teil der Items erfordert dies nicht.

Carstensen weist darauf hin, dass die innerhalb der klassischen Testtheorie üblichen Methoden auf bivariaten Zusammenhängen zwischen Variablenpaaren beruhen (Carstensen 2000, S. 22). Damit beruht die Testung der Dimensionen der Mobilitätssubskala mit einer CFA ebenfalls auf bivariaten Kovarianzmatrizes. Es sind Zusammenhänge zwischen den Variablenpaaren wie „Positionswechsel im Bett“ und „stabile Sitzposition halten“, die eine Rolle für die Modellgeltungstests einer CFA spielen, nicht aber gleichzeitig der Bezug dieses Variablenpaares zu einer dritten Variablen wie „Aufstehen aus sitzender Position/Umsetzen“ oder einer vierten Variablen wie „Treppensteigen“.

Probabilistische Modelle sind in der Lage, das Verhältnis aller Variablen zueinander zu berücksichtigen, was dem komplexen Konstrukt der Mobilität besser gerecht würde.

Ordinales Rasch Modell: Keine Anpassung an empirische Daten – Quantifizierung des Konstrukts Mobilität gelingt nicht

Als probabilistisches Verfahren verweist das ordinale Rasch-Modell auf eine nicht genügend exakte Operationalisierung der Mobilität. Die Skala sollte nach den Ergebnissen der Prüfung des Rasch-Modells nicht einfach summiert werden, da statistisch signifikante Prüfgrößen, z. B. ein hoher Chi-Quadrat-Wert bei vielen Freiheitsgraden dafür spricht, dass das getestete Modell von den empirischen Daten abweicht und damit das Rasch-Modell nicht gilt.

Die Annahmen des Rasch-Modells, d.h. die Eindimensionalität, die lokale stochastische Unabhängigkeit, die spezifische Objektivität, suffiziente Statistiken[26] und gleiche Trennschärfen (vgl. Bensch 2012, Kapitel 5 in diesem Band) sind im NBA-Modul „Mobilität“ offenbar verletzt worden. Da diese Annahmen jedoch die wesentlichen Voraussetzungen für die Summation von Items sind, bildet der Summenwert der Mobilitätsskala keinen validen Index für die Selbständigkeit/Unselbständigkeit des Probanden in Bezug auf seine Mobilität.

Die Prüfung ergibt also, dass das ordinale Rasch-Modell für die Mobilitätsskala des NBA eindeutig zurückgewiesen wird. Der Gesamt-Chi²-Wert liegt bei der Schätzung

[26] Der Summenscore einer suffizienten Statistik enthält alle (qualitativen) Informationen über die Person.

der Modellpassung mit dem pairwise-Algorithmus[27] mit 45 Freiheitsgraden bei 677 und ist damit extrem hoch.

Neben diesem Gesamtmodellgeltungstest mit Chi² wurden weitere Modellgeltungstests eingesetzt, so z. B. der Wald-Test. Beim Wald-Test zur Testung der Gesamtmodellgüte wird die empirische Stichprobe in zwei Untergruppen nach ambulantem und stationärem Setting unterteilt, um zu untersuchen, ob sich die Itemparameter in den beiden Personengruppen signifikant unterscheiden (Mair et al 2009, S. 11). Sollte dies der Fall sein, ist das ein Hinweis darauf, dass sich die Gruppen durch weitere Merkmale zu unterscheiden scheinen und Mobilität mit dieser Skala nicht Setting-unabhängig gemessen werden kann.

Es wurde verglichen, ob die Skala der beiden Gruppen „ambulante Versorgung“ und „stationäre Versorgung“ gleiche Itemvektoren über alle Personen hinweg produziert, also ob sich die Itemparameter im Ergebnis nicht unterscheiden. Es zeigt sich, dass die Items im ambulanten und stationären Setting nicht gleichermaßen funktionieren. Drei („Stabile Sitzposition halten“, „Fortbewegen innerhalb des Wohnbereichs“, „Treppensteigen“) der fünf Items unterscheiden sich abhängig vom Setting (Tab. 1.5):

Nr	Item	z-statistic	p-value
1	Positionswechsel im Bett	-0.833	0.405
2	Stabile Sitzposition halten	2.490	0.013
3	Aufstehen aus sitzender Position/Umsetzen	0.344	0.731
4	Fortbewegen innerhalb des Wohnbereichs	2.977	0.003
5	Treppensteigen	-2.371	0.018

Tab. 1.5 Wald-Test auf Itemlevel (z-Werte)

[27] Rasch hat eine Trennung der Schätzung von Personen- und Item-Parametern favorisiert. Für die Schätzung der Itemparameter kann die so genannte conditional Maximum Likelihood Methode (cML), eingesetzt werden, bei der die Wahrscheinlichkeit von Antwortpattern bei gegebenen Summenscores geschätzt wird. Diese ist dann alleine abhängig von den Item-Parametern, also letztlich den Item-Schwierigkeiten, die aus ihren Lösungswahrscheinlichkeiten heraus berechnet werden. Für die Schätzung der Personen-Parameter favorisiert Rost die so genannten Weighted (gewichteten) Likelihood Estimates, WLE abgekürzt (Rost, 2004, S. 314). Rost geht davon aus, dass es sich bei den WLE um die Schätzer der Personenparameter mit der geringsten Standardabweichung handelt (Rost, 2004, S. 315). Die WLE stellen das einzige Schätzverfahren für Personenparameter dar, das für Extremwerte, die also z.B. alle Kriterien maximal erfüllt haben oder aber alle Kriterien nicht erfüllt haben, überhaupt einen endlichen Personenparameter schätzt. Andere Verfahren, wie der so genannte „Pair-Wise“-Algorithmus (Rost, 2004, S. 311), schließen Probanden mit Extremwerten aus der Berechnung des Modells aus und schätzen die Modellpassung für alle Probanden, die keine Extremwerte aufweisen mit Hilfe der Item-Parameter. Begründet wird dieses Vorgehen damit, dass die Fähigkeit der Personen, die alle oder keines der Kriterien erfüllen, nicht mehr geschätzt werden kann, da der Test für die einen zu leicht und für die anderen zu schwierig ist. Die Unconditional Maximum Likelihood (UML)–Methode (Quelle) produziert bei der Schätzung der Personenparameter nur geringe Abweichungen von der WLE-Methode.

Keiner der Rasch-Modellgeltungstests ließe die getestete Mobilitätsskala ohne Änderungen passieren. Aufgrund der geringen Anzahl von Items kann auch keine alternative Mobilitätsskala aus der aktuellen extrahiert werden.

Damit ist der Aussage, dass sich in der Beurteilung eines Instruments klassisch testtheoretische Verfahren wie Cronbachs alpha und eine Item-Selektion nach dem Rasch-Modell nicht unterscheiden, wie sie von Erhart et al. (2009, S. 1) gemacht wird, eindeutig zu widersprechen. Beide Methoden unterscheiden sich im Einsatz wahrscheinlich nur dann nicht, wenn eine Skala bereits valide quantitative Daten liefert. Dies kann aber kein Hinweis darauf sein, dass beide Methoden einen identischen Beitrag zur Instrumentenentwicklung leisteten.

Die Abstände der ordinalen Skala zwischen „überwiegend selbständig“ (= 1) und „überwiegend unselbständig“ (= 2) sind bei den beiden schwierigsten Kriterien sehr viel kleiner (siehe Abb. 1.8) als bei den leichten Kriterien (siehe Abb. 1.7), so dass hier im Vergleich zu den leichten Kriterien keine gleichen Merkmalsabstände durch die Skala abgebildet werden. Bei der Weiterentwicklung sollte deshalb in einer nächsten Version besser mit einer dichotomen Antwortskala begonnen werden.

Die fünf Items sind eindeutig unterschiedlich schwierig. Vom leichtesten bis zum schwersten Item geordnet ergibt sich die folgende Rangreihe:

Nr	Item
2	Stabile Sitzposition halten
1	Positionswechsel im Bett
3	Aufstehen aus sitzender Position/Umsetzen
4	Fortbewegen innerhalb des Wohnbereichs
5	Treppensteigen

Tab. 1.6 Rangfolge der Items nach Schwierigkeit

Für die beiden leichtesten Items differenziert die eingesetzte Antwort-Skala hinreichend. Für die beiden schwersten Items, „Fortbewegen“ und „Treppensteigen“, unterscheiden die beiden Kategorien „überwiegend selbständig“ und „überwiegend unselbständig“ das Merkmal Mobilität allerdings nicht sehr deutlich. In den folgenden

Abbildungen sind die Wahrscheinlichkeiten der drei Schwellen der vierstufigen Likert-Skala für das Item „Positionswechsel im Bett“ (Abb. 1.7) und Treppensteigen (Abb. 1.8)

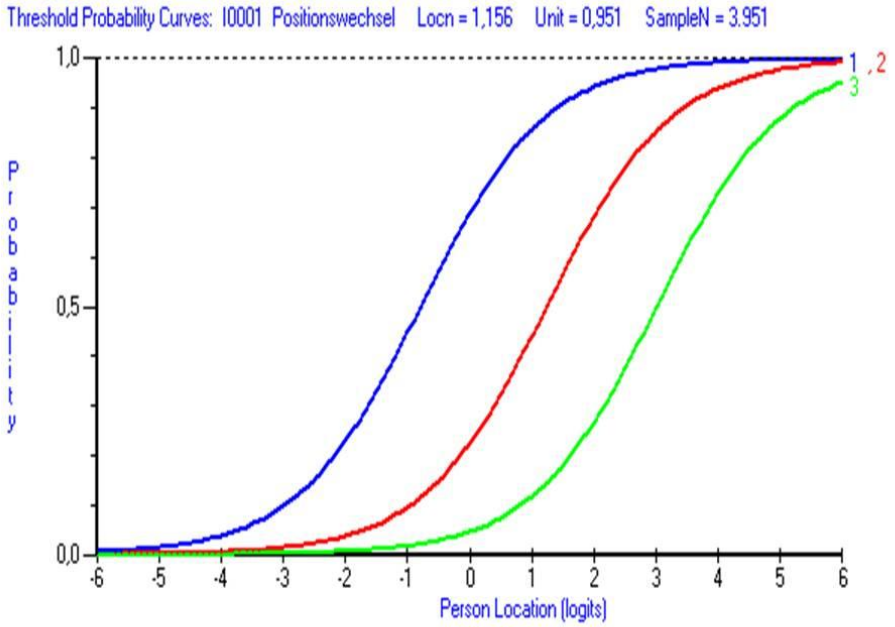


Abb. 1.7 Wahrscheinlichkeitsfunktion der Schwellenüberschreitung „Positionswechsel im Bett“

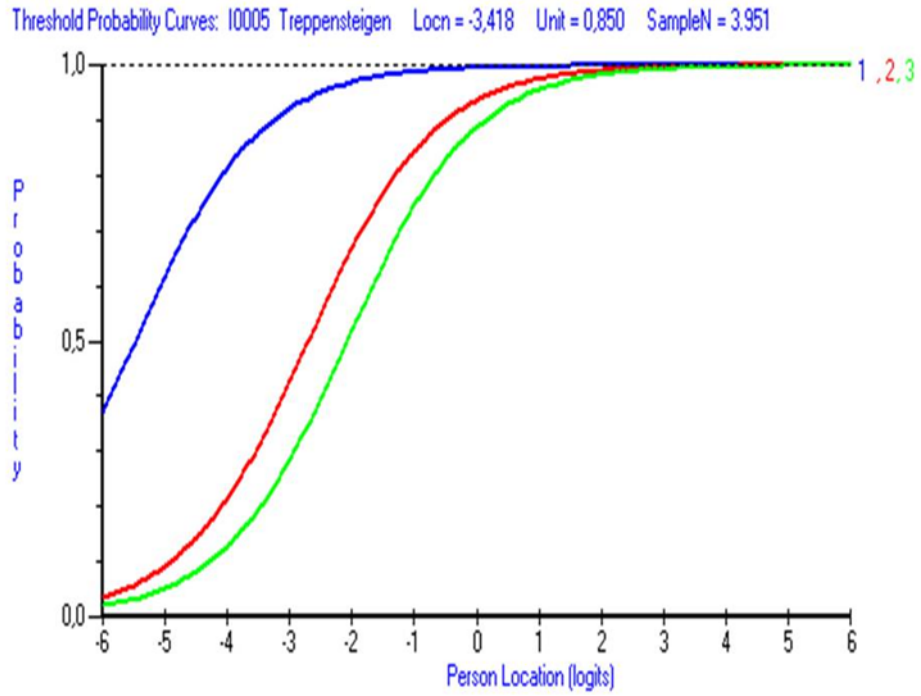


Abb. 1.8 Wahrscheinlichkeitsfunktion der Schwellenüberschreitung „Treppensteigen“

Die linke Linie symbolisiert die Schwelle, die überwunden werden muss, um nicht mehr „selbständig“ sondern „überwiegend selbständig“ anzukreuzen, die mittlere Linie symbolisiert die Schwelle, die von „überwiegend selbständig“ zu „überwiegend unselbständig“ überwunden werden muss und die rechte Linie symbolisiert die Schwelle, die von „überwiegend unselbständig“ zu „unselbständig“ überwunden werden muss. „Rechts von 6 werden alle drei Schwellen mit einer Wahrscheinlichkeit von nahe 1 überwunden, so dass „unselbständig“ die Ausprägung ist, die jemand mit der Unselbständigkeit von 6 zugeordnet bekommt und links von einer Unselbständigkeit von -6 wird keine der drei Schwellen mit einer Wahrscheinlichkeit von >0 überwunden, so dass hier „selbständig“ die Ausprägung ist, die jemand zugeordnet bekommt.

Für „Treppensteigen“ ergibt sich mit der ordinalen Skala demnach keine gute Merkmalsdifferenzierung (vgl. Abb. 1.8).

Abb. 1.9 demonstriert, wie die Item-Eigenschaften ausgedrückt in den drei Antwortschwellen ausgeprägt sein müssten, wenn eine valide Quantifizierung mit der Skala gelingen sollte. Tatsächlich sieht der Versuch der Quantifizierung des Merkmals „Mobilität“ mit der aktuell eingesetzten Antwortskala empirisch aus wie in Abb. 1.10 dargestellt.

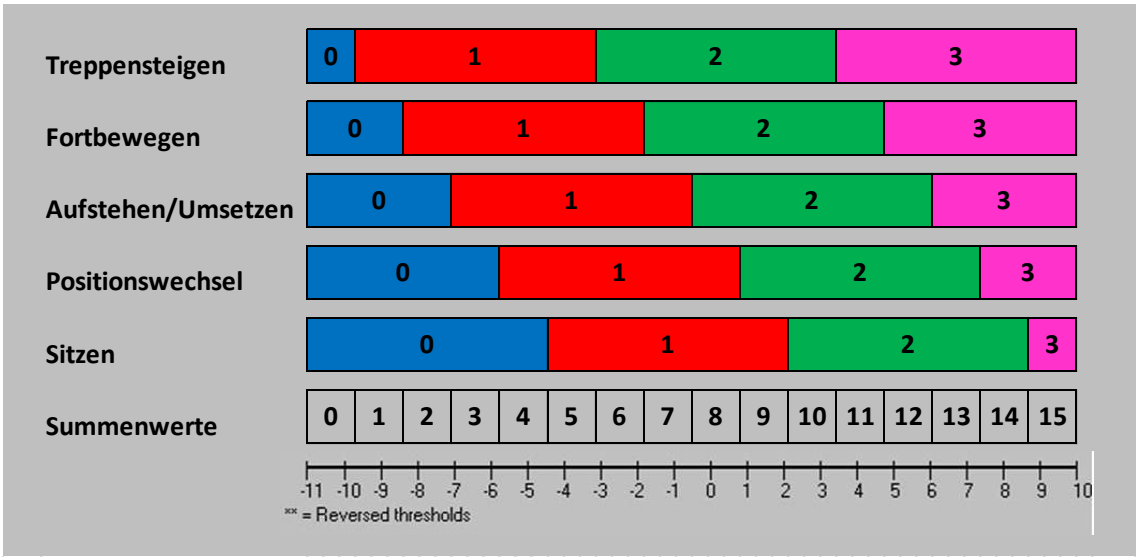


Abb. 1.9 Optimale Schwellenwerte für das Modul 1 „Mobilität“ des NBA

Abb. 1.10 zeigt die eindeutig unterschiedliche Schwierigkeit der fünf Items der Subskala „Mobilität“ in Bezug zur Personenfähigkeit und die großen Unterschiede in der Merkmalsdifferenzierung, die mit der gleichen Antwortskala zur Selbständigkeit bei allen fünf Items produziert wird.

Im Ergebnis kann nicht allen Summenwerten auch eine eindeutige Fähigkeits-ausprägung auf der intervallskalierten Logit-Skala von -11 (selbständig) bis zu 10 (unselbständig) zugeordnet werden.

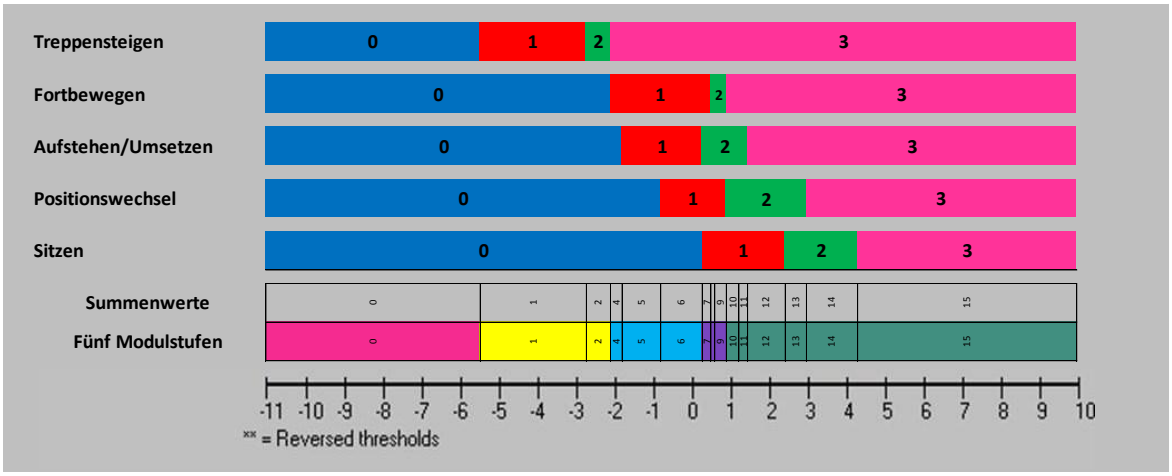


Abb. 1.10 Tatsächliche Schwellenwerte des Moduls 1 „Mobilität“ des NBA

Das aber wäre das Ziel einer Anpassung eines Rasch-Modells, das hier verfehlt wird.

Da im zweiten Quantifizierungsschritt den Summenwerten von 10-15 die gleiche Stufe (4) zugeordnet wird, werden in dieser höchsten Stufe der Mobilitätseinschränkung pflegerisch extrem unterschiedliche Personen zusammengefasst. Personen in Stufe 4 des Moduls „Mobilität“ können z. B. „überwiegend selbständig“, „überwiegend unselbständig“ oder „unselbständig“ sitzen. Personen in der Stufe 0 und 1 unterscheiden sich aber nur im Treppensteigen.

Für die Mobilitätssummenwerte kann die Hypothese aufgestellt werden, dass eindeutig nur diejenigen, die auf den eigenen Füßen stehen können von denjenigen unterschieden werden können, die dies nicht können.

Ein weiterer Hinweis auf die mangelnde Passung des Rasch-Modells sind die Anpassungs-Werte einzelner Items. Sie können z. B. in so genannten Fit-Residuen gemessen werden.

Die Fit-Residuen nach dem pairwise-Algorithmus der einzelnen Items sind eine Möglichkeit, um zu bewerten, welches Item in die Reihe von Items der Mobilitätsskala passt und welches Item nicht.

Sie können im Programm RUMM2030 berechnet werden. Sie sollten für ein Item nicht über 2 oder unter -2 liegen (Andrich 2004, S. 100f).

Nr	Item	Fit-Residuen	Signifikanz
1	Positionswechsel im Bett	0,22	.000
2	Stabile Sitzposition halten	-3,22	.000
3	Aufstehen aus sitzender Position/Umsetzen	-11,81	.000
4	Fortbewegen innerhalb des Wohnbereichs	-4,885	.000
5	Treppensteigen	-1	.000

Tab. 1.7 Fit-Residuen

Drei der fünf Items weisen Fit-Residuen auf (siehe Tab. 1.7), die unter -2 liegen. Alle Items haben unter der Annahme eines ordinalen Rasch-Modells eine Wahrscheinlichkeit, die gegen null geht.

Das nach dem „pairwise“-Algorithmus vorgehende Schätz-Verfahren in RUMM2030 ist hier besser geeignet, da es die extremen Antwortmuster ausschließt. Es weist für den vorliegenden Datensatz eine Modellpassung der empirischen Daten an ein Rasch-Modell eindeutig zurück. Der pairwise-Algorithmus ist im Vergleich zum Weighted Likelihood Estimate (WLE)-Algorithmus das strengere Verfahren, da er alle diejenigen Fälle aus der Berechnung ausschließt, die Extremwerte aufweisen. Das sind im vorliegenden Datensatz immerhin 630 Fälle von 5080 Fällen, die in allen Items vollkommen selbständig und 531 Fälle von 5080 Fällen, die in allen Items vollkommen unselbständig sind.

ZUSAMMENFASSUNG

Zusammenfassend kann festgehalten werden, dass Cronbachs alpha keine Hinweise für eine Veränderung der Mobilitätsskala liefert, die CFA Hinweise auf ein möglichweise nicht passendes eindimensionales Modell des Konstrukts „Mobilität“ liefert aber große Schätzprobleme auftreten und ein Rasch-Modell eindeutig nicht angepasst werden kann.

Das nicht passende Rasch-Modell verweist darauf, dass ein Summenwert nicht gebildet werden sollte.

Die ordinale Antwort-Skala führt nicht zu gleichen Merkmalsabständen zwischen den Antwortstufen und die Items der Skala werden unterschiedlich beantwortet je nachdem

in welchem Setting die Daten erhoben wurden (vgl. Bensch 2012, in diesem Band, S. 134ff).

Der Einsatz probabilistisch fundierter Methoden führt also zu anderen Ergebnissen als der Einsatz klassisch testtheoretisch fundierter Verfahren.

Das liegt daran, dass probabilistische Methoden die vorgenommenen Quantifizierungsschritte prüfen und klassisch testtheoretische Methoden eine valide Quantifizierung voraussetzen.

4. Welche Konsequenzen ergeben sich aus einer Erweiterung des Methodenspektrums für die Entwicklung von Erhebungsinstrumenten in der Pflege?

Die Ergebnisse der Analyse zeigen, dass

1. *keine validen intervallskalierten Daten vorliegen*

2. *Verfahren wie Cronbachs alpha und Faktorenanalysen nicht sinnvoll eingesetzt werden können, da sie bereits wenigstens ordinal-skalierbare Daten voraussetzen, die als intervallskalierte, d. h. äquidistante Daten behandelt werden dürfen.*

3. *probabilistische Verfahren in der Testung der Struktur einer Skala strenger sind als klassisch testtheoretisch fundierte, weil sie wechselseitige Beziehungen der Variablen höherer Ordnung abbilden können, wo hingegen faktorenanalytische Verfahren auf linearen Zusammenhängen zwischen Variablenpaaren basieren, was keine Prüfung einer gesamten Ordnung aller Items zueinander sicherstellt.*

4. *die Theorie eines Konstrukts und seine valide Messung in einem komplementären Zusammenhang zu verstehen sind und die Entwicklung valider Instrumente ein zirkuläres Verständnis von Instrumenten- und Theorieentwicklung voraussetzt*

5. *ein unpassendes Messmodell erst einmal identifiziert werden muss. Das gelingt mit Verfahren der klassischen Testtheorie nicht, wenn kategoriale Daten vorliegen, die vorschnell quantifiziert werden. Dann besteht keine Möglichkeit mehr, aus Instrumentenprüfungen etwas über die Angemessenheit einer Theorie zu erfahren.*

6. *die Testung von Reliabilität (im Beispiel mit Cronbachs alpha) kann positiv ausfallen, ohne dass dies etwas über die Qualität eines Instruments aussagte, wenn keine validen quantitativen Daten vorliegen.*

7. *Instrumente wesentlich mehr Zeit für Ihre Entwicklung brauchen, wenn empirische Ergebnisse dafür genutzt werden sollen.*

Prüfungen mit Verfahren aus der KTT, die valide quantitative Daten voraussetzen aber bei im Kern kategorialen Daten angewandt werden, sagen nicht viel über das Messmodell und die Struktur und damit über die Konstruktvalidität einer Skala aus.

Probabilistische Verfahren wie Rasch-Modelle helfen, die Quantifizierbarkeit durch Summenbildung zu prüfen und können in ihren dazu passenden Varianten auch die Qualität ordinaler Antwortskalen testen.

Für eine Quantifizierung des Konstrukts Mobilität sollten zusätzlich zu den Methoden der klassischen Testtheorie Methoden der probabilistischen Testtheorie Eingang in den Methodenkanon der Pflegewissenschaft finden, um bei kategorialen Daten bereits während der Instrumentenentwicklung zur Auswahl der Items eingesetzt zu werden und damit notwendige Änderungen des Instruments vornehmen zu können.

Sie helfen aber nur weiter, wenn sie während des Entwicklungsprozesses eingesetzt werden. Testet man eine bereits fertig gestellte Skala und stellt fest, dass die vorgenommen Quantifizierungen durch die Summierung von Items und die Skalierung von Antworten nicht valide sind, so muss die Weiterentwicklung des Instruments erfolgen, die dann weitere Methoden integrieren sollte. Hierauf gehen wir im Abschlusskapitel dieses Bandes ein.

Für eine Quantifizierung von Mobilität in der Form eines Summenwertes ist wahrscheinlich eine erneute inhaltliche Ausdifferenzierung und Operationalisierung des Konstrukts „Mobilität“ erforderlich, wie sie in dem kurzen Zeitraum der bisherigen Instrumentenentwicklung nicht vorgenommen werden konnte. Eine wichtige Frage ist die, wie wir setting-unabhängige Mobilitätsitems identifizieren können.

Zur Verbesserung der Operationalisierung von Theorien werden wir im Abschlusskapitel auf die Facettentheorie eingehen, von der wir uns präzisere Umsetzungen von Theorien in Instrumente versprechen. Außerdem brauchen wir empirisch gehaltvolle Erklärungsansätze für Pflegebedürftigkeit. Wir sehen hier eine

Beziehung zur Entwicklung von empirischen Erklärungsansätzen des tatsächlich produzierten Pflegeaufwands.

Wahrscheinlich müssen für eine Quantifizierung des Subkonstrukts Mobilität zusätzliche, neue Items in eine solche Skala integriert werden. Könnte mit einer größeren Anzahl von Items in eine Weiterentwicklung eingestiegen werden, so wäre eine Auswahl von Items, die in ein Rasch-Modell passen eher denkbar, als mit fünf Items eine „Punktlandung“ zu erreichen.

Ob die Summation von Items verschiedener Dimensionen (Kognition, Kommunikation, Mobilität, Hilfebedarf etc.) dann der richtige Weg ist, um Pflegebedürftigkeit zu bestimmen, ist aktuell eher fraglich.

Wahrscheinlich sind Kombinationen von dichotomen Items verschiedener Dimensionen nötig, wenn Pflegebedürftigkeit valide gemessen werden soll. Summierte Indices verschiedener Subdimensionen zu summieren ist deshalb nicht sinnvoll, weil einzelne Verbindungen einzelner und mehrere Items verschiedener Dimensionen relevant für die Ausprägung von Pflegebedürftigkeit sein können.

Um eine möglichst große Offenheit im Entwicklungsprozess zu produzieren, sind auch Struktur- und Messmodell entdeckende und prüfende Verfahren frühzeitig einzusetzen. Hierzu behandeln wir im Abschlusskapitel mögliche Alternativen bei der Entwicklung von Klassifikationsmodellen für Pflegeaufwand. Hierbei werden wir auf die von uns bereits eingesetzten Verfahren aus dem Bereich das so genannten „Data Mining“ (vgl. Hastie et al. 2009) eingehen, bei denen aus empirischen Datensätzen Klassifikationsmodelle entwickelt werden können. Solche Datensätze produzieren wir aktuell z. B. im Projekt „Pflegebedarf im Saarland“ (PiSaar) in Zusammenarbeit mit der Saarländischen Pflegegesellschaft (SPG). Diese werden eine deutlich niedrigere Anzahl von Items und einfachere Messmodelle benötigen, als sie das NBA aktuell für seine Bedarfsgrade vorschlägt, um Aufwandsgruppen zu unterscheiden. Wir ziehen die Konsequenz aus dem gescheiterten Versuch, Pflegebedürftigkeit zu quantifizieren und bemühen uns erst einmal darum, aktuelle (Zeit)Aufwände zu erklären.

Passt ein Modell nicht, muss erneut über die theoretische Differenzierung des Instruments nachgedacht werden. Notwendige Verbesserungen müssen Teil des Prozesses der Instrumentenentwicklung sein. Darüber hinaus ist das Ziel exakt zu benennen: soll ein Instrument entwickelt werden, um zu klassifizieren (Gruppen zu bilden) oder zu quantifizieren (einen Index zu bilden)? Diese klare Festlegung auf einen Zweck des Instruments erleichtert die Validierung des Instruments im Verlaufe seiner Entwicklung und trägt dazu bei, den Entwicklungsprozess methodisch zu strukturieren. Die Feststellung der Defizite eines bereits publizierten Instruments ist von

geringem Vorteil, wenn keine Möglichkeiten der Änderung und Verbesserung existieren. Deshalb plädieren wir dafür, Instrumente anhand von Datenerhebungen entwickeln zu können. Instrumente erst zu entwickeln, um sie mit Verfahren zu prüfen, die quantitative Daten und Merkmalskonstanz voraussetzen, und damit keine aussagekräftigen Ergebnisse liefern und sie dann gegenüber allen empirischen Prüfungen des Mess- und Strukturmodells zu verteidigen ist eine Strategie, die ein Lernen bei der Instrumentenentwicklung nicht vorsieht.

Wir schlagen vor, dass Datenerhebungen zukünftig dazu genutzt werden dürfen, Struktur- und Messmodelle und die Aufgabenerfüllung von Instrumenten zu prüfen, um daraus für die Weiterentwicklung von Instrumenten zu lernen.

Für das NBA bedeutet das, die Aufgabe zu vereinfachen und erst einmal realen Pflegeaufwand zu erklären und das Mess- und Strukturmodell zu vereinfachen.

LITERATUR

Andrich, David (2004): An Introduction to Rasch Models for Measurement and traditional Test Theory. Instrument Design with Rasch IRT and Data Analysis I. EDU435/635. Herausgegeben von Murdoch University. Pearth, Western Australia.

Becker, Clemens; Blinkert, Baldo; Dietz, Berthold; Döhner, Hanneli; Frommelt, Mona; Klie, Thomas; Kruse, Andreas; Rothgang, Heinz (2007): Memorandum. Die Quadratur des Kreises in der Begutachtung der Pflegebedürftigkeit – Forschung statt Politik – Instrument vor Verfahren. http://www.bapp.info/texte/Memorandum-Pflegebeduerftigkeit.pdf zuletzt eingesehen am 02.07.2012

Bensch, Sandra (2012): Konstruktvalidität der Module „Mobilität und „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessments zur Feststellung von Pflegebedürftigkeit. Dissertation an der Philosophisch-Theologischen Hochschule Vallendar, Pflegewissenschaftliche Fakultät

Brühl, Albert; Berger, Bianca (2011): Mit weniger Kriterien besser differenzieren: Warum bei der Messung von Qualität weniger mehr sein kann. Selektion von 15 validen bewohnerbezogenen Kriterien der Pflegetransparenzvereinbarung nach §115 SGB XI. Pflegewissenschaft 10/2011. 525-534

Bühner, Markus (2006): Einführung in die Test- und Fragebogenkonstruktion. 2., aktualisierte Auflage. München: Pearson Studium.

Carstensen, Claus H. (2000) Mehrdimensionale Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik, IPN 171, Universität Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

Erhart M, Hagquist C, Auquier P, Rajmil L, Power M, Ravens-Sieberer U and the European KIDSCREEN Group (2009): A comparison of Rasch item-fit and Cronbachs alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents. Blackwell Publishing Ltd, Child: care, health and development, 36, 4, 473–484

Franken, Georg (2010): Konstruktvalidität der Subskala „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit (NBA). Masterarbeit. Vallendar: Philosophisch Theologische Hochschule. http://opus.bsz-bw.de/kidoks/volltexte/2012/66/ zuletzt geprüft am 20.08.2012

Hagquist, Curt; Bruce, Malin; Gustavsson, J. Petter (2009): Using the Rasch model in nursing research: An introduction and illustrative example. International Journal of Nursing Studies, 46 (2009) 380-393

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009): The Elements of Statistical Learning. Springer

Huber, Evelyn (2008): OLA: Optimierung der Lebensqualität im Alter: Entwicklung eines Fragebogens zur Zufriedenheit der Angehörigen von Bewohnerinnen in Altersinstitutionen. Pflege 21 (5) 319-326

Jöreskog, Karl G. (2002): Structural Equation Modeling with Ordinal Variables using LISREL. Online verfügbar unter http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf, zuletzt aktualisiert am 10.02.2005, zuletzt geprüft am 03.08.2011.

Jöreskog, Karl G.; Sörbom, Dag (2003): PRELIS 2. User's Reference Guide. A program for multivariate data screening and data summarization; a preprocessor for LISREL. 3rd ed., updated to PRELIS 2. Lincolnwood, IL: Scientific Software International, Inc.

Krohwinkel, Monika (2007): Rehabilitierende Prozesspflege am Beispiel von Apoplexiekranken. Fördernde Prozesspflege als System. 2., überarbeitete und erweiterte Auflage. Bern: Huber

Mair, Patrick; Hatzinger, Reinhold (2009): Extended Rasch Modeling: The R

Mai, Markus (2010): Das Sturzrisiko von Patienten im Krankenhaus. Entwicklung eines konstruktvaliden Sturzrisikoeinschätzungsinstruments unter dem Einsatz von Modellen aus dem Bereich der probabilistischen Testtheorie. München: Dr. Hut-Verlag

Moosbrugger, Helfried; Kelava, Augustin (2007): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer

Moosbrugger, Helfried; Schermelleh-Engel, Karin (2008): Exploratorische (EFA) und Konfirmatorische Faktorenanalyse (CFA). In: Moosbrugger, Helfried; Kelava, Augustin (Hg.): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer Medizin, S. 307–324.

Müller-Staub, Maria; Lumney, Margaret; Lavin, Mary Ann; Needham, Ian; Odenbreit, Matthias; van Achterberg, Theo (2010): Testtheoretische Gütekriterien des Q-DIO, eines Instruments zur Messung der Qualität der Dokumentation von Pflegediagnosen, -interventionen und –ergebnissen. Pflege 23 (2) 119-128

Package eRm. PDF - Dateianhang zum Programmpaket eRm.

Panfil, Eva-Maria (2004): Fokus: Klinische Pflegeforschung. Hannover: Schlütersche

Pospeschill, Markus (2010): Testtheorie, Testkonstruktion, Testevaluation. München: Ernst Reinhard Verlag

Rost, Jürgen; Davier, Matthias von (1994): A Conditional Item-Fit Index for Rasch Models. Applied Psychological Measurement 18 (2) 171-182

Rost, Jürgen (1999): Was ist aus dem Rasch-Modell geworden? Psychologische Rundschau 50 (3), 140-156

Rost, Jürgen (2004): Testtheorie und Testkonstruktion. 2. Auflage, Bern: Huber

Schermelleh-Engel, Karin; Moosbrugger, Helfried; Müller, Hans (2003): Evaluating the Fit of Structural Equation Models: Test of Significance and Descriptive Goodness-of-Fit Measures. Methods of Psychological Research Online 8 (2) 23-74 http://mpr-online.de

Schröder, Martina (2010): Konstruktvalidität der Subskala Mobilität des Neuen Begutachtungsassessments für Pflegebedürftigkeit (NBA). Masterarbeit. Betreut von Prof. Dr. Albert Brühl. Vallendar. Philosophisch-Theologische Hochschule Vallendar. Online verfügbar unter http://www.dip.de/datenbankwise/detail/?no_cache=1&tx_dipwise_pi2[uid]=499 zuletzt geprüft am 09.08.2010.

Strobl, Carolin (2010): Das Rasch-Modell. München: Rainer Hampp

Walter, Oliver (2005): Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten. Lengerich: Pabst Science Publishers

Wingenfeld, Klaus; Büscher, Andreas; Gansweid, Barbara (2008): Das neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Projekt: Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Abschlussbericht zur Hauptphase 1: Entwicklung eines neuen Begutachtungsinstruments. Studie im Rahmen des Modellprogramms nach § 8 Abs. 3 SGB XI im Auftrag der Spitzenverbände der Pflegekassen. Bielefeld: IPW

Wingenfeld, Klaus; Engels, Dietrich (2011): Entwicklung und Erprobung von Instrumenten zur Beurteilung der Ergebnisqualität in der stationären Altenhilfe. Abschlussbericht vom 31. Januar 2011 Bielefeld/Köln: Institut für Pflegewissenschaft an der Universität Bielefeld (IPW); Institut für Sozialforschung und Gesellschaftspolitik GmbH (ISG)

2. Das implizite Struktur- und Messmodell des Neuen Begutachtungsassessments (NBA)

Katarina Planer, Albert Brühl

Im Jahr 2005 initiierte das Bundesministerium für Gesundheit eine grundlegende Überarbeitung des Pflegebedürftigkeitsbegriffs. Hierzu wurde im Oktober 2006 ein Beirat einberufen, dessen Aufgabe es ist, den Entwicklungs- und möglicherweise auch Umsetzungsprozess zu begleiten (BMG 2009). Parallel etablierte das Ministerium einen Steuerungskreis zur unmittelbaren Betreuung des Projekts. Die Spitzenverbände der Pflegekassen beauftragten im Rahmen des Modellprogramms nach § 8 (3) SGB XI das Institut für Pflegewissenschaft an der Universität Bielefeld (IPW) eine Recherche und Analyse von Pflegebedürftigkeitsbegriffen und Einschätzungsinstrumenten durchzuführen (Wingenfeld et al. 2007). Die Auftraggeber folgten der Empfehlung des IPW, eine Neuentwicklung eines Begutachtungsinstruments in Anbetracht des Zeitplans anzugehen. Die in einem Memorandum geäußerte Kritik an der Projektausschreibung (Becker et al. 2007) fand keine Berücksichtigung. Vor diesem Hintergrund entwickelte das IPW in Kooperation mit dem Medizinischen Dienst der Krankenversicherung Westfalen-Lippe (MDK-WL) im Rahmen des Projekts „Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI“ in einer ersten Hauptphase modellhaft das „Neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit“ (NBA) (Wingenfeld et al. 2008). In einer zweiten Hauptphase des Projekts wurden vom Institut für Public Health und Pflegeforschung der Universität Bremen (IPP Bremen) in Kooperation mit dem Medizinischen Dienst des Spitzenverbandes Bund der Krankenkassen e.V. (MDS) die wissenschaftliche Beurteilung der Güte (Windeler et al. 2008) sowie die Abschätzung der inhaltlichen und finanziellen Folgen (Rothgang et al. 2008 und Windeler et al. 2008) vorgenommen.

Analyse

Die in diesem Kapitel vorgenommene Analyse des Struktur- und Messmodells vor dem Hintergrund der Definition von Pflegebedürftigkeit (Wingenfeld et al. 2008, S. 28) beschränkt sich auf das standardisierte Verfahren des NBA zur Feststellung von

Pflegebedürftigkeit von Erwachsenen (Module 1 - 6), das derzeit vorliegt. Sie wird aus einer Perspektive der Instrumentenentwicklung vorgenommen, die eine pflege(wissenschaftliche) Perspektive mit messtheoretischen Aspekten zu verbinden sucht. Sie berücksichtigt nicht die Rahmenbedingungen und die Entstehungsgeschichte des Instruments sowie mögliche politische Implikationen, sondern versteht sich als Beitrag einer vertieften, konstruktiv kritischen Auseinandersetzung mit den Bedingungen einer validen Messung komplexer, pflegerelevanter Phänomene. Darüber hinaus seien interessierten Lesern die umfassenden Berichte der einzelnen Projektphasen empfohlen[28].

Die Analyse des NBA folgt der Struktur des heuristischen Rahmens (Brühl 2012, in diesem Band, S. 14f) die dazu dient, ein Instrument im Kontext seines Anwendungszwecks zu definieren.

Aufgabe

Entsprechend der Projektausschreibung hat das IPW die „Entwicklung eines neuen, modularen, praktikablen, standardisierten […] Begutachtungsinstruments mit Pretest unter Berücksichtigung der gleichzeitigen Erarbeitung oder Zugrundelegung eines vom Gesetzgeber noch nicht entschiedenen alternativen Pflegebedürftigkeitsbegriffs“ vorgenommen (Wingenfeld et al. 2008, S. 5). Dieser Auftrag beinhaltet sowohl die Entwicklung einer Theorie zur Pflegebedürftigkeit (Strukturmodell) als auch die Entwicklung einer Bewertungssystematik (Messmodell) zur Einschätzung des Gesamtausmaßes der Abhängigkeit von personeller Hilfe, das in einem Zeitrahmen von 60 Minuten standardisiert anwendbar ist.

Daher wurden mit der Entwicklung des NBA unterschiedliche Anforderungen verknüpft. Das Instrument soll dazu dienen

- Pflegebedürftigkeit zu erfassen
- Hilfebedürftigkeit zu erheben
- durch Erfassung von Risiken den Präventionsbedarf zu ermitteln
- Rehabilitationsbedürftigkeit und Rehabilitationsfähigkeit einzuschätzen
- Hilfsmittelversorgung zu erfassen
- besondere Problem- und Bedarfskonstellationen zu ermitteln (Bedarfsstufe 5)
- Hinweise für den individuellen Pflegeplan zu liefern

(Wingenfeld et al. 2008, S. 20f)

[28] Bundesministerium für Gesundheit (2009); Windeler et al (2008); Wingenfeld et al (2007 und 2008); Rothgang et al (2008)

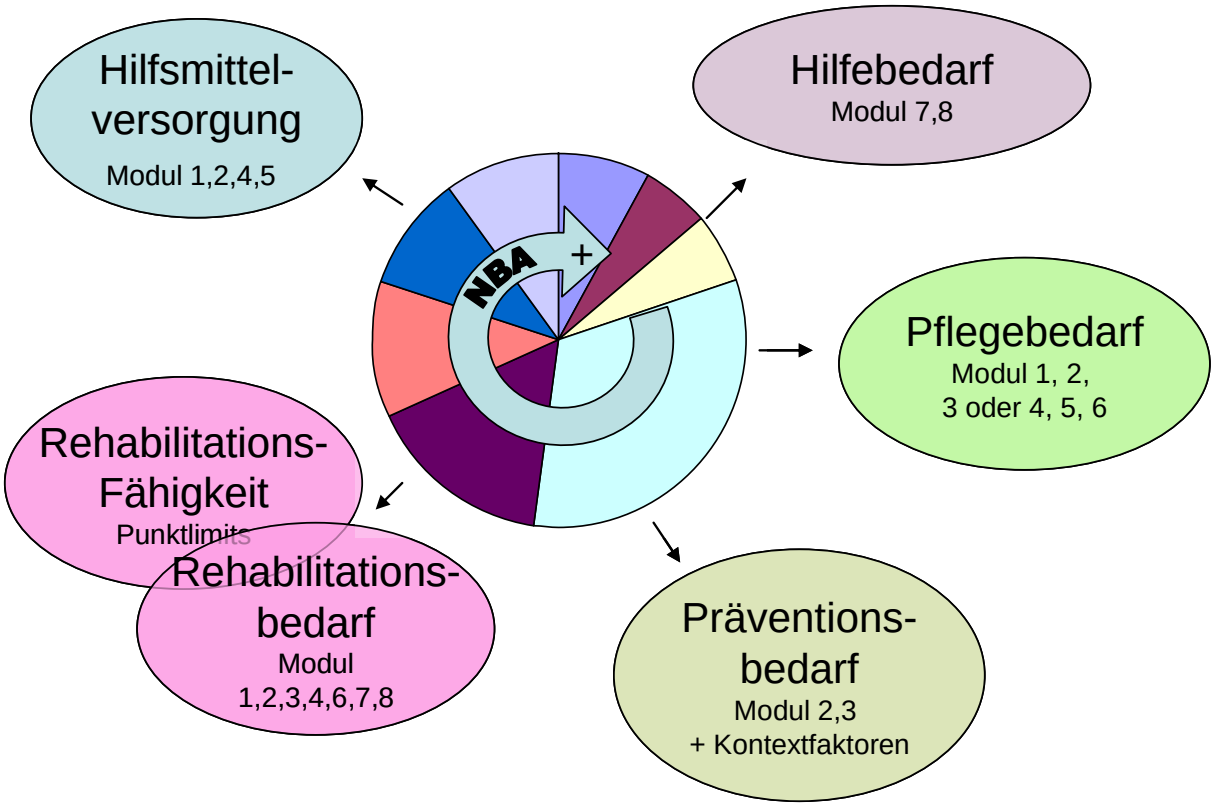


Abb. 2.1 Funktionen des NBA

Die einzelnen Module begründen also nicht nur das Konstrukt der Pflegebedürftigkeit sondern sollen auch die (nicht explizierte) strukturtheoretische Grundlage für weitere Anwendungen bilden. Die Elemente des NBA, bzw. einzelne Module werden als „kleinster gemeinsamer Nenner“ für die Ermittlung weiterer Bedarfe angesehen. Die Module 1 – 6 des Begutachtungsassessments sollen vorrangig der Ermittlung von unterschiedlichen Leistungsansprüchen dienen (Wingenfeld et al. 2008, S. 33). Zusätzlich sollen die Ergebnisse des neuen Begutachtungsverfahrens im Rahmen der individuellen Hilfe- oder Pflegeplanung nutzbringend sein.“ (Wingenfeld et al. ebd.)

Inhalt

Das Neue Begutachtungsassessment legt folgende Definition von Pflegebedürftigkeit zu Grunde:

„[…] Eine Person ist als pflegebedürftig zu bezeichnen, wenn sie

- infolge fehlender personaler Ressourcen, mit denen körperliche oder psychische Schädigungen, die Beeinträchtigung körperlicher oder kognitiver/psychischer Funktionen, gesundheitlich bedingte Belastungen oder Anforderungen kompensiert oder bewältigt werden könnten,
- dauerhaft oder vorübergehend

- zu selbständigen Aktivitäten im Lebensalltag, selbständiger Krankheitsbewältigung oder selbständiger Gestaltung von Lebensbereichen und sozialer Teilhabe
- nicht in der Lage und daher auf personelle Hilfe angewiesen ist."

(Wingenfeld et al. 2007, S. 43; 2008, S. 28)

Diese Definition als theoretisches Inhaltsmodell von Pflegebedürftigkeit listet Aspekte und Elemente auf, die als leistungsrelevant für das Phänomen der Pflegebedürftigkeit erachtet werden.

Die Instrumente EASY Care, FACE (Functional Assessment of the Care Environment for Older People), RAI HC 2.0 (Resident Assessment Instrument Home Care 2.0) und das "alternative Begutachtungsverfahren" der MDK-Gemeinschaft wurden als Referenzen herangezogen (Wingenfeld et al. 2008, S. 9).

Für die o.g. zahlreichen Aufgaben wurde das NBA in einer Weise strukturiert, die es ermöglicht, den jeweiligen Anforderungen mithilfe der Daten der einzelnen Teile und Module des Instruments gerecht zu werden. Der Entwurf eines entsprechenden Begutachtungsformulars ist wie folgt strukturiert (Wingenfeld et al. 2008, A-1):

A. Neues Begutachtungsformular

1. Angaben zur Person und Begutachtungssituation
2. Anamnese
3. Wohn- und Lebenssituation
4. Versorgungssituation
5. Befunderhebung zu Schädigungen und Beeinträchtigungen

B. Erhebungsbogen des neuen Begutachtungsassessment zur Bestimmung der Pflegebedürftigkeit (nur Modul 1-6)

1. Mobilität
2. Kognitive und kommunikative Fähigkeiten
3. Verhaltensweisen und psychische Problemlagen
4. Selbstversorgung
5. Umgang mit krankheits-/therapiebedingten Anforderungen und Belastungen
6. Gestaltung des Alltagslebens und soziale Kontakte
7. Außerhäusliche Aktivitäten (für Hilfebedarf)
8. Haushaltsführung (für Hilfebedarf)
9. Präventionsbedarf

C. Ergebnisse und Empfehlungen

Eine Interpretation des Inhaltsmodells des NBA wird in Abb. 2.2 grafisch dargestellt. Entsprechend der Definition (Wingenfeld et al. 2008, S. 28) werden „körperliche oder psychische Schädigungen“, die „Beeinträchtigung körperlicher oder kognitiver/psychischer Funktionen“ und „gesundheitlich bedingte Belastungen oder Anforderungen“ als Bedingungen von Pflegebedürftigkeit in der Grafik auf der linken Seite abgebildet. Die Definition differenzierend werden „körperliche oder psychische Schädigungen“ und „gesundheitlich bedingte Belastungen oder Anforderungen als Ursachen für „Beeinträchtigungen körperlicher oder kognitiver/psychischer Funktionen“ dargestellt. Hier folgt das Inhaltsmodell dem Anspruch des NBA, die Dimension der „Selbständigkeit“ zu erfassen.

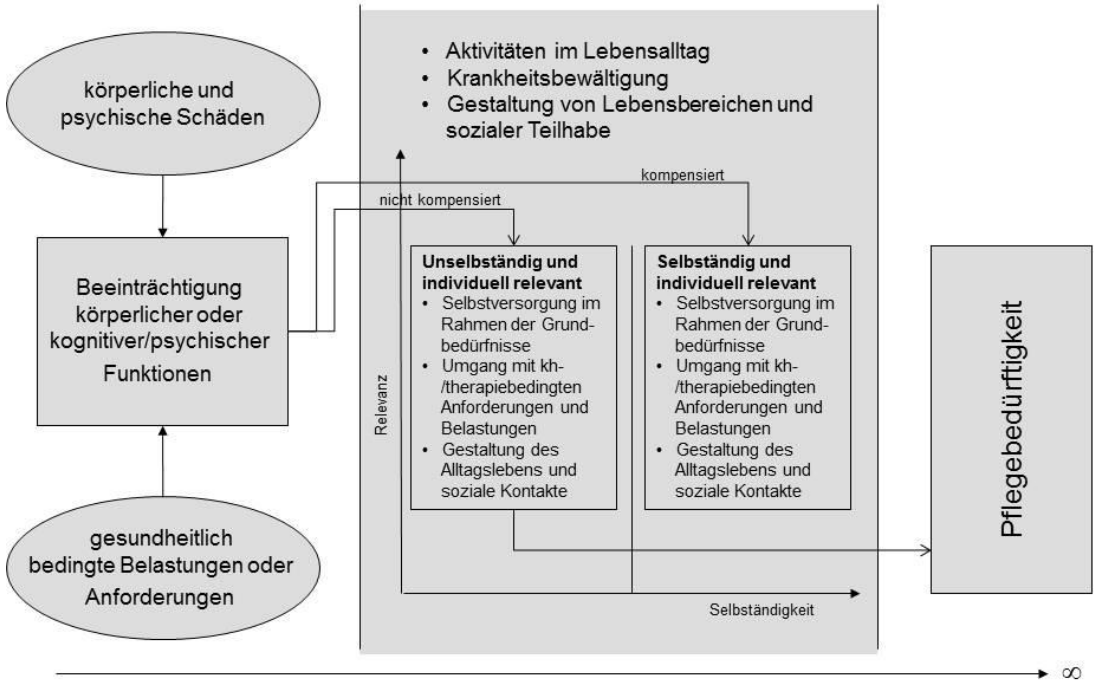


Abb. 2.2 Interpretation des Inhaltsmodells des NBA

Ist eine Person aufgrund ihrer personalen Ressourcen in der Lage, „Beeinträchtigungen körperlicher oder kognitiver/psychischer Funktionen zu kompensieren, gilt sie als selbständig in den „Aktivitäten im Lebensalltag“, der „Krankheitsbewältigung“ und der „Gestaltung von Lebensbereichen und sozialer Teilhabe“ und ist damit nicht auf personelle Hilfe angewiesen, selbst wenn „körperliche und psychische Schädigungen“ oder „gesundheitlich bedingte Belastungen oder Anforderungen“ vorliegen sollten (deren Auswirkungen kompensiert werden).

„Das Modul 1 „Mobilität“ umfasst die Fähigkeit zur Fortbewegung sowie zur Lageveränderung des Körpers.“ (Wingenfeld et al. 2008, S. 35). Und wird sowohl als Aktivität als auch als Funktion verstanden.

Die Auswahl der fünf Items dieses Moduls, die mit der vierstufigen Selbständigkeitsskala erfasst werden, wird inhaltlich und nicht empirisch begründet.

Weil Selbständigkeitseinbußen, die auf die Beeinträchtigung „kognitiver und kommunikativer Fähigkeiten“ (Modul 2) sowie „Verhaltensweisen und psychische Problemlagen“ (Modul 3) zurückzuführen sind, einen umfangreichen Unterstützungsbedarf im Sinne allgemeiner Betreuung und Beaufsichtigung nach sich ziehen können, wurden die Module 2 und 3 ins NBA aufgenommen (Wingenfeld et al. 2008, S. 20).

STRUKTURMODELL

Das Strukturmodell beschreibt die Relationen und Wechselwirkungen der Inhalte des Instruments zueinander.

Für das Strukturmodell ist nicht nur die inhaltliche Beziehung der Elemente des Konstrukts von Interesse sondern darüber hinaus interessieren die Annahmen über die konkreten Relationen, die sich z.B. in der Skalierung oder Kategorisierung der einzelnen Items, bzw. ihrer Gewichtung auf Modulebene, bzw. im Rahmen des Gesamtindex widerspiegeln.

Die Definition von Pflegebedürftigkeit als Abhängigkeit von personeller Hilfe beinhaltet das Konzept der Selbständigkeit/Unselbständigkeit, das im Rahmen des Assessments in den zumeist vierstufigen Antwortkategorien der Skala quantifiziert wurde. In seinem Endergebnis macht das NBA eine „Aussage über Art und Ausmaß der Beeinträchtigungen von Selbständigkeit bei der Durchführung von Aktivitäten und der Gestaltung von Lebensbereichen [in dem der Gutachter […] den Grad der Selbständigkeit der betreffenden Person ermittelt. (Wingenfeld et al. 2008, S. 28)].“ (Wingenfeld et al. 2008, S. 23)

Neben einer quantitativen Größe soll das Endergebnis auch qualitative Aussagen ermöglichen, wie „u.a. [ist] eine explizite Einschätzung des Bedarfs an allgemeiner Betreuung und Beaufsichtigung [vorgesehen]“. (Wingenfeld et al. 2008, S. 20)

Die Grundannahme, dass sich Pflegebedürftigkeit über den Grad der Selbständigkeit ermitteln lässt, wird nicht inhaltstheoretisch begründet, sondern als Abkehr von der

kritisierten Zeitberechnung erforderlicher Pflegeleistungen und Orientierung an „renommierten internationalen Vorbildern“ argumentiert (Wingenfeld et al. 2008, S. 20).

Bestätigt wird dies durch einen Blick auf das Messmodell: Die Kodierung der Selbständigkeits-Skala zeigt, dass der Kategorie „selbständig“ 0 Punkte zugeordnet werden und über eine vierstufige Skala die Kategorie „unselbständig“ mit 3 Punkten codiert wird (Wingenfeld et al. 2008, S. 29). Da im Endergebnis eine hohe Punktzahl einen hohen Pflegebedarf abbilden soll wird somit nicht Selbständigkeit sondern Unselbständigkeit gemessen, was dem Phänomen Pflegebedürftigkeit und der NBA-Definition (Wingenfeld et al 2008, S. 29) entspricht.

Abb. 2.3 Konzept der Selbständigkeitsskala des NBA

Dieser inhaltstheoretische Ansatz, Selbständigkeit als Ressource wahrzunehmen und zu messen ist im Kontext eines Leistungskonzepts des SGB XI das vorrangig präventive Leistungen fördert, sehr sinnvoll und schlüssig. Aufgrund der Definition wird Pflegebedürftigkeit aber als „Angewiesensein auf personelle Hilfe“ in Folge des Fehlens personaler Ressourcen zur Kompensation oder Bewältigung von Defiziten (vgl. Wingenfeld et al. 2008, S. 28) als Kompensations- und nicht als Präventionsmodell verstanden (ebd. S. 33, 92). Damit führt ein ressourcenorientierter Messansatz struktur- und messtheoretisch ohne weitere Definition oder Referenzwert von „Unselbständigkeit“[29] zu unbewältigten Problemen. Darüber hinaus wird dieser Herausforderung mit zwei wesentlichen Abweichungen vom theoretischen Konzept auf der Ebene der Kriterien begegnet. Die NBA-Skalen zeigen, dass die Entwickler eher davon auszugehen scheinen, dass Pflegebedürftigkeit von der Handlungs(un)fähigkeit des Betreffenden (Selbständigkeits-Skala), von der Hilfeleistung eines Dritten und von speziellen Personenmerkmalen des Antragstellers abhängig sei (vgl. Franken 2012, in diesem Band, S. 79). Damit wird der eigene Anspruch, den „Grad der Selbständigkeit

[29] Wenn Pflegebedürftigkeit mit dem Maß des Angewiesenseins auf personelle Hilfe definiert wird, das Instrument aber Selbständigkeit messen soll (obwohl die Skala umgekehrt codiert ist), ist das Ausmaß der Pflegebedürftigkeit von der Differenz zwischen dem Grad der Selbständigkeit und einer absoluten Unselbständigkeit abhängig (siehe Abb. 2.3). Die Schwierigkeit bei der Definition von Unselbständigkeit liegt in der Problematik der Güte und Intensität von (teilweise) selbständig durchgeführtem Handeln. Im Unterschied zur Selbständigkeit über die der Pflegebedürftige meist selbst und eigenständig entscheiden kann (weil er dann nicht pflegebedürftig ist), nimmt der Anteil der Fremdbewertung durch Dritte bei zunehmender Unselbständigkeit zu. Damit ist die Wahrscheinlichkeit hoch, dass der Referenzwert der absoluten Unselbständigkeit maßgeblich durch die Pflegeperson/Gutachter bestimmt wird.

bei der Durchführung von Aktivitäten oder der Gestaltung von Lebensbereichen“ (Wingenfeld et al. 2008, S. 20) als Maßstab zu definieren, nicht konsequent eingehalten.

Das Konzept der Selbständigkeit wird nicht in allen Modulen angewandt. Die Items des Moduls 2 „kognitive und kommunikative Fähigkeiten“ werden anhand einer ebenfalls vierstufigen Skala zum Ausmaß der erfassten Fähigkeit in der Differenzierung „vorhanden/unbeeinträchtigt“ – „größtenteils vorhanden“ – „in geringem Maße vorhanden“ – „nicht vorhanden“ erfasst. Die Items des Moduls 3 „Verhaltensweisen und psychische Problemlagen“ werden entsprechend ihrer Auftretenshäufigkeit erfasst: „nie“ – „selten“ (ein- bis zweimal innerhalb von zwei Wochen) – „häufig“ (zweimal oder mehrmals wöchentlich, aber nicht täglich) – „täglich“.

Die sich aus der fehlenden Spezifikation des Strukturmodells nach sich ziehenden Probleme für das Messmodell beschreibt Franken (Franken 2012 in diesem Band, S. 82f).

MESSMODELL

Messtheoretische Grundlagen

Die Ausprägung von Pflegebedürftigkeit lässt sich nicht direkt beobachten und ist mutmaßlich von vielfältigen Faktoren und Kombinationen dieser Faktoren abhängig. Die Definition von Konstrukten geschieht in einem gesellschaftlichen Kontext und wird meist dann notwendig, wenn viele Menschen in Bezug auf einen Eigenschaftskomplex zu einem bestimmten Zweck miteinander verglichen werden sollen.

Entsprechend seiner Definition soll das NBA Menschen in ihrer Pflegebedürftigkeit aufgrund ihrer Selbständigkeit/ Unselbständigkeit unterscheiden, um daraus den individuellen Leistungsanspruch gegenüber den Sozialversicherungen ermitteln zu können.

Die Vergleichbarkeit der Personen geschieht meist (so auch beim NBA) mittels eines Index, der sich aus der regelrechten Anwendung des Instruments als dessen Ergebnis errechnen lässt. Ein Index, bzw. ein Konstrukt gilt dann als valide, wenn die Berechnungssystematik (Messmodell) geeignet ist, die Menschen in Bezug auf den interessierenden Eigenschaftskomplex in ihrer Verschiedenartigkeit richtig zu unterscheiden.

Soll das Instrument die Pflegebedürftigkeit bei sehr vielen Menschen (beim NBA alle Erwachsenen, bzw. Kinder) valide unterscheiden, muss die Theorie (Beziehungen der inhaltlichen Elemente zueinander) über das Konstrukt so allgemeingültig sein, dass die Aspekte für den potentiellen Personenkreis relevant und geeignet sind, sie in Bezug auf das Konstrukt (Pflegebedürftigkeit) zu differenzieren.

Ist hingegen die inhaltliche Struktur in zu hohem Maße differenziert besteht das Risiko, dass diese inhaltlichen Spezifikationen die Verallgemeinerbarkeit begrenzen.

Das Messen latenter Konstrukte setzt zunächst die Operationalisierung des inhaltlichen Strukturmodells in ein Messinstrument (Test) voraus. Durch das Festlegen von Messanweisungen oder Regeln als Abbildungsfunktion der theoretischen Relationen in numerische Relationen entsteht ein Messmodell.

In der Praxis bedeutet dies, dass den inhaltlichen Aspekten in Form von Fragen oder Aufgaben (Items) ein Beantwortungsschema (Skala) zugeordnet wird, dessen Antwortmöglichkeiten (Objekte als Kategorien, Ereignisse, Merkmale, Häufigkeiten) Zahlen zugeordnet werden.

Die zugeordneten Zahlen dienen als Stellvertreter der Objekte, um mathematische Operationen überhaupt erst zu ermöglichen. Diese Zuordnung wird in Form einer Skala vorgenommen, die die Zuordnungsregel von Maßzahlen zu den Objekten darstellt. Die verwendete Skala stellt also einen Teil der durch das Instrument operationalisierten Theorie des Konstrukts dar.

Da es sich im Grunde genommen um eine willkürliche Zuordnung von Zahlen zu den Objekten handelt, die auf ersten theoriegeleiteten Hypothesen beruhen, kann nicht mit Sicherheit davon ausgegangen werden, dass die numerischen Relationen des Messmodells die empirischen Relationen bzw. die theoretischen Annahmen des inhaltlich-theoretischen Modells strukturerhaltend abbilden.

Bei einer Validierung eines Instruments gilt es daher nicht nur die Validität der Items zu prüfen, sondern auch die Validität der Skala und des (Gesamt-) Indexes in Frage zu stellen (Brühl 2012, in diesem Band, S. 34).

Um zu überprüfen, ob die Zuordnungs- und Verrechnungsregeln für den Index geeignet sind, die empirischen Beziehungen der inhaltlichen Aspekte zueinander mithilfe der numerischen Relationen des Messmodells abzubilden, werden statistische Methoden benötigt, deren Grundlage die Testtheorie bildet. Im Vorfeld ist es die

Aufgabe der (klassischen) Messtheorie die logisch-mathematischen Voraussetzungen dieser Zuordnungen und die Spezifizierung von Zuordnung unter Berücksichtigung des Repräsentationsproblems, des Eindeutigkeitsproblems und des Bedeutsamkeitsproblems zu erklären (vgl. Bortz et al. 2006, S. 65).

Repräsentationsproblem

Das Repräsentationsproblem bezieht sich auf die Frage, ob die dem Instrument zugrunde gelegte Skala tatsächlich die behaupteten Bedingungen erfüllt. Die Selbständigkeits-Skala mit der die Module „Mobilität", „Selbstversorgung" und „Gestaltung des Alltagslebens und soziale Kontakte" des NBA gemessen werden geht davon aus, dass Selbständigkeit (Wert = 0) zur Unselbständigkeit (Wert = 3) über die beiden Abstufungen „überwiegend selbständig" (Wert = 1) und „überwiegend unselbständig" (Wert = 2) jeweils mit einem zugeordneten numerischen Wert +1 linear ansteigt. Die weiteren verwendeten Skalen für die übrigen Module gehen ebenfalls von einer linearen Quantifizierbarkeit von Fähigkeit, bzw. von Häufigkeiten aus. Für das NBA bedeutet dies, dass für alle Schritte des Messprozesses geprüft werden muss, ob die explizierte lineare Ordnungsrelation der als intervallskaliert verwendeten Itemwerte, Modulwerte und transformierter Modulwerte sowie der Index für Pflegebedürftigkeit die messtheoretischen Axiome der Reflexivität (a=a), der Symmetrie (wenn a>b, dann b<a), der Transitivität (wenn a>b und b>c, dann a>c) und der Konnexität (a<b oder b<a) erfüllen. Für das NBA ist dies davon abhängig, dass die Skalen (die numerischen Relationen) der einzelnen Module geeignet sind, die empirischen Relationen strukturgetreu abzubilden. Darüber hinaus stellt das erfüllte Repräsentationstheorem die Grundlage der Beurteilung des Eindeutigkeits-, bzw. des Bedeutsamkeitsproblems für das NBA dar.

Eindeutigkeitsproblem

Das Eindeutigkeitsproblem stellt in Frage, inwiefern gemessene Eigenschaften bei einer Transformation von Skalenwerten unverändert erhalten bleiben (siehe Abb. 2.4 und Tab. 2.1). Im NBA finden auf dem Berechnungsweg zu einem Gesamtindex für Pflegebedürftigkeit vier Transformationen der erhobenen Skalenwerte statt, deren Transformationsregeln normativ (aufgrund inhaltlicher Überlegungen) festgelegt wurden, ohne empirisch zu prüfen, ob die transformierte Skala gegenüber der ursprünglichen Skala in der Relation ihrer Differenzen unverändert (invariant) bleibt.

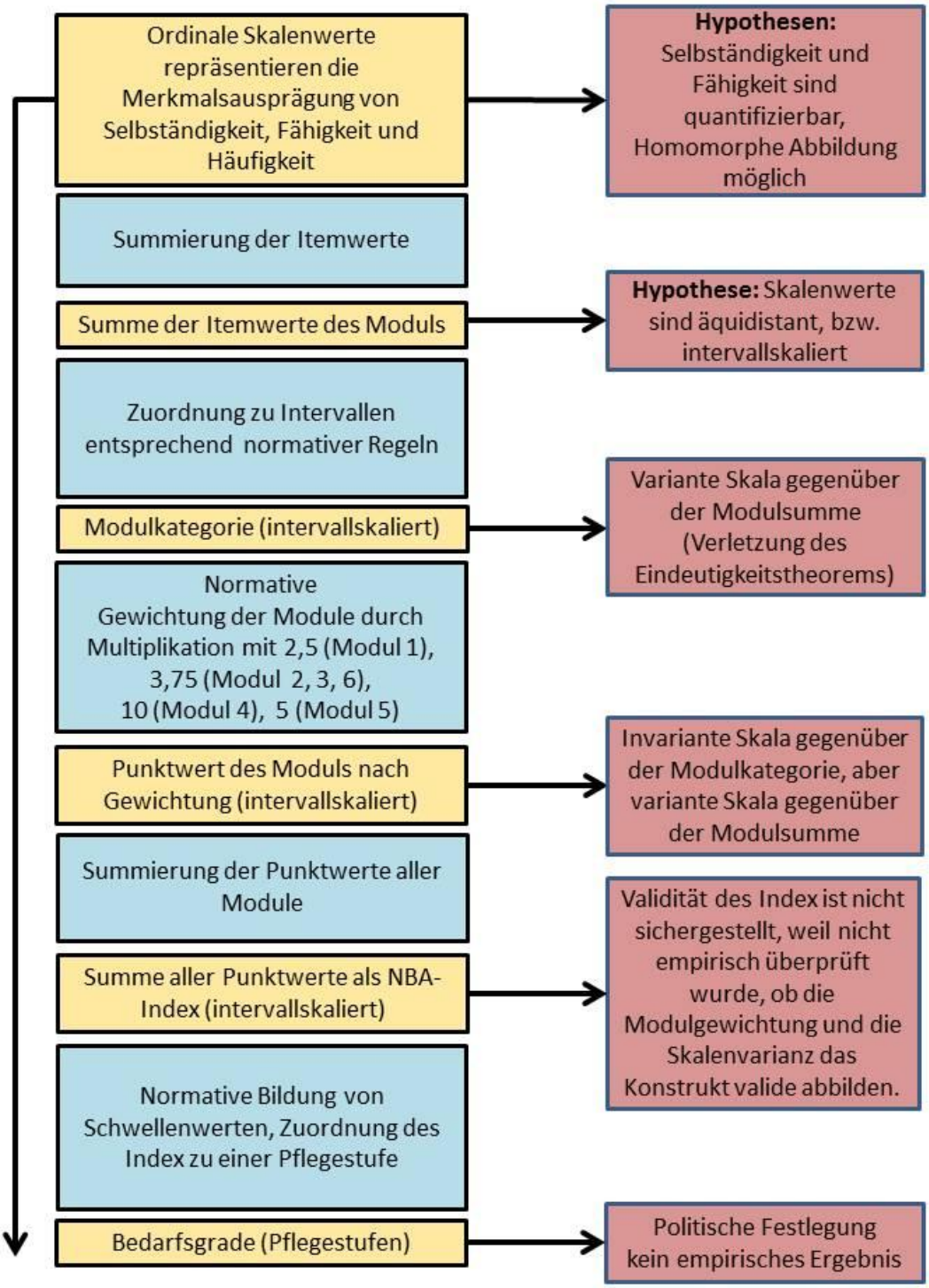


Abb. 2.4 Berechnungsschritte (gelb) anhand der Transformationsregeln (blau) und den daraus resultierenden Probleme (rot) des NBA-Messmodells

Bedeutsamkeitsproblem

Im Rahmen des Bedeutsamkeitsproblems gilt es zu klären, welche mathematischen und statistischen Verfahren bei welchem Messniveau (Skalenniveau) zulässig sind, weil sie aussagekräftige Ergebnisse bei der Prüfung der Konstruktvalidität liefern können. Viele allgemein übliche statistische Verfahren setzen für ihre Berechnungen intervallskalierte Daten voraus. Das Bedeutsamkeitsproblem für das NBA lässt sich also nur lösen, wenn im Zuge der Lösung des Repräsentationsproblems geklärt werden kann, welches Datenniveau die empirischen Daten, die mit den ordinal angelegten Skalen der Module erhoben wurden, tatsächlich aufweisen. Lässt sich eine Intervallskala nachweisen, können statistische Verfahren der Klassischen Testtheorie zur Konstruktvalidierung eingesetzt werden (z. B. Faktorenanalysen) (vgl. Brühl 2012, in diesem Band, S. 44). Stellt sich das Datenniveau als niedriger heraus (ordinal, nominal), müssen zur Konstruktvalidierung andere statistische Verfahren eingesetzt werden (z.B. logistische Rasch-Modelle) (vgl. Bensch 2012, in diesem Band, S. 117).

EXKURS KONSTRUKTVALIDITÄT

Ein Index gilt dann als valide, wenn er das Ergebnis des Messmodells strukturerhaltend in Bezug auf die Relationen des Strukturmodells abbilden kann.

Ein valider Index ist dementsprechend abhängig von der Übereinstimmung der Relationen des Strukturmodells mit den empirischen Relationen, die mithilfe des numerischen Relativs einer passenden Skala, bzw. Messmodells abgebildet werden. Struktur- und Messmodell sind nur gedanklich bei der Entwicklung und Validierung eines Instruments oder für dessen Analyse zu trennen und bilden gemeinsam ein komplementäres Ganzes. Wie das NBA zeigt, kann ein Konstrukt das Ergebnis theoretischer Überlegungen sein, die in ein entsprechendes theoretisches Messmodell überführt werden. Das NBA stellt damit eine Operationalisierung der theoretischen Annahmen über das Konstrukt der Pflegebedürftigkeit, als Ausmaß der Abhängigkeit von personeller Hilfe dar. Eine detaillierte Ausdifferenzierung der Bedingungen, Wechselwirkungen und Abhängigkeiten der inhaltlichen Aspekte auf das Maß des Angewiesenseins auf personelle Hilfe im Sinne eines nomologischen Netzes[30] wird für

[30] Ein nomologisches Netz stellt die Beziehungen, Wechselwirkungen und Abhängigkeiten der beobachtbaren Variablen, die auch Prädiktoren genannt werden, in Beziehung zum latenten Konstrukt dar. Die Beziehungen der beobachtbaren Variablen zum latenten Konstrukt lassen sich durch die Prüfung von Korrespondenzhypothesen konkretisieren und dienen damit sowohl der Entwicklung valider Tests oder Instrumente als auch der Präzisierung der Theorie.

das NBA nicht vorgenommen[31]. Die dem NBA implizite Theorie über die Relationen der Elemente von Pflegebedürftigkeit wird also nicht ex ante expliziert, sondern es kann nur versucht werden, mit Hilfe der Definition von Pflegebedürftigkeit und mit Hilfe des theoretischen Messmodells die impliziten theoretischen Beziehungen der Inhaltselemente zueinander ex post zu beschreiben.

Die Berechnungsregeln des Messmodells stellen implizite Hypothesen über die Relationen der Elemente des Konstrukts untereinander und in Bezug auf den Index als Gesamtergebnis dar. Ob diese Hypothesen gültig sind oder widerlegt werden können, lässt sich bedingt[32] durch die statistische Auswertung der empirischen Daten klären, die die Anwendung des Instruments hervorbringen. Die damit untersuchte Konstruktvalidität ist die Übereinstimmung des Strukturmodells mit dem empirischen Modell und stellt das wichtigste Gütekriterium eines Messinstruments dar. Besteht ein Instrument die Prüfung der Konstruktvalidität nicht, sind weitere Gütekriterien wie Zuverlässigkeit (Reliabilität) und Objektivität oder Praktikabilität zwangsläufig wenig aussagekräftig. Bei der Auswahl angemessener statistischer Verfahren zur Prüfung der Konstruktvalidität sind testtheoretische Grundsätze zu beachten (vgl. Brühl 2012, in diesem Band, S. 44ff). Die Validitätsprüfung des NBA (Windeler et al. 2008, S. 52ff) beschränkt sich auf die Kriteriumvalidität des Moduls „Kognitive und kommunikative Fähigkeiten“ und prüft nicht die Konstruktvalidität des gesamten NBA. Es wird also nicht getestet, ob das Instrument in sich konsistent ist (Skalierung, Dimensionen, Facetten), sondern ob sich die Einschätzung des Moduls „Kognitive und kommunikative Fähigkeiten“ vom Ergebnis des eingesetzten Kontrollverfahrens (TFDD[33]) signifikant unterscheidet. Eine besondere Brisanz in Bezug auf ein fehlendes Strukturmodell des Konstrukts Pflegebedürftigkeit ergibt sich daraus, dass das NBA an der bescheinigten Pflegestufe (Definition von Pflegebedürftigkeit nach § 14 SGB XI) der Studienteilnehmer validiert wurde (Windeler et al. 2008, S. 10). Seit der Einführung der Pflegeversicherung steht diese Definition der Pflegebedürftigkeit aufgrund ihrer

[31] „Dubin gibt zu bedenken, dass es zu vielen Problemen kommen wird, wenn summierende Konzepte in einer Theorie benutzt werden. Ein ganzes Phänomen wird durch ein oder zwei Worte erklärt. Jedes kumulative Konzept besteht aus zahlreichen, assoziativen und relationellen Konzepten und ihrem Zusammenwirken, aber keines wird benannt oder definiert. Man kann nur annehmen, welches dieser weniger komplexen Konzepte zu der kumulativen Größe beiträgt, wie sie zusammen wirken und unter welchen Bedingungen sie zu dem Phänomen beitragen. Dubin sagt, dass kumulative Konzepte geringen Nutzen in der Theorieentwicklung haben, da die Konzepte nicht genau definiert werden können und ihre Wechselbeziehungen nicht sofort feststellbar sind.“ (Keck 1992, 43)

[32] Wenn kein Strukturmodell zum Zweck systematisch statistisch prüfbarer Hypothesen entwickelt wurde, kann eine Prüfung der Konstruktvalidität meist nur einzelne Zusammenhänge beleuchten, die sich anhand impliziter Annahmen ableiten lassen. Fehlt ein explizierter Gesamtzusammenhang des Konstrukts (theoretische Relationen als nomologisches Netz) ist eine Validierung des Instruments als Ganzes unmöglich, weil keine Korrespondenzhypothesen formuliert werden können (siehe Franken, 2012 Kapitel 3 in diesem Band).

[33] Test zu Früherkennung von Demenzen mit Depressionsabgrenzung von Ihl et al, 2000

fehlenden theoretischen Fundierung und ihres Zeitbezugs bei der Erfassung erforderlicher Pflegeleistungen in der Kritik (Wingenfeld et al. 2007, S. 4ff).

Implizite Hypothesen

Das NBA wird von der impliziten Strukturhypothese dominiert, dass alle Items aller sechs Module zum Ausmaß von Pflegebedürftigkeit einen bestimmten Beitrag leisten.

Die jeweilige Gewichtung des Modulwertes im Gesamtindex wird anhand plausibler Abhängigkeiten der Module untereinander (z.B. Mobilität und kognitive und kommunikative Fähigkeiten) normativ festgelegt. Diese implizite Hypothese „je mehr Aspekte auf einen Menschen zutreffen (je höher der Index), desto pflegebedürftiger ist er“ spiegelt sich messtheoretisch in der Bildung von Summenscores wieder, daher kann von einem additiven und kompensatorischen Modell gesprochen werden.

Auf Modulebene kann die implizite Strukturhypothese geprüft werden, dass es sich bei der Modulstruktur um ein additives Modell handelt und dass die Summierung der als intervallskaliert angenommenen Messwerte trotz des ordinalen Skalenniveaus der Antworten ein valides Modulergebnis darstellt (siehe Franken 2012 in diesem Band, S. 112; Bensch 2012, in diesem Band, S. 138).

Eine weitere implizite Hypothese verbirgt sich in der Argumentation der numerischen Relationen der Items zu Aspekten der Ernährung, des Trinkens und der Ausscheidung im Modul „Selbstversorgung“ (siehe Tab. 2.1). Obwohl die Autoren in ihrem Bericht den Vorzug des NBA betonen, Pflegebedürftigkeit nicht anhand des Zeitaufwands für Pflegeleistungen einzuschätzen, wird die doppelte, bzw. dreifache Gewichtung dieser Items statt mit einer inhalts-theoretischen Begründung im Rahmen des Strukturmodells mit dem für die Kompensation erforderlichen Zeitaufwand gerechtfertigt (Wingenfeld et al. 2008, S. 50).

Diese Argumentation folgt der Hypothese, dass das Ausmaß der Pflegebedürftigkeit dann höher sei, wenn der Zeitaufwand zur Kompensation der Unselbständigkeit größer ist. Genau diese Betrachtungsweise wird bei der Entwicklung des NBA mit Blick auf die aktuelle Erfassung der Pflegebedürftigkeit kritisiert. (Wingenfeld et al. 2008, S. 28, 34)

Im Zuge des Umsetzungsberichts wird vorgeschlagen, die Skalen der Module „Mobilität“, „kognitive und kommunikative Fähigkeiten“, „Selbstversorgung“ und „Gestaltung des Alltagslebens und sozialer Kontakte“ zu modifizieren: die Kategorie „überwiegend selbständig“, die mit dem numerischen Wert „1“ verbunden ist, soll nicht

in die Bewertung einfließen (BMG 2009, S.19). In der Konsequenz bedeutet dies, dass ein Pflegebedürftiger, für den der Gutachter in allen fünf Items des Moduls „Mobilität“ z.B. aufgrund von chronischem Schwindel „überwiegend selbständig“ attestiert zwar fünf Punkte in diesem Modul erzielt, aber aufgrund der fehlenden Wertung dieser Kategorie „1“ bei der Transformation der Itemwerte in den Modulwert (siehe Tab. 2.1) leer ausgeht und mit null Modulpunkten als „selbständig“ eingeschätzt wird. Für die weiteren Berechnungen spielen die im Modul „Mobilität“ erfassten Einzelwertungen also keine Rolle mehr. Ein Pflegebedürftiger hingegen, der in allen vier Items als „selbständig“ eingeschätzt wird und ausschließlich beim Item „Treppensteigen“ eine „3“ für „unselbständig“ erzielt, erhält bei der Transformation der Itemwerte zum Modulwert eine „1“ („geringe Beeinträchtigung der Selbständigkeit“), die in die weitere Bewertung mit einfließt.

Probleme des NBA-Messmodells

Das Messmodell des NBA wird nicht als komplementäres Pendant eines theoretischen, empirisch prüfbaren Strukturmodells entwickelt.

Es besteht zwar eine Theorie über das, was für das Konstrukt der Pflegebedürftigkeit relevant sei, allerdings werden wahrscheinlich vorhandene Wechselwirkungen zwischen einzelnen Items und Modulen nicht in ihren Relationen spezifiziert.

Die inhaltlichen Relationen der Elemente, die das Konstrukt „Pflegebedürftigkeit“ konstituieren werden nicht expliziert. Das Problem, dass ein Kriterium Pflegebedürftigkeit sowohl begründen kann (z.B. die Items der Module Mobilität oder kognitive und kommunikative Fähigkeiten) als auch Pflegebedürftigkeit widerspiegeln kann (vgl. Franken 2012, in diesem Band, S. 81), wird im Messmodell ausschließlich im Rahmen der normativ festgelegten Gewichtung berücksichtigt.

Damit ist eine Validierung des Instruments, die zeitgleich eine Prüfung der theoretischen Annahmen darstellt, schwierig und in Anbetracht der Komplexität des Konstrukts der Pflegebedürftigkeit und der fehlenden Explikationen im Ganzen unmöglich.

Für die Prüfung der theoretischen Annahmen, die das Konstrukt konstituieren ist es erforderlich, prüfbare Hypothesen aus dem Strukturmodell heraus generieren zu können, um für die Validierung eine angemessen Methodologie und Methode zu wählen. Implizite und damit häufig unklare theoretische Strukturen wie sie das NBA aufweist sind dazu nicht geeignet (vgl. Brühl 2012, in diesem Band, S. 17; Franken

2012 in diesem Band, S. 82). Ohne explizierte theoretische Relationen bleibt als Möglichkeit der Validierung des Instruments nur die Kriteriumsvalidität. Die Kriteriumsvalidität sagt nicht direkt etwas über die „Binnenstruktur" des Instruments und damit der Gültigkeit der Wechselwirkungen theoretischer Inhalte untereinander aus, sondern ermittelt die Übereinstimmung der Indizes und die Homogenität der Bedarfsgrade. Geprüft wurde die Kriteriumsvalidität des NBA anhand des derzeit gültigen Pflegebedürftigkeitsbegriffs (Windeler et al. 2008, S. 52). Aufgrund der begründeten Kritik an der „Theorie" des geltenden Pflegebedürftigkeitsbegriffs können die aktuellen Pflegestufen aber weder aufgrund ihrer Validität noch ihrer theoretischen Fundierung als Referenzwert, bzw. als „Goldstandard" aufgefasst werden.

Aus messtheoretischer Perspektive ist es relevant zu überprüfen, ob die Theoreme der Messtheorie (Repräsentativität, Eindeutigkeit und Bedeutsamkeit) verletzt werden. Die Problematik des NBA in Bezug auf das Skalenniveau liegt neben der normativen Gewichtung der Module in der Behandlung der ordinalen Daten als intervallskalierte Daten. Das ordinale Datenniveau führt zu einer unbestimmten Gewichtung der Summanden, wodurch eine gewichtete Summe entsteht, deren Gewicht nicht bekannt ist und das in weiteren Berechnungen ignoriert wird. Mit diesem Verfahren gehen zum einen Informationen der Daten verloren, zum anderen wird mit Daten gerechnet, die nicht den ursprünglich numerischen Relationen und damit nicht mehr den empirischen Relationen entsprechen müssen und somit das Ergebnis erheblich verfälschen können. Die Untersuchungen von Bensch und Franken (Kapitel 4 und 5 in diesem Band) beziehen sich auf die Frage, ob für die Module „Mobilität" und „Kognitive und kommunikative Fähigkeiten" sowohl das Repräsentationsproblem als auch das Bedeutsamkeitsproblem gelöst werden konnten.

Dass innerhalb der Transformationsschritte von Modulwerten des NBA das Eindeutigkeitstheorem verletzt wird, lässt sich bereits in der Analyse der Transformationsregeln erkennen (siehe Abb. 2.4 und 2.5 sowie Tab. 2.1 im Anhang).

Im NBA werden ordinale Daten wie intervallskalierte Daten behandelt, um sie dann wieder in ordinale Daten zu transformieren[34]. Da dies nach unterschiedlichen Regeln geschieht, entstehen in der Berechnungssystematik zwei aufeinanderfolgende variante Skalen. Die Transformation der einzelnen Modulsummen in einen Modulwert/Modulkategorie führt zu einer varianten Skala, d. h., die ursprünglichen Relationen werden in der transformierten Skala (mit der weitergerechnet wird) nicht mehr korrekt abgebildet. Damit wird deutlich, dass die anhand der Plausibilität von Einzelfällen vorgenommene Gewichtung der Module, die aufgrund von qualitativen

[34] Bei der Aggregation der Items innerhalb der Module und bei der Integration der Modulwerte in den Gesamtindex.

Merkmalsprofilen (Ebene der Kriterien) entwickelt wurde, große Chancen hat, das Eindeutigkeitstheorem zu verletzen.

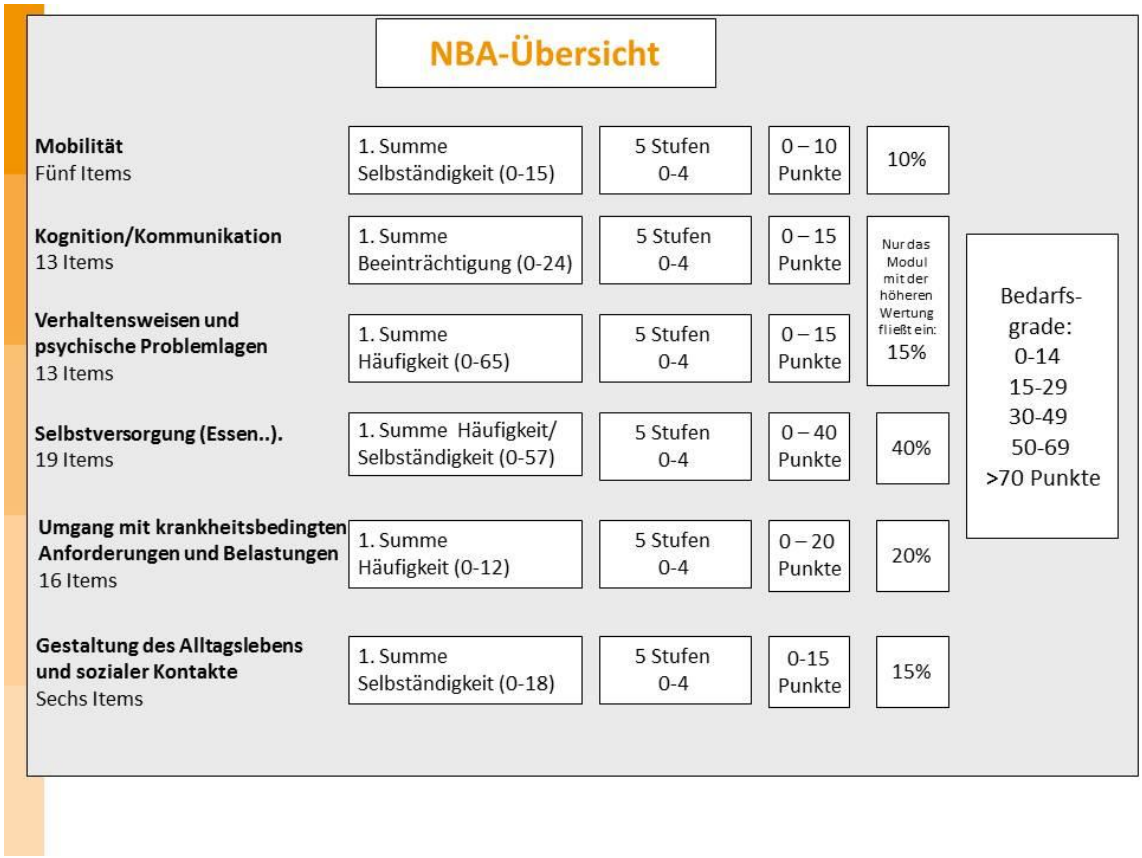


Abb. 2.5 Transformationsschritte des NBA von der Item-Ebene bis zum Bedarfsgrad

FAZIT

Obwohl Wingenfeld et al ihr Messmodell als „kompliziert“[35] beschreiben, bezieht sich die „Kompliziertheit“ eher auf die unterschiedlichen Transformations- und Gewichtungsverfahren in den einzelnen Schritten von der Einschätzung der einzelnen Items bis zum Index für Pflegebedürftigkeit (siehe Tab. 2.1 im Anhang) als auf das genutzte Messmodell. Die Bewertungssystematik beschränkt sich im Wesentlichen auf Additionen und kann damit als „einfaches Messmodell“ beschrieben werden, das mittels Summen bzw. Mittelwerten und prozentualen Gewichtungen Klassifikationen erzeugt.

[35] „Da die Bewertungssystematik keine bloße Summierung von Punktwerten, sondern ein relativ kompliziertes Verfahren darstellt, ist das Einschätzungsergebnis in der Begutachtungssituation selbst nur dann unmittelbar verfügbar, wenn die Begutachtung (wie heute vielfach üblich) EDV-gestützt erfolgt.“ (Wingenfeld et al 2008, S. 22)

Ein solches „einfaches Messmodell“ testtheoretisch als gültiges Modell identifizieren zu können wäre für die praktische Anwendbarkeit ideal. Liegt ein solches (annähernd ideales) Messmodell vor, bedeutet dies, dass es mess- und testtheoretisch recht hohen und sehr restriktiven Anforderungen gerecht wird. Dass das Messmodell des NBA die notwendigen Bedingungen zum Teil nicht erfüllt zeigen die Arbeiten von Franken und Bensch (Kapitel 4 und 5 in diesem Band) sowie die Analyse der Skalentransformationen (Abb. 2.4 und 2.5 sowie Tab. 2.1 im Anhang).

Aufgrund der fehlenden Explikationen eines zugehörigen Strukturmodells lassen sich nur begrenzt prüfbare Hypothesen ableiten, damit lässt sich das NBA als gesamtes nicht validieren.

Bei der Entwicklung von (Mess)Instrumenten ist die Komplementarität von Struktur- und Messmodell zu berücksichtigen, um eine Weiterentwicklung sowohl der Theorie des latenten Konstrukts als auch der Verbesserung der Validität des Instruments zu ermöglichen.

LITERATUR

Becker, Clemens; Blinkert, Baldo; Dietz, Berthold; Döhner, Hanneli; Frommelt, Mona; Klie, Thomas; Kruse, Andreas; Rothgang, Heinz (2007): Memorandum. Die Quadratur des Kreises in der Begutachtung der Pflegebedürftigkeit – Forschung statt Politik – Instrument vor Verfahren. http://www.bapp.info/texte/Memorandum-Pflegebeduerftigkeit.pdf zuletzt eingesehen am 02.07.2012

Bensch, Sandra (2012): Konstruktvalidität der Module „Mobilität und „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessments zur Feststellung von Pflegebedürftigkeit. Dissertation an der Philosophisch-Theologischen Hochschule Vallendar, Pflegewissenschaftliche Fakultät

Bortz, J.; Döring, N. (2006): Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler. 4. Aufl., Nachdr. Heidelberg: Springer

Bundesministerium für Gesundheit (2009): Bericht des Beirats zur Überprüfung des Pflegebedürftigkeitsbegriffs vom 26. Januar 2009. Berlin: BMG

Bundesministerium für Gesundheit (2009): Umsetzungsbericht des Beirats zur Überprüfung des Pflegebedürftigkeitsbegriffs vom 20. Mai 2009. Berlin: BMG

Brühl, Albert; Berger, Bianca (2011): Mit weniger Kriterien besser differenzieren: Warum bei der Messung von Qualität weniger mehr sein kann. Pflegewissenschaft 2011 (10) 525-534

Franken, Georg (2010): Konstruktvalidität der Subskala „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit (NBA). Masterarbeit an der Philosophisch Theologischen Hochschule Vallendar. http://opus.bsz-bw.de/kidoks/volltexte/2012/66/ zuletzt geprüft am 20.08.2012

Garms-Homolová, Vjenka; Theiss, Katrin (2007): Expertise „Möglichkeiten der Berücksichtigung von RAI 2.0 und/oder RAI HC bei der Erarbeitung eines zukünftigen Begutachtungsinstruments“ erstellt im Kontext von den Spitzenverbänden der Pflegekassen durchgeführten Modellvorhabens: „Maßnahmen zur Schaffung eines neuen reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI“. Berlin: Alice Salomon Fachhochschule

Grubitzsch, Siegfried (1991): Testtheorie-Testpraxis. Psychologische Tests und Prüfverfahren im kritischen Überblick. Reinbek: Rowohlt

Ihl, R.; Grass-Kapanke, B.; Lahrem, P.; Brinkmeyer, J.; Fischer, S.; Gaab, N.; Kaupmannsennecke, C. (2000): Entwicklung und Validierung eines Tests zur Früherkennung der Demenz mit Depressionsabgrenzung (TFDD). Fortschritte der Neurologie Psychiatrie 68, 413-422

Keck, Juanita Fogel (1992): Terminologie der Theorieentwicklung. In: Marriner-Tomey, Anne: Pflegetheoretikerinnen und ihr Werk. Basel: Recom

Rost, Jürgen (2004): Lehrbuch Testtheorie – Testkonstruktion. Zweite, überarbeitete und erweiterte Auflage. Bern: Huber

Rothgang, H.; Holst, M.; Kulik, D.; Unger, R. (2008): Finanzielle Auswirkungen der Umsetzung des neuen Pflegebedürftigkeitsbegriffs und des dazugehörigen Assessments für die Sozialhilfeträger und die Pflegekassen. Ergänzungsprojekt zum Modellprojekt „Entwicklung und Erprobung eines neuen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit“. Abschlussbericht unter Mitwirkung von Schneekloth, U. (TNS Infratest. Bremen: Zentrum für Sozialpolitik (ZeS)

Schaeffer, Doris; Wingenfeld, Klaus; Büscher, Andreas; Heine, U.; Gansweid, Barbara (2008): Das neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Anlagenband. Ergänzte und korrigierte Fassung vom 25. März 2008. Bielefeld: IPW; Münster: MBK WL

Steyer, Rolf; Eid, Michael (2001): Messen und Testen. Berlin; Heidelberg: Springer

Wingenfeld, Klaus; Büscher, Andreas; Schaeffer, Doris (2007): Recherche und Analyse von Pflegebedürftigkeitsbegriffen und Einschätzungsinstrumenten. Studie im Rahmen des Modellprogramms nach § 8 Abs. 3 SGB XI im Auftrag der Spitzenverbände der Pflegekassen. Bielefeld: IPW

Wingenfeld, Klaus; Büscher, Andreas; Gansweid, Barbara (2008): Das neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Projekt: Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI

Abschlussbericht zur Hauptphase 1: Entwicklung eines neuen Begutachtungsinstruments. Studie im Rahmen des Modellprogramms nach § 8 Abs. 3 SGB XI im Auftrag der Spitzenverbände der Pflegekassen. Bielefeld: IPW; Münster: MDK WL

Windeler, Jürgen; Görres, Stefan; Thomas, Stefanie (2008): Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen

Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Hauptphase 2 Abschlussbericht Endfassung. Bremen: IPP; Essen: MDS

Modulnr	Modul	Itemanzahl	Itemnr	Item	Codierung 1	Kategorien 1	Item-Messwert	Itemgewichtung	Summe aller Itemwerte des Modul s	Modulkategorie Codierung 2 (an FACE orientiert, 32)	Kategorien 2	Punktwert des Modul s aufgrund der Kategorie Codierung 3	Modulgewichtung für Gesamtwert (=Bedarfsgrad)
1	Mobilität	5	1.1. - 1.5	1.1 Positionswechsel im Bett	0	selbständig	0	Gleichgewichtung aller Items "überwiegend selbständig" fließt nicht in die Bewertung ein (Umsetzungsbericht S.19)	0	0	selbständig	0	10 %
				1.2 Stabile Sitzposition halten	1	überwiegend selbständig	1		1-3	1	geringe Beeinträchtigung der Selbständigkeit	2,5	
				1.3 Aufstehen aus sitzender Position/Umsetzen	2	überwiegend unselbständig	2		4-6	2	erhebliche Beeinträchtigung der Selbständigkeit	5	
				1.4 Fortbewegen innerhalb des Wohnbereichs	3	unselbständig	3		7-9	3	schwere Beeinträchtigung der Selbständigkeit	7,5	
				1.5 Treppensteigen					10-15	4	völliger Selbständigkeitsverlust	10	
2	Kognitive und kommunikative Fähigkeiten	13	2.1 - 2.13	2.1 Personen […] erkennen	0	vorhanden/unbeeinträchtigt	0	Nur Items 1 -8 fließen in die Bewertung ein	0	0	keine Beeinträchtigung	0	15 % es fließt nur das Modul mit dem höheren Punktwert ein
				2.2 Örtliche Orientierung 2.3 Zeitliche Orientierung	1	größtenteils vorhanden	1		1-4	1	geringe Beeinträchtigung kognitiver Fähigkeiten	3,75	
				2.4 Gedächtnis 2.5 […] Alltagshandlungen ausführen	2	in geringem Maße vorhanden	2	"überwiegend selbständig" fließt nicht in die Bewertung ein (Umsetzungsbericht S. 19)	5-8	2	erhebliche Beeinträchtigung kognitiver Fähigkeiten	7,5	
				2.6 Entscheidungen […] treffen 2.7 Sachverhalte+Infos verstehen	3	nicht vorhanden	3		9-13	3	schwere Beeinträchtigung kognitiver Fähigkeiten	11,25	
				2.8 Risiken+Gefahren erkennen					14-24	4	völliger/weitgehender Fähigkeitsverlust	15	
3	Verhaltensweisen und psychische Problemlagen	13	3.1 - 3.13	3.1 Motorisch geprägte Verhaltensweisen	0	nie	0	wenn 3.2 tgl. oder 3.3 wenigstens häufig auftritt oder komatöser Bewusstseinszustand, dann Stufe 4 (Codierung 2) keine Begründung für höhere Gewichtung	0	0	keine Beeinträchtigung	0	
				3.2 nächtliche Unruhe 3.3 selbstschädigendes und autoaggressives Verhalten	1	selten (1-2 x in 14 Tagen)	1		1-2	1	geringe Beeinträchtigung	3,75	
				3.4 Beschädigung von Gegenständen 3.5 Physisch aggr. Verhalten geg. Anderen 3.6 Verbale Aggression 3.7 Andere vokale Auffälligkeiten 3.8 Abw ehr pfleger./and. […]Maßnahmen 3.9 Wahnvorstellungen, Sinnestäuschungen 3.10 Ängste	2	häufig (2 - 6x wö, aber nicht tgl.)	3		3-4	2	erhebliche Beeinträchtigung	7,5	
				3.11 Antriebslosigkt, dep. Stimmungslage 3.12 soz. Inadäqu. Verhaltensweisen 3.13 Sonstige Inadäqu. Verhaltensweisen	3	täglich	5		5-6	3	schwere Beeinträchtigung	11,25	
									>6	4	schwerste Beeinträchtigung der Selbststeuerungskompetenz	15	

Tab. 2.1 Tabellarische Darstellung der Bewertungssystematik des Neuen Begutachtungsinstruments für Pflegebedürftigkeit (NBA) (Wingenfeld et al, 2008)
(Abweichungen von der allgemeinen Regel sind orange hinterlegt)

Modulnr	Modul	Itemanzahl	Itemnr	Item	Codierung 1	Kategorien 1	Item-Messwert	Itemgewichtung	Summe aller Itemwerte des Moduls	Modulkategorie Codierung 2 (an FACE orientiert, 32)	Kategorien 2	Punktwert des Modul aufgrund der Kategorie Codierung 3	Modulgewichtung für Gesamtwert (=Bedarfsgrad)
4	Selbstversorgung	19	B1 - B6 und 4.1 - 4.12	4.1 vorderen OK waschen 4.2 Kämmen/Zahn-Prothesenpflege/Rasieren 4.3 Intimbereich waschen 4.4 Duschen oder Baden 4.5 OK an-/auskleiden 4.6 UK an-/auskleiden 4.7 Nahrung mundgerecht zubereiten/Getränk eingießen	0 1 2 3	selbständig überwiegend selbständig überwiegend unselbständig unselbständig	0 1 2 3	"überwiegend selbständig" fließt nicht in die Bewertung ein (Umsetzungsbericht S.19) 4.7 geht nicht in die Modulsumme ein, wenn B1 mit "ausschliesslich oder nahezu ausschliesslich Sondennahrung" und B2 mit "Bedienung mit	0	0	keine Beeinträchtigung	0	40 %
			B1 Sondenernährung	zur oralen Nahrungsaufn. Gelegentl. Sondennahrung tgl. orale Nahr. + 1-3x tgl. Sondennahrung min. 4xtgl. Sondennahr. + tgl. geringe Mengen orale Nahr. (nahezu) ausschließlich Sondennahrung	Kategorial 0 5 9 12			wenn nur Sondenern. ohne orale Nahr. entfallen 4.7 - 4.9	1-9	1	geringe Beeinträchtigung	10	
			B2 Parenterale Ernährung	teilweise komplett	Kategorial 5 12			wenn komplett parenteral ernährt entfällt 4.7 - 4.9					
			4.8 Essen	selbständig überwiegend selbständig überwiegend unselbständig unselbständig	0 1 2 3		0 3 6 9	aufgrund des Unterstützungsumfangs (Zeit) dreifach gewichtet Geht nicht in die Modulsumme ein, wenn B1 mit "ausschliesslich oder nahezu ausschliesslich Sondennahrung" und B2 mit "Bedienung mit Fremdhilfe" bewertet wurde	10-24	2	erhebliche Beeinträchtigung	20	
			4.9 Trinken	selbständig überwiegend selbständig überwiegend unselbständig unselbständig	0 1 2 3		0 2 4 6	aufgrund des Unterstützungsumfangs (Zeit) zweifach gewichtet geht nicht in die Modulsumme ein, wenn B1 mit "ausschliesslich oder nahezu ausschliesslich Sondennahrung" und B2 mit "Bedienung mit Fremdhilfe" bewertet wurde					
			B3 Blasenkontrolle/Harninkontinenz B4 Darmkontrolle/Stuhlkontinenz B5 Künstliche Harnableitung B6 Colo-/Ileostoma		kein Kategorial			Items regeln die Bedingungen für die Berücksichtigung der Items 4.11 und 4.12	25-39	3	schwere Beeinträchtigung	30	
			4.10 Toilette/Toilettenstuhl benutzen	selbständig überwiegend selbständig überwiegend unselbständig unselbständig	0 1 2 3		0 2 4 6	aufgrund des Unterstützungsumfangs (Zeit) zweifach gewichtet					
			4.11.Folgen einer Harninkontinenz bewältigen, Umgang mit DK/Urostoma	selbständig überwiegend selbständig	0 1		0 1	geht in die Modulsumme nur ein, wenn B3/B4 mit "überwiegend-" oder "komplett inkontinent" beantwortet wurde oder künstl. Ableitung von Harn/Stuhl erfolgt (B5/B6)	>39	4	schwerste Beeinträchtigung der Selbststeuerungskompetenz	40	
			4.12 Folgen einer Stuhlinkontinenz bewältigen, Umgang mit Stoma	überwiegend unselbständig unselbständig	2 3		2 3						

Tab. 2.1 Tabellarische Darstellung der Bewertungssystematik des Neuen Begutachtungsinstruments für Pflegebedürftigkeit (NBA) (Wingenfeld et al, 2008) (Abweichungen von der allgemeinen Regel sind orange hinterlegt)

Modulnr	Modul	Itemanzahl	Itemnr	Item	Codierung 1	Kategorien 1	Item-Messwert	Itemgewichtung	Summe aller Itemwerte des Moduls	Modulkategorie Codierung 2 (an FACE orientiert, 32)	Kategorien 2	Punktwert des Moduls aufgrund der Kategorie Codierung 3	Modulgewichtung für Gesamtwert (=Bedarfsgrad)
5	Umgang mit krankheits-/therapiebedingten Anforderungen und Belastungen	16	5.1 - 5.7	5.1Medikation 5.2 Injektionen 5.3 Versorgung i.v. Zugänge 5.4 Absaugen oder Sauerstoffgabe 5.5 Einreibungen, Kälte-/Wärmeanw endung 5.6 Messung und Deutung von Körperzuständen 5.7 Umgang mit körpernahen Hilfsmitteln	entfällt, selbständig, tgl., wöchentl., monatl., < 6	Seltener als 1x tgl.	0	"Die Bündelung der Items bzw. die Unterteilung dieser Bereiche orientiert sich größtenteils am Aufwand, den sie mit sich bringen." [Zeit, Anm. KP] (Wingenfeld et al 2008, 54f)	0	0	selbständig/unabhängig	0	20 %
						1-3 x tgl.	1						
						4-8 x tgl.	2	Es fließt nur der jeweils höchste Wert jedes der vier "Bündel" in den Modulwert ein, daher sind max. 12 Punkte möglich	1	1	geringe Beeinträchtigungen der Selbständigkeit	5	
						> 8 x tgl.	3	Für die Items 5.1 -5.16 gilt: ausgeschlossen w erden Aktivitäten und Maßnahmen, die über einen kürzeren Zeitraum als sechs Monate erforderlich sind					
			5.8 - 5.11	5.8 Verband w echsel/ Wundversorgung 5.9 Wundversorgung bei Stoma 5.10 Regelm. Einmalkatheterisierung, Nutzung von Abführmethoden 5.11 Therap.maßn. in häusl. Umgebung	entfällt, selbständig,	seltener als 1x w ö.	0		2-3	2	erhebliche Beeinträchtigungen der Selbständigkeit	10	
						1 - mehrmals w ö.	1						
						1-2 x tgl.	2						
						min. 3 x tgl.	3						
			5.12 - 5.15	5.12 zeitl. Ausgedehnte technikintens. Maßn. In häusl. Umgebung 5.13 Arztbesuche 5.14 Besuch anderer med./therap. Einrichtungen (< 3 h) 5.15 zeitl. ausgedehnter Besuch med./therap. Einrichtungen (> 3 h)	entfällt, selbständig, tgl., w öchentl., monatl., < 6 Mon.	0 - <4,3 Punkte	0	monatliche Besuche bei Ärzten und med. Einrichtungen = 1 Punkt monatliche, zeitaufw ändige Besuche sow ie zeit- und technikintensive Maßnahmen in häuslicher Umgebung w erden mit 2 Punkten bew ertet					
						4,3 - <8,6 Punkte	1	w öchentliche Besuche bei Ärzten und med. Einrichtungen = 4,3 Punkte	4-5	3	schw ere Beeinträchtigung der Selbständigkeit	15	
						8,6 - <12,9 Punkte	2	w öchentliche, zeitaufw ändige Besuche sow ie zeit- und technikintensive Maßnahmen in häuslicher Umgebung w erden mit 8,6 Punkten bew ertet					
						ab 12,9 Punkte	3						
			5.16	Einhaltung einer Diät oder anderer Verhaltensvorschriften	0	entfällt/nicht erforderlich oder selbständig	0		6-12	4	w eitgehende Abhängigkeit von anderen Personen	20	
					1	überw iegend selbständig	1						
					2	überw iegend unselbständig	2						
					3	unselbständig	3						
6	Gestaltung des Alltagslebens & soziale Kontakte	6	6.1 - 6.6	6.1 Tagesabl. gestalten und anpassen 6.2 Ruhen und Schlafen 6.3 Sich beschäftigen 6.4 In die Zukunft gerichtete Planungen vornehmen 6.5 Interaktion mit Personen im direkten Kontakt 6.6 Kontaktpflege zu Personen außerhalb des direkten Umfelds	0	selbständig	0	Gleichgew ichtung aller Items "überw iegend selbständig" fließt nicht in die Bew ertung ein (Umsetzungsbericht S.19)	0	0	selbständig	0	15 %
					1	überw iegend selbständig	1		1-3	1	geringe Beeinträchtigung der Selbständigkeit	3,75	
					2	überw iegend unselbständig	2		4-6	2	erhebliche Beeinträchtigung der Selbständigkeit	7,5	
					3	unselbständig	3		7-11	3	schw ere Beeinträchtigung der Selbständigkeit	11,25	
									12-18	4	völliger Selbständigkeitsverlust	15	

Tab. 2.1 Tabellarische Darstellung der Bewertungssystematik des Neuen Begutachtungsinstruments für Pflegebedürftigkeit (NBA) (Wingenfeld et al, 2008)
(Abweichungen von der allgemeinen Regel sind orange hinterlegt)

3. Zur Verwendung reflektiver und formativer Indikatoren am Beispiel des NBA

Georg Franken

Einleitung

Assessmentinstrumente sollen relevante Phänomene zuverlässig und gültig erfassen. Um die entsprechende Güte eines Instruments zu beurteilen, werden häufig Verfahren zur internen Konsistenz oder Konstruktvalidität eingesetzt, die im Rahmen der Klassischen Testtheorie voraussetzen, dass ein nicht unmittelbar beobachtbares und daher latentes Konstrukt die Ausprägungen in den beobachtbaren Variablen, den Indikatoren des Konstrukts, bestimmt (vgl. Brühl 2012, in diesem Band, S. 18). Sofern die Indikatoren einer Skala dasselbe Konstrukt erfassen, korrelieren dabei ihre Ausprägungen und es kann zur Einschätzung der Güte einzelner Indikatoren oder der ganzen Skala die Stärke dieses Zusammenhangs herangezogen werden. Nicht immer aber erscheint es sinnvoll, diese Annahmen für Assessmentinstrumente vorauszusetzen. So ist es fraglich, ob ADL-Skalen (Activity of Daily Living) den Grad der Selbständigkeit mit Indikatoren erheben, deren Ausprägungen notwendig miteinander korrelieren müssen (Streiner 2003, S. 217; Bartholomeyczik, Halek 2009, S. 18). Eine Anwendung gängiger Verfahren zur Beurteilung der Reliabilität und Konstruktvalidität, die die Annahmen der KTT voraussetzen, führt dann aber zur Fehleinschätzung der Güte solcher Instrumente (Bollen, Lennox 1991; Jarvis et al. 2003; MacKenzie et al. 2005).

Für die Entwicklung und Beurteilung eines Instruments muss daher geklärt werden, welche Messmodelle in den Beziehungen zwischen Indikatoren und dem erfassten Konstrukt konzeptionell vorausgesetzt werden.

Im Folgenden werden dazu zwei Modelle dargestellt, die Unterschiede in der Beurteilung ihrer Güte erläutert und vor diesem Hintergrund das NBA diskutiert (Franken 2010, S. 28–33).

Reflektive und formative Messmodelle

Aus Sicht der Klassischen Testtheorie und Faktorenanalyse werden Indikatoren als Effekte einer latenten Variable betrachtet, die ein Konstrukt repräsentiert (Bollen, Lennox 1991):

$x_i = \lambda_i \xi + \delta_i.$ (1)

Dabei wird der Wert eines einzelnen Indikators x_i mit $i = 1, 2, \ldots, n$ durch den Einfluss λ_i der latenten Variablen ξ und einem Messfehler δ_1 bestimmt. Im Sinne der Klassischen Testtheorie wird angenommen, dass der Erwartungswert[36] des Messfehlers δ_i Null ist und die Messfehler der einzelnen Variablen voneinander und von der latenten Variable unabhängig sind. Indikatoren, deren Ausprägung so durch eine latente Variable bedingt sind, werden als reflektiv bezeichnet (Backhaus et al. 2011, S. 528). Wird test- und messtheoretisch angenommen, dass es sich um genau eine latente Variable für *n* beobachtbare Indikatoren handelt, können die so verbundenen Indikatoren als eindimensional betrachtet werden (Bühner 2006, S. 302). Das folgende Pfaddiagramm stellt ein entsprechendes reflektives Messmodell mit n = 3 Indikatoren dar.

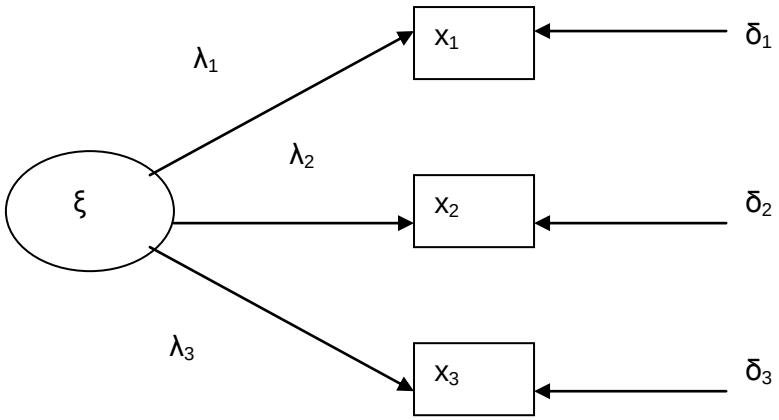


Abb. 3.1 Eindimensionales Modell reflektiver Indikatoren (Quelle: Edwards, Bagozzi 2000, S. 161)

Der Anteil gemeinsamer Varianz aller Indikatoren kann als „wahrer Wert" des Konstrukts verstanden werden. Die Kovarianzen zwischen den Indikatoren werden ausschließlich aus der Wirkung der gemeinsamen latenten Variablen erklärt (Edwards, Bagozzi 2000, S. 161).[37] Der systematische Zusammenhang zwischen den Indikatoren lässt sich so empirisch an ihren Korrelationen ablesen. Er „verschwindet", wenn der Einfluss der latenten Variablen auf die Indikatoren „ausgeschaltet" wird, indem beispielsweise die latente Variable zwischen den Messungen konstant gehalten wird. Die Indikatoren sind in diesem Sinne lokal unabhängig (Bühner 2006, S. 21). Umgekehrt sollten die Indikatoren möglichst hoch miteinander korrelieren, wenn sich

[36] Unter dem Erwartungswert versteht man den Mittelwert einer theoretischen Verteilung (Bühner 2006, S. 27), zur Unterscheidung von Mittelwert und Erwartungswert vgl. Bortz 2005, S. 64f

[37] Insofern enthält das Konstrukt auch alle weiteren Faktoren, die den Zusammenhang zwischen den Indikatoren bestimmen (Streiner 2003, S. 219).

die Konstruktausprägung ändert. Daher sind die Indikatoren bei gleicher Reliabilität austauschbar und das Hinzufügen oder Streichen einzelner Indikatoren verändert zwar die Messgenauigkeit, nicht aber das Konstrukt (Jarvis et al. 2003, S. 200).

Nicht immer aber folgt das Verhältnis zwischen den Indikatoren und der latenten Variablen diesem Modell. So enthalten Assessmentinstrumente zur gesundheitsbezogenen Lebensqualität häufig krankheits-, behandlungs- oder versorgungsbezogene Indikatoren wie Schlafstörungen, Appetitlosigkeit oder finanzielle Belastungen aufgrund von Krankheitskosten, die die Lebensqualität beeinflussen, ohne dass dies im umgekehrten Sinn gelten würde (Fayers, Hand 1997; Streiner 2003; Bollen et al. 2009). Verringert sich dann beispielsweise durch Leistungen der Sozialversicherung das Ausmaß der finanziellen Belastungen, so erhöht dies die Lebensqualität, ohne dass sich die Ausprägungen der übrigen Merkmale verändern müssen. In diesem Messmodell bestimmen die Indikatoren das Konstrukt.

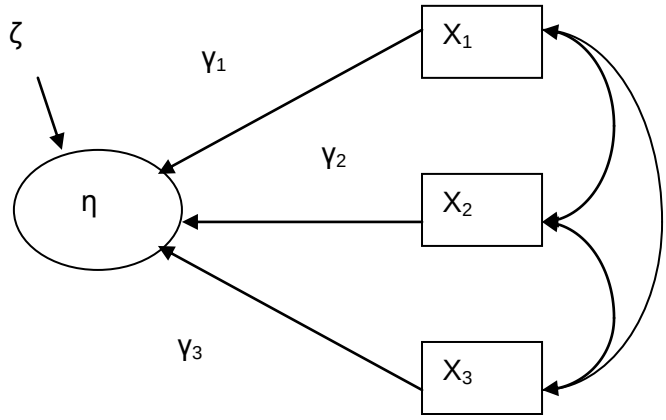


Abb. 3.2 Eindimensionales Modell formativer Indikatoren (Quelle: (Edwards, Bagozzi 2000, S. 162)

Abbildung 3.2 zeigt die Spezifikation eines Konstrukts η, dessen Ausprägung durch die Indikatoren x_i, mit i = 1 bis 3, und einem Messfehler ζ bestimmt wird. Die Pfeile zeigen wiederum die Richtung der Beeinflussung an. Das Modell lässt sich durch die folgende Gleichung ausdrücken:

$$\eta = \sum \gamma_i x_i + \zeta.. \qquad (2)$$

Die latente Variable ist als Linearkombination ihrer Indikatoren definiert. ζ kennzeichnet den Anteil der Verteilung der latenten Variablen, der nicht durch die Indikatoren erklärt

werden kann und bezeichnet damit den Fehlerterm auf der Ebene des Konstrukts.[38] Die Varianz des Konstrukts ist die vom Fehlerterm bereinigte Varianz aller Indikatoren. Umgekehrt erklärt das Konstrukt aufgrund der angenommenen Kausalbeziehungen nicht die Varianz der Indikatoren oder deren Kovarianz (Edwards, Bagozzi 2000, S. 162)[39]. Die Indikatoren können daher miteinander korrelieren, müssen dies aber nicht. Interne Konsistenz wird daher nicht verlangt. Die Indikatoren sind zudem nicht austauschbar, da sie den Umfang des Konstrukts bestimmen. Sie formieren vielmehr als „Bausteine“ oder „Dimensionen“ das Konstrukt und werden daher auch als formative Indikatoren bezeichnet (Backhaus et al. 2011, S. 528).[40]

Formative Messmodelle entsprechen nach Auffassung einiger Autoren in empirischen Studien häufig eher den Grundannahmen eines Konstrukts, auch wenn die jeweiligen Forscher in ihrem methodischen Vorgehen reflektive Modelle voraussetzen (Jarvis et al. 2003; Eggert, Fassott 2003; Fassott 2006). Fehlspezifikationen führen jedoch zu methodischen Mängeln in der Testkonstruktion, der Beurteilung von Modellen und Einschätzung von Gütekriterien (Bollen, Lennox 1991; Fayers, Hand 1997; Diamantopoulos, Winklhofer 2001; Streiner 2003; MacKenzie et al. 2005).

Reliabilität und Validität reflektiver und formativer Messmodelle

Während die Güte reflektiver Messmodelle mit Hilfe der geläufigen Verfahren zur Beurteilung der Reliabilität und Validität eingeschätzt werden kann, lassen sich die im Rahmen der KTT entwickelten Methoden nicht auf die Entwicklung und Bewertung formativer Instrumente anwenden. Da formative Indikatoren nicht miteinander korrelieren müssen, ist insbesondere eine Skalenbereinigung oder Einschätzung der Reliabilität über das Kriterium interner Konsistenz unangemessen[41] und kann zu einer

[38] Im Unterschied zu den Messfehlern einzelner Indikatoren eines reflektiven Messmodells fließen in den Fehlerterm eines formativen Messmodells Messfehler, Wechselwirkungen zwischen den Indikatoren sowie Dimensionen des Konstrukts ein, die mit den Indikatoren nicht erfasst werden (MacKenzie et al. 2005, S. 712).

[39] Insofern wird von den Indikatoren eines formativen Messmodells auch nicht verlangt, dass sie lokal unabhängig sind.

[40] Das in Abbildung 2 dargestellte Modell ist in der vorliegenden Form unteridentifiziert. Für die Schätzung der Parameter muss es in einem umfassenderen Modell eingeordnet werden, in dem die latente Variable mit reflektiven Indikatoren verbunden ist (Bollen, Lennox 1991, S. 312; MacKenzie et al. 2005, S. 726).

[41] Wenn die Variablen in den Gleichungen (1) und (2) mit einem Mittelwert von 0 und einer Standardabweichung von 1 standardisiert werden, entspricht die Korrelation zweier Indikatoren y_1 und y_2 in Gleichung (1) dem Produkt der Koeffizienten $\lambda_1\lambda_2$. Da nach dem Modell die Indikatoren mit der latenten Variablen korrelieren sollen, sind alle Koeffizienten bei gleicher Messrichtung entsprechend der Forderung interner Konsistenz positiv. In Gleichung (2) ist dagegen die Korrelation zweier Indikatoren x_1 und x_2 vom Modell her unbestimmt. Die Korrelationen zwischen formativen Indikatoren können daher positiv, Null oder negativ sein (Bollen, Lennox 1991, S. 307; Streiner 2003, S. 218f.).

irreführenden Veränderung der Bedeutung eines Konstrukts führen.[42] Hohe interne Konsistenz zwischen formativen Indikatoren lassen vielmehr vermuten, dass die Indikatoren einer Skala das Konstrukt zu eng fassen, statt es im Sinne seiner Definition inhaltlich umfassend zu operationalisieren (Streiner 2003, S. 220). Hohe Korrelationen zwischen den Indikatoren erschweren es zudem, den spezifischen Beitrag eines einzelnen Indikators zu bestimmen (Bollen, Lennox 1991, S. 307; Diamantopoulos, Winklhofer 2001, S. 272; MacKenzie et al. 2005, S. 712; Giere et al. 2006, S. 687). Um die Zuverlässigkeit einer formativen Skala zu prüfen, bleiben daher nur die Retestmethode, Interrater-Reliabilität oder Korrelation einzelner Items mit reflektiven Erhebungen desselben Aspekts (Bagozzi 1994, S. 333; MacKenzie et al. 2005, S. 727).

Unter den Kriterien zur Gültigkeit formativer Messmodelle kommt ihrer Inhaltsvalidität besondere Bedeutung zu. Während in reflektiven Modellen die Indikatoren ein „Universum“ von möglichen Indikatoren repräsentieren, aus denen sie prinzipiell frei wählbar sind, sind formativ gemessene Konstrukte von ihren Indikatoren abhängig. Ihre Indikatoren müssen daher die gesamten inhaltlichen Bereiche erfassen, die den spezifischen Umfang eines Konstrukts definieren (Bollen, Lennox 1991, S. 308). Es liegen jedoch bislang keine statistischen Verfahren vor, um insbesondere die Konstruktvalidität eines einzelnen formativen Messmodells einzuschätzen. Vielmehr bedarf es dazu theoretischer Annahmen, die das formative Konstrukt mit reflektiven Modellen verbinden (Diamantopoulos, Winklhofer 2001; MacKenzie et al. 2005). Lässt sich das interessierende Konstrukt so in ein nomologisches Netz einordnen, in dem das Konstrukt den Einfluss der Indikatoren auf reflektive Variablen vermittelt, können die Parameter des gesamten Modells eingeschätzt werden (multiple indicators multiple causes model (MIMIC), vgl. Jöreskog, Goldberger 1975) oder der theoretisch postulierte Zusammenhang zu anderen reflektiven Konstrukten oder Variablen genutzt werden, um die Kriteriumsvalidität des interessierenden Konstrukts oder einzelner seiner Indikatoren zu prüfen (Diamantopoulos, Winklhofer 2001, S. 272; Fassott, Eggert 2005, S. 41; MacKenzie et al. 2005, S. 726–728).

[42] Hribek und Schmalen (2000) entwickeln ein Instrument zur Patientenzufriedenheit im Krankenhaus, in dem der Indikator „Freundlichkeit des Pflegepersonals“ aus dem Konstrukt „Interaktionsqualitäten des Pflegepersonals“ aufgrund mangelnder Reliabilität entfernt wurde. Für das Fehlen dieses Indikators können die Autoren allerdings nach eigenem Bekunden „keine inhaltlich plausible Erklärung“ geben (Hribek, Schmalen 2000, S. 225 Anm. 29; Fassott, Eggert 2005, S. 45). Fayer und Hand (1997) weisen zudem darauf hin, dass die Korrelation formativer Indikatoren im besonderen Maße stichprobenabhängig ist. Dies gilt insbesondere für Skalen zur gesundheitsbezogenen Lebensqualität oder ADL-Skalen, die Krankheitssymptome oder die Nebenwirkungen von Behandlungen umfassen (Fayers, Hand 1997; Streiner 2003, S. 220). Eine Itemselektion mit Hilfe einer Faktorenanalyse, bei der die Indikatoren ausgewählt werden, die besonders hoch auf eine spezifische Komponente laden, kann dann dazu führen, dass wesentliche Aspekte eines Konstrukts gestrichen werden (Juniper et al. 1994).

IDENTIFIKATION DER MESSMODELLE IM NBA

Die Unterschiede zwischen den Messmodellen machen es notwendig, für die Entwicklung und Beurteilung eines Assessmentinstruments die Relationen zwischen Konstrukten und deren Indikatoren zu bestimmen. Zur Einschätzung einer Spezifikation geben Jarvis, MacKenzie, Podsakoff et al. vier Kriterien an, um reflektive und formative Modelle voneinander zu unterscheiden[43] (Jarvis et al. 2003, S. 203, vgl. MacKenzie et al. 2005, S. 713): (1) die kausale Richtung zwischen Konstrukt und Indikatoren, [44] (2) die Austauschbarkeit der Indikatoren, (3) die Kovarianz unter den Indikatoren sowie (4) ihre Einbindung in ein nomologisches Netz. [45] Um die Voraussetzungen für eine Konstruktvalidierung des NBA zu klären, werden im Folgenden die darin enthaltenen Messmodelle anhand dieser Kriterien eingeschätzt. Dabei wird „Pflegebedürftigkeit" als mehrdimensionales Konstrukt höherer Ordnung, in dem der systematische Zusammenhang verschiedener Konstrukte durch die Beziehung zu einem gemeinsamen Konstrukt höherer Ordnung spezifiziert wird, auf dieser höheren Ebene analog untersucht.

Die theoretische Grundlage zum NBA bildet der vom Institut für Pflegewissenschaft an der Universität Bielefeld ausgearbeitete Pflegebedürftigkeitsbegriff (Wingenfeld et al. 2007). Pflegebedürftigkeit wird darin durch körperliche oder psychische Schädigungen, die Beeinträchtigung körperlicher oder kognitiv/psychischer Funktionen sowie gesundheitlich bedingte Belastungen oder Anforderungen verursacht, die nicht durch persönliche Ressourcen bewältigt oder kompensiert werden können. Sie manifestiert sich in verschiedenen Aktivitäten im Lebensalltag, der Krankheitsbewältigung, der Gestaltung von Lebensbereichen und sozialen Teilhabe, die zusammen den Umfang der Pflegebedürftigkeit konstituieren (Abb. 3.3). Dabei erwarten die Autoren, im nachfolgenden Prozess der Instrumentenentwicklung die Aktivitäten und Bereiche der Pflegebedürftigkeit im Einzelnen zu bestimmen und damit den Pflegebedürftigkeitsbegriff weiter zu klären (Wingenfeld et al. 2007, S. 107f).

---

[43] Einen empirischen Modelltest zur Prüfung alternativer Spezifikationen schlagen Bollen und Ting vor. Der Confirmatory Tetrad Analysis Test (CTA) oder auch Vanishing Tetrad Test (VTT) beruht darauf, dass unter Annahme eines reflektiven Messmodells die Differenzen paarweise gebildeter Produkte der Kovarianzen, sogenannte Tetraden $\tau_{ghij} = \sigma_{gh}\sigma_{ij} - \sigma_{gi}\sigma_{hi}$, gleich Null sein müssen (Bollen, Ting 1993, Bollen, Ting 2000; Bollen et al. 2009).

[44] Obwohl häufig die Beziehung zwischen formativen Indikatoren und ihren Konstrukten analog zu denen in reflektiven Messmodellen als „kausal" und so auch die Indikatoren u. a. als „causal indicators" bezeichnet werden, wird damit zunächst nur die Messrichtung charakterisiert, in der ein Konstrukt und seine Indikatoren einander bestimmen, ohne dass damit die weiterreichenden Bedeutungen eines Kausalitätsbegriffs beinhaltet sein müssen (Bollen, Lennox 1991, S. 306; Fayers, Hand 1997, S. 145; Fayer et al. 1997, S. 395; Streiner 2003, S. 219 Anm.4; zum spezifischen Verständnis kausaler Beziehungen Bollen 1989; Edwards, Bagozzi 2000; zur erkenntnistheoretischen Diskussion Borsboom et al. 2003).

[45] Die Unterscheidung der Spezifikationen ist häufig dadurch erschwert, dass das Verhältnis von Ursache und Wirkung mangels theoretisch fundierter Modelle nicht eindeutig bestimmt werden kann. Hinzukommt, dass einzelne Skalen sowohl reflektive wie auch formative Indikatoren enthalten können (Streiner 2003, S. 221, vgl. Fayer et al. 1997, S. 395f).

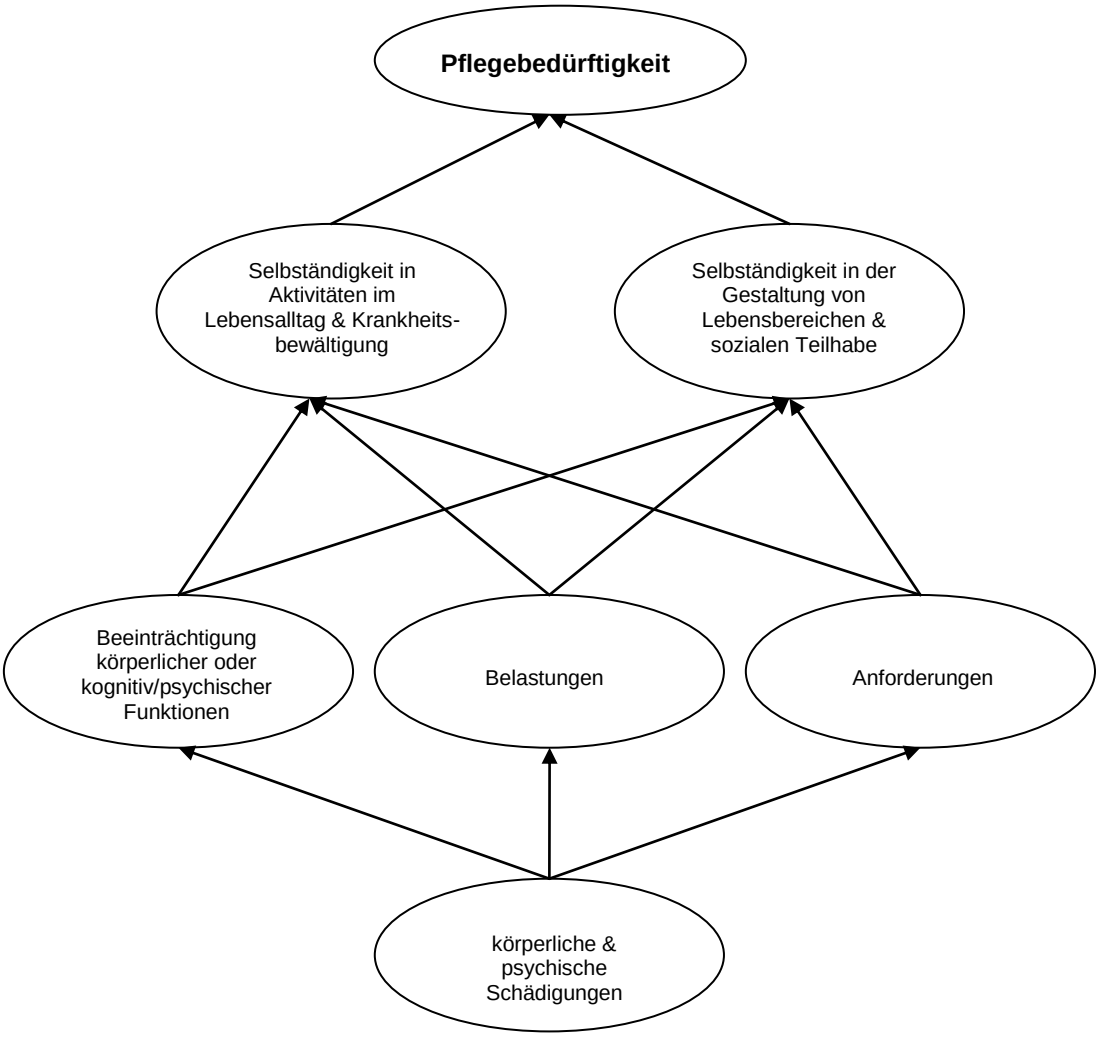


Abb. 3.3 Identifizierte und grafisch dargestellte Struktur des Pflegebedürftigkeitsbegriffs des NBA (vgl. Wingenfeld et al. 2007, S. 40-43)

Die Konstruktion des NBA setzt bei den Aktivitäten und Lebensbereichen als Wirkungsbereichen an und überträgt sie in Module. Ursächliche Funktionen, Belastungen und Anforderungen werden in eigenständigen Modulen oder Indikatoren einzelner Wirkungsbereiche erfasst. Die Module bilden dabei Dimensionen oder Facetten der Pflegebedürftigkeit als einem Konstrukt höherer Ordnung. Die Struktur des Modells der Pflegebedürftigkeit wird dabei jedoch nicht explizit bestimmt. Zieht man die oben genannten Kriterien zur Unterscheidung der Messmodelle heran, so scheint eine formative Perspektive zu überwiegen.[46]

Die Kausalität bezieht sich darin auf den Informationsfluss zwischen Indikatoren/Dimensionen und Konstrukt: Sind die Indikatoren/Dimensionen Ausprägungen des Konstrukts oder bestimmen sie es? Verursachen Veränderungen in

[46] Die folgenden Überlegungen beziehen sich auf die Module 1–6 (NBA), in denen die Pflegebedürftigkeit erhoben wird (Wingenfeld et al. 2008, S. 75).

den Indikatoren bzw. Dimensionen Änderungen im Konstrukt oder umgekehrt? Die Indikatoren der Module „Mobilität“, „Kognitive und kommunikative Fähigkeiten“, „Gestaltung des Alltagslebens und soziale Kontakte“ oder in den Bereichen zu Körperpflege, An- und Ausziehen in der „Selbstversorgung“ können danach reflektiv interpretiert werden, in anderen Bereichen der „Selbstversorgung“ oder dem „Umgang mit krankheits-/therapiebedingten Anforderungen und Belastungen“ wird teilweise erst erhoben, ob bestimmte Anforderungen vorliegen, bevor die Selbständigkeit in ihrer Bewältigung eingeschätzt wird. Hier verläuft der Informationsfluss zunächst von den Indikatoren zum Konstrukt. Im Modul 3 wiederum bedingen aus Sicht der Autoren unbewältigte psychische Problemlagen die erhobenen Verhaltensweisen (Wingenfeld et al. 2008, S. 43). Andererseits verursachen Änderungen im Verhalten Änderungen im Konstrukt, während umgekehrt eine Änderung im Konstrukt nicht auf das Vorkommen oder die Häufigkeit bestimmter Verhaltensweisen schließen lässt. Hier lässt sich aus Beschreibung und zu erwartender Messrichtung kein übereinstimmendes Messmodell bilden. Eine Stufe höher lässt sich Pflegebedürftigkeit selbst dagegen als Funktion ihrer Bereiche verstehen (vgl. Abb. 3.3).

In einem reflektiven Modell sollten des Weiteren die Indikatoren bzw. Dimensionen prinzipiell austauschbar sein, in einem formativen Modell bestimmen dagegen die Indikatoren bzw. Dimensionen das Konstrukt konzeptionell. Für „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ können in diesem Sinne reflektive Modelle angenommen werden. So selektieren die Autoren auch ausschließlich im Modul 2 die Items, indem sie die „stärker kommunikationsbezogenen Merkmale“ (Wingenfeld et al. 2008, S. 41) aufgrund ihrer hohen Korrelation mit der restlichen Skala aus der Berechnung des Modulwertes ausschließen. In den Modulen zu „Verhaltensweisen und psychischen Problemlagen“ sowie dem „Umgang mit krankheits-/therapiebedingten Anforderungen und Belastungen“ scheint dagegen die Perspektive zu überwiegen, die Bereiche umfassend abzubilden, in denen Menschen gesundheitsbedingt auf personelle Hilfe angewiesen sind. Werden demnach bestimmte Verhaltensweisen oder Anforderungen hinzugefügt oder gestrichen, verändert sich auch der Bedeutungsumfang der Fähigkeit, sie zu bewältigen. In den Modulen zu „Gestaltung des Alltagslebens und soziale Kontakte“ und einzelnen Bereichen der „Selbstversorgung“ bleibt dagegen offen, ob mit den Indikatoren aus Sicht der Autoren eine ursächliche Fähigkeit erhoben oder ein wesentlicher Bereich möglicher Pflegebedürftigkeit definiert werden soll. Wie die Diskussion um den Umfang der Pflegebedürftigkeit (Wingenfeld et al. 2007, S. 43–50, 59f, 108) und deren Definition (Wingenfeld et al. 2008, S. 75) zeigen, soll demgegenüber die Bedeutung von „Pflegebedürftigkeit“ über ihre Module festgelegt werden. Der modulare Aufbau des Instruments dient gerade dazu,

Pflegebedürftigkeit nach sozialpolitischen Entscheidungen definieren zu können, und impliziert damit ein formatives Modell.

Eng mit dem vorhergehenden Kriterium verbunden ist die Erwartung, dass die Indikatoren bzw. Dimensionen miteinander korrelieren. In reflektiven Modellen wird dies angenommen, in formativen können die Indikatoren oder Dimensionen miteinander korrelieren, müssen es aber nicht. Für die Indikatoren der Module „Mobilität“, „Kognitive und kommunikative Fähigkeiten“, „Gestaltung des Alltagslebens und soziale Kontakte“ sowie den bereits genannten Bereichen der „Selbstversorgung“ kann angenommen werden, dass sie miteinander korrelieren. Demgegenüber wird eine Korrelation der Indikatoren im „Umgang mit krankheits-/therapiebedingten Anforderungen und Belastungen“ ausdrücklich ausgeschlossen (Wingenfeld et al. 2008, S. 55). Aber auch von den im Modul 3 aufgeführten Verhaltensweisen muss nicht erwartet werden, dass sie miteinander korrelieren.

Auf Modulebene werden dagegen Ausprägungen in den Modulen „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ als voneinander unabhängig betrachtet, während von anderen Modulen erwartet wird, dass ihre Werte miteinander korrelieren (Wingenfeld et al. 2008, S. 103f.). Dies liegt daran, dass nach dem Pflegebedürftigkeitsbegriff *ursächliche Funktionen* für eine eingeschränkte Selbständigkeit im NBA auf einer Ebene mit ihren *Auswirkungen* auf Aktivitäten und Lebensbereiche eingeordnet werden. Während sie jedoch so in ihrer Eigenschaft als erklärende Variablen nicht spezifiziert werden, dienen sie in der Darstellung von Qualitätsindikatoren als Erklärungen zu erwartender Korrelationen (Wingenfeld et al. 2008, S. 103f.). Die Autoren unterscheiden dabei Beeinträchtigungen der motorischen Funktionen und kognitiven Fähigkeiten als Ursachen für eingeschränkte Selbständigkeit. Das NBA wäre danach mehrdimensional, ohne dass dies weiter spezifiziert ist.

Das letzte Kriterium zur Unterscheidung reflektiver und formativer Modelle betrifft das nomologische Netz, in dem sie eingebettet sind. Insofern reflektive Indikatoren oder Dimensionen demselben Konstrukt unterliegen und austauschbar sind, haben sie dieselben Voraussetzungen bzw. Konsequenzen. Konzeptionell wird dies bei den Indikatoren zu „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ sowie der „Gestaltung des Alltagslebens und soziale Kontakte“ und einzelnen Bereichen der „Selbstversorgung“ erwartet, sofern motorische und kognitive Fähigkeiten diese Merkmalsbereiche als ganze bestimmen. Den einzelnen Indikatoren der Module 3 und 5 können dagegen unterschiedliche Anforderungen oder Problemlagen zugrunde liegen. Auch auf Modulebene können für die einzelnen Bereiche unterschiedliche Voraussetzungen angenommen werden.

Dies stützt die Interpretation, im NBA „Pflegebedürftigkeit“ als formatives Konstrukt zu verstehen, dem Bereiche zugeordnet sind, die teilweise reflektiv, teilweise zumindest in einzelnen Subskalen formativ spezifiziert werden müssen. Als eindeutig reflektiv lassen sich aber nur die Messmodelle in den Modulen zu „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ einschätzen. Dies mag daran liegen, dass in anderen Modulen motorische und kognitive Fähigkeiten Indikatoren zugrunde liegen, die konzeptionell wesentliche Bereiche möglicher Pflegebedürftigkeit definieren. Damit überschneiden sich jedoch Perspektiven, die messtheoretisch getrennt werden müssen. Hier zeigt sich die Notwendigkeit, in der Entwicklung insbesondere komplexer Instrumenten die zugrunde liegenden Modelle eindeutig zu spezifizieren, um zu verdeutlichen, was wie gemessen werden soll.

ZUSAMMENFASSUNG

In der Entwicklung und Beurteilung der Güte von Assessmentinstrumenten müssen reflektive und formative Messmodelle voneinander unterschieden werden.

In reflektiven Messmodellen sind die beobachtbaren Variablen als Indikatoren des Konstrukts durch eine latente Variable bedingt, in formativen Messmodellen bestimmen dagegen die Indikatoren Inhalt und Ausmaß des Konstrukts.

Reflektive Messmodelle können mit Hilfe der geläufigen Verfahren zur Beurteilung der Reliabilität und Validität eingeschätzt werden. Formative Messmodelle entsprechen dagegen nicht den Grundannahmen der Klassischen Testtheorie, so dass sich die in diesem Rahmen entwickelten Methoden nicht auf die Entwicklung und Bewertung formativer Instrumente anwenden lassen. Die Verwendung gängiger Verfahren zur internen Konsistenz oder Konstruktvalidität auf formative Modelle führt so zu methodischen Mängeln in der Instrumentenentwicklung, der Beurteilung von Modellen und Einschätzung von Gütekriterien. Zu Beginn einer Instrumentenentwicklung sollten daher die Struktur eines Konstrukts und die Beziehungen zu seinen Indikatoren bestimmt werden. In der Literatur werden Kriterien vorgeschlagen, um reflektive und formative Modelle voneinander zu unterscheiden, doch erst die theoretische Ausarbeitung eines zu erfassenden Phänomens ermöglicht es, die Relationen eines Modells eindeutig zu spezifizieren, alternative Spezifikationen zu testen und das Modell in ein nomologisches Netz einzuordnen, um es zu validieren. Vor diesem Hintergrund lässt sich „Pflegebedürftigkeit“ im NBA als formatives Konstrukt verstehen, dem Bereiche zugeordnet sind, die teilweise reflektiv, teilweise zumindest in einzelnen Subskalen formativ spezifiziert werden müssen. Zentrale Bestandteile bilden dabei die Module „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“. Während sie in

ihren Ausprägungen als voneinander unabhängig betrachtet werden, wirken sich Beeinträchtigungen in diesen Funktionen aus Sicht der Entwickler in nahezu allen Lebensbereichen aus und sollen so auch den Zusammenhang zwischen den Ausprägungen anderer Module erklären. Beide Module lassen sich zudem als reflektive Messmodelle spezifizieren und erfüllen damit die testtheoretischen Voraussetzungen, um ihre Validität zu prüfen. Während daher das NBA erst auf seine Konstruktvalidität geprüft werden kann, wenn es theoretisch fundiert in ein nomologisches Netz eingebettet wird, lassen sich mit diesen beiden Modulen die wesentlichen Momente von Pflegebedürftigkeit auf ihre Validität untersuchen.

LITERATUR

Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; Weiber, Rolf (2011): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 13., überarb. Auflage. Berlin, Heidelberg: Springer

Bagozzi, Richard P. (1994): Structural Equation Models in Marketing Research. Basic Principles. In: Bagozzi, Richard P. (Hg.): Principles of Marketing Research. Cambridge, Mass.: Blackwell, S. 317–385

Bartholomeyczik, Sabine; Halek, Margareta (2009): Assessmentinstrumente in der Pflege. Möglichkeiten und Grenzen ; überarbeitete, erweiterte und ergänzte Beiträge einer Fachtagung zu diesem Thema am Institut für Pflegewissenschaft der Universität Witten/Herdecke in Zusammenarbeit mit der "Nationalen Pflegeassessmentgruppe Deutschland". 2., [aktualisierte], völlig überarb. Aufl. Hannover: Schlüter (Wittener Schriften)

Bollen, Kenneth; Lennox, Richard (1991): Conventional wisdom on measurement. A structural equation perspective. In: Psychological Bulletin, Jg. 110, H. 2, S. 305–314

Bollen, Kenneth A.; Lennox, Richard D.; Dahly, Darren L. (2009): Practical application of the vanishing tetrad test for causal indicator measurement models. An example from health-related quality of life. In: Statistics in medicine, Jg. 28, H. 10, S. 1524–1536

Bollen, Kenneth A.; Ting, Kwok-fai (1993): Confirmatory tetrad analysis. In: Sociological Methodology, Jg. 23, S. 147–175

Bollen, Kenneth A.; Ting, Kwok-fai (2000): A Tetrad Test for Causal Indicators. In: Psychological Methods, Jg. 5, H. 1, S. 3–22

Borsboom, Denny; Mellenbergh, Gideon J.; van Heerden, Jaap (2003): The Theoretical Status of Latent Variables. In: Psychological Review, Jg. 110, H. 2, S. 203–219

Bortz, Jürgen (2005): Statistik für Human- und Sozialwissenschaftler. 6. vollst. überarb. und aktu. Auflage. Heidelberg: Springer

Bühner, Markus (2006): Einführung in die Test- und Fragebogenkonstruktion. 2. aktual. u. erw. Auflage. München, Boston, San Francisco u.a.: Pearson Studium

Diamantopoulos, Adamantios; Winklhofer, Heidi M. (2001): Index construction with formative indicators. An alternative to scale developement. In: Journal of Marketing Research, Jg. 38, H. 2, S. 269–277

Edwards, Jeffrey R.; Bagozzi, Richard P. (2000): On the nature and direction of relationships between constructs and Measures. In: Psychological Methods, Jg. 5, H. 2, S. 155–174

Eggert, Andreas; Fassott, Georg (2003): Zur Verwendung formativer und reflektiver Indikatoren in Strukturgleichungsmodellen. Ergebnisse einer Metaanalyse und Anwendungsempfehlungen. Kaiserslautern (Kaiserslauterer Schriftenreihe Marketing, 20)

Fassott, Georg (2006): Operationalisierung latenter Variablen in Strukturgleichungsmodellen. Eine Standortbestimmung. In: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, Jg. 58, H. 1, S. 67–88

Fassott, Georg; Eggert, Andreas (2005): Zur Verwendung formativer und reflektiver Indikatoren in Strukturgleichungsmodellen. Bestandsaufnahme und Anwendungsempfehlungen. In: Bliemel, Friedhelm; Eggert, Andreas; Fassott, Georg; Henseler, Jörg (Hg.): Handbuch PLS-Pfadmodellierung. Methoden, Anwendung, Praxisbeispiel. Stuttgart: Schäfer-Poeschel, S. 31–47

Fayer, P. M.; Hand, D. J.; Bjordal, K.; Groenvold, M. (1997): Causal indicators in quality of life research. In: Quality of Life Research, Jg. 6, H. 5, S. 393–406

Fayers, Peter M.; Hand, D. J. (1997): Factor analysis, causal indicators and quality of life. In: Quality of Life Research, Jg. 6, H. 2, S. 139–150

Franken, Georg (2010): Konstruktvalidität der Subskala "Kognitive und kommunikative Fähigkeiten" des Neuen Begutachtungsassessments zur Feststellung von Pflegebedürftigkeit (NBA). Masterarbeit. Betreut von Prof. Dr. Albert Brühl. Vallendar. Philosophisch-Theologische Hochschule Vallendar, Pflegewissenschaftliche Fakultät. http://opus.bsz-bw.de/kidoks/volltexte/2012/66/ zuletzt geprüft am 20.08.2012

Giere, Jens; Wirtz, Bernd W.; Schilke, Oliver (2006): Mehrdimensionale Konstrukte. Konzeptionelle Grundlagen und Möglichkeiten ihrer Analyse mithilfe von Strukturgleichungsmodellen. In: Die Betriebswirtschaft, Jg. 66, H. 6, S. 678–694

Hribek, Günther; Schmalen, Helmut (2000): Konzeptualisierung und Operationalisierung der Patientenzufriedenheit mit stationärer Versorgung. Entwicklung multiattributiver Messinstrumente für Krankenhäuser und Rehabilitationseinrichtungen. In: Marketing ZFP, Jg. 22, H. 3, S. 208–225

Jarvis, Cheryl Burke; MacKenzie, Scott B.; Podsakoff, Philip M.; Mick, David Glen; Bearden, William O. (2003): A critical review of construct indicators and measurement model. Misspecification in marketing and consumer research. In: Journal of Consumer Research, Jg. 30, H. 2, S. 199–218

Jöreskog, Karl G.; Goldberger, A. S. (1975): Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. In: Journal of the American Statistical Association, Jg. 10, S. 631–639

Juniper, E. F.; Guyatt, G. H.; King, D. R. (1994): Comparison of methods for selecting items for a disease-specific quality-of-life questionnaire - importance versus factor-analysis. In: Quality of Life Research, Jg. 3, H. 1, S. 52f

MacKenzie, Scott B.; Podsakoff, Philip M.; Jarvis, Cheryl Burke (2005): The problem of measurement model misspecification in behavioural and organizational research and some recommended Solutions. In: Journal of Applied Psychology, Jg. 90, H. 4, S. 710–730

Streiner, David L. (2003): Being inconsistent about consistency. When Coefficient Alpha does and doesn't matter. In: Journal of Personality Assessment, Jg. 80, H. 3, S. 217–222

Wingenfeld, K.; Büscher, A.; Gansweid, B. (2008): Das neue Begutachtungsinstrument zur Feststellung von Pflegebedürftigkeit. Überarbeitete, korrigierte Fassung. Projekt: Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Abschlussbericht zur Hauptphase 1: Entwicklung eines neuen Begutachtungsinstruments. Institut für Pflegewissenschaft an der Universität Bielefeld (IPW). Medizinischer Dienst der Krankenversicherung Westfalen-Lippe (MDK-WL). Bielefeld, Münster. Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/ag6/downloads/Abschlussbericht_IPW_MDKWL_25.03.08.pdf, zuletzt geprüft am 16.04.2010

Wingenfeld, K.; Büscher, A.; Schaeffer, D. (2007): Recherche und Analyse von Pflegebedürftigkeitsbegriffen und Einschätzungsinstrumenten. Überarbeitete, korrigierte Fassung. Studie im Rahmen des Modellprogramms nach §8 Abs. 3 SGB XI. Im Auftrag der Spitzenverbände der Pflegekassen. Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/ag6/downloads/ipw_bericht_20070323.pdf, zuletzt geprüft am 16.04.2010

4. Konstruktvalidität der Subskala „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessments (NBA)

Georg Franken

Einleitung

Allgemein wird unter „Pflegebedürftigkeit“ ein komplexes Konstrukt verstanden, das Ursachen wie Bereiche umfasst, in denen Menschen aufgrund gesundheitlicher Beeinträchtigungen auf personelle Hilfe angewiesen sind (Abt-Zegelin 2000; Wingenfeld 2000; Werner 2004; Bartholomeyczik 2004; Braatz, Gansweid 2005; Hassler, Görres 2005; Menning, Hoffmann 2009). Nach der vom Institut für Pflegewissenschaft an der Universität Bielefeld durchgeführten Vorstudie zur Analyse und Bewertung von Pflegebedürftigkeitsbegriffen und Begutachtungs- bzw. Einschätzungsinstrumenten kann dieses Konstrukt so spezifiziert werden, dass es die „maßgeblichen Faktoren [erfasst], die das Ausmaß der Abhängigkeit von personeller Hilfe bzw. Art und Umfang der erforderlichen Hilfen bestimmen“ (Wingenfeld et al. 2007 S. 60). Zentrale Bestandteile bilden dabei im Neuen Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit (NBA) die Module „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ (Franken im vorliegenden Band).

Ziel und Fragestellungen

Mit dem NBA liegt aus Sicht seiner Autoren „ein ausgearbeitetes Verfahren vor, das für eine breite Erprobung und Testung seiner methodischen Güte bereit ist“ (Wingenfeld et al. 2008a, S. 129). Betrachtet man die Evaluation des NBA, soweit sie sich auf das zugrunde liegende Konstrukt und dessen Validität bezieht, so bleiben Fragen zum grundlegenden Skalenniveau und Konstrukt des Moduls „Kognitive und kommunikative Fähigkeiten“ offen (Franken 2010, S. 49–69). An der Pflegewissenschaftlichen Fakultät der Philosophisch-Theologischen Hochschule Vallendar wurde daher im Rahmen einer empirischen Studie die Konstruktvalidität der Subskala untersucht (Franken 2010). Forschungsleitend waren dabei die folgenden Fragestellungen:

1. Welche Dimensionen liegen der Subskala „Kognitive und kommunikative Fähigkeiten“ zugrunde?
2. Welches Modell kann für die Subskala identifiziert werden?

3. Bildet der im Rahmen der vorliegenden Bewertungssystematik ermittelte Wert der Subskala die empirischen Verhältnisse ab?

Vorgehen und Methode

Insofern unter Validität verstanden wird, dass ein Test misst, was er zu messen vorgibt, entspricht eigentlich nur Inhaltsvalidität dieser Definition (Murphy, Davidshofer 2005, S. 155; Bühner 2006, S. 36) und auch eine hinreichende Prüfung der Struktur eines Modells ersetzt noch nicht dessen angemessene inhaltliche Interpretation (Hartig et al. 2008, S. 154).

Umgekehrt verweist jedoch die Prüfung der Inhaltsvalidität auch auf die Validität des interessierenden Konstrukts und seine Einbettung in ein nomologisches Netz, denn um zu wissen, ob ein Instrument misst, was es messen soll, muss bekannt sein, was und wie es dies messen soll: Das Instrument muss hinsichtlich des zu messenden Konstrukts dimensional bestimmt und in den Beziehungen zu seinen Indikatoren und zu anderen Konstrukten spezifiziert sein.

Nachfolgend wird daher das Modul 2 (NBA) kurz dargestellt und mit den zu seiner Entwicklung herangezogenen Referenzinstrumenten verglichen, bevor Hypothesen zur inhaltlichen Struktur expliziert sowie die empirische Datenerhebung und Datenanalyse vorgestellt werden. Um dazu den methodischen Stand zur Überprüfung der Konstruktvalidität sowie des Skalenniveaus standardisierter Assessmentinstrumente zu bestimmen und Hypothesen zur Struktur der Subskala „Kognitive und kommunikative Fähigkeiten“ aus der Literatur herzuleiten, wurde in den Datenbanken Pubmed, Cinahl, Cochrane Library, CC Med, Psycinfo, Psyndex, der Deutschen Nationalbibliothek sowie den Landes- und Hochschulbibliotheken über die Bibliotheksverbünde SWB, BVB, HBZ, HEBIS, KOBV und GBV nach Veröffentlichungen in den Themenbereichen „Konstruktvalidität“, „Assessmentinstrumente“, „kognitive und kommunikative Einschränkungen“ sowie „Skalenniveau“ recherchiert.

Das Modul „Kognitive und kommunikative Fähigkeiten“ erhebt die Intensität der Beeinträchtigung geistiger Funktionen im Alltag bzw. das Ausmaß, in dem eine entsprechende Fähigkeit vorhanden ist. Inhaltlich gliedert sich das Modul in acht Items zu kognitiven und drei Items zu kommunikativen Fähigkeiten:

Nr.	Item	Variablenlabel*	vorhanden/ unbeeinträchtigt	überwiegend vorhanden	in geringem Maße vorhanden	nicht vorhanden
1	Personen aus dem näheren Umfeld erkennen	PERSONEN	☐ 0	☐ 1	☐ 2	☐ 3
2	Örtliche Orientierung	ORT	☐ 0	☐ 1	☐ 2	☐ 3
3	Zeitliche Orientierung	ZEIT	☐ 0	☐ 1	☐ 2	☐ 3
4	Gedächtnis	ERINNERN	☐ 0	☐ 1	☐ 2	☐ 3
5	Mehrschrittige Alltagshandlungen ausführen	HANDELN	☐ 0	☐ 1	☐ 2	☐ 3
6	Entscheidungen im Alltagsleben treffen	ENTSCH	☐ 0	☐ 1	☐ 2	☐ 3
7	Sachverhalte und Informationen verstehen	INFOS	☐ 0	☐ 1	☐ 2	☐ 3
8	Risiken und Gefahren erkennen	GEFAHREN	☐ 0	☐ 1	☐ 2	☐ 3
9	Mitteilung elementarer Bedürfnisse	MITTEILEN	☐ 0	☐ 1	☐ 2	☐ 3
10	Verstehen von Aufforderungen	AUFFORDERN	☐ 0	☐ 1	☐ 2	☐ 3
11	Beteiligung an einem Gespräch	GESPRÄCH	☐ 0	☐ 1	☐ 2	☐ 3

Tab. 4.1 Modul 2 „Kognitive und kommunikative Fähigkeiten“ des NBA (Wingenfeld et al. 2008a, S. 39f; *Ergänzung, GF)

Die Merkmale werden in einer vierstufigen Ratingskala erfasst, deren Ausprägungen Punktwerte von 0 bis 3 zugeordnet werden. Zur Berechnung des Modulwertes werden die Punktwerte der Merkmalsausprügungen zu kognitiven Items addiert und einer fünfstufigen Bewertungsskala zugeordnet, die die Beeinträchtigung bzw. die Ausprägung der Fähigkeit ausdrücken soll. (Wingenfeld et al. 2008a, S. 39–42).

Für die inhaltliche Ausarbeitung des Moduls wurden von den Autoren Referenzinstrumente hinzugezogen.[47] Im Vergleich zu diesen Instrumenten stellt das

[47] Allgemein wird für die inhaltliche Ausarbeitung der Module darauf verwiesen, dass die als Referenz empfohlenen Instrumente FACE, RAI 2.0, RAI HC, EASY-Care, ABV sowie die eingeschränkt empfohlenen Instrumente CANE, RCN Assessment und RUM daraufhin befragt wurden, „wie sie die

Modul eine umfassende Erhebung der kognitiven und kommunikativen Leistungsfähigkeit älterer Menschen dar. Kategoriale Überschneidungen zwischen einzelnen kognitiven Fähigkeiten in den verwendeten Instrumenten können als Ausdruck konzeptioneller Unterschiede oder als Hinweis auf unterschiedliche Schwierigkeitsgrade der erhobenen Merkmale verstanden werden. Kommunikative Fähigkeiten werden dabei ausser in der Alzheimer's Disease Assessment Scale" (ADAS) in allen Instrumenten konzeptionell als eigenständige Dimension betrachtet.

Um die zugrunde liegenden Konzepte kognitiver Assessmentinstrumente zu prüfen, wird in der Literatur wie auch in der Evaluation des Moduls 2 "Kognitive und kommunikative Fähigkeiten" (NBA) häufig der Ansatz einer kriterienbezogenen Validierung gewählt. Kennziffern für die Kriteriumsvalidität wie Sensitivität oder Likelihood Ratio beziehen sich jedoch auf Gesamtskalen und bieten bei komplexen Skalen mit unterschiedlichen Facetten wie in neuropsychologischen Tests keine Hinweise auf Stärken und Grenzen eines Instruments (McDowell 2006, S. 31–34). Um detaillierte Einsichten von komplexen Instrumenten, für die es kein einzelnes Außenkriterium als Standard gibt, zu erhalten, werden Verfahren der Konstruktvalidierung verwendet.

Am Beginn einer Konstruktvalidierung steht die konzeptionelle Definition des gemessenen Konstrukts, der internen Struktur seiner Komponenten und seines Verhältnisses zu anderen Konstrukten. Hinsichtlich des hier untersuchten Moduls 2 (NBA) betrifft dies das Verständnis von Kognition, die Struktur kognitiver Fähigkeiten und ihr Verhältnis zur Kommunikation sowie die Beziehung der erhobenen Fähigkeiten und deren Beeinträchtigungen zur Selbständigkeit einer Person.

Im NBA wird das zugrundeliegende Verständnis von Kognition nur für die Kinderbegutachtung hinsichtlich der Altersgrenzen kognitiver Entwicklung expliziert und mit Intelligenz gleichgesetzt (Wingenfeld et al. 2008b, E-11). Kognition kennzeichnet danach „die Fähigkeit, sich in der gegebenen Umwelt zu behaupten, Informationen miteinander zu verknüpfen, aufgrund von Gedächtnisleistungen Erfahrungen zu sammeln, Handlungen zu planen und Entscheidungen zu treffen sowie diese dann auszuführen" (Neuhäuser 2004 zitiert in Wingenfeld et al. 2008b, a. a. O.). Während aber Intelligenztests in den genannten Bereichen die allgemeine Leistungsfähigkeit ermitteln, suchen die Testautoren des NBA im Modul 2 die Fähigkeiten zu erfassen, deren Einschränkungen die Selbständigkeit beeinträchtigen (Wingenfeld et al. 2008a,

---

verschiedenen Aspekte der Pflegebedürftigkeit zum Zweck des Assessments operationalisieren" (Wingenfeld et al. 2008a, S. 10). Zusätzlich wurden für die Einschätzung spezifischer Aspekte der Pflegebedürftigkeit weitere Instrumente herangezogen. Im Anhang zur „Recherche und Analyse von Pflegebedürftigkeitsbegriffen und Einschätzungsinstrumenten" wird dazu hinsichtlich kognitiver Fähigkeiten ausdrücklich jedoch nur die ADAS empfohlen. Um Art, Umfang und Geltungsbereich der Subskala „Kognitive und kommunikative Fähigkeiten" (NBA) einschätzen zu können, wurden die entsprechenden Bereiche der genannten Instrumente verglichen. Das „Alternative Begutachtungsverfahren" der MDK-Gemeinschaft konnte dazu nicht recherchiert werden (Franken 2010, S. 75–83).

S. 38). Die Merkmale werden dabei operational definiert und orientieren sich an neuropsychologischen Tests und Assessmentinstrumenten, die kognitive Dysfunktionen oder Schädigungen und damit Demenz als äußerste Form kognitiven Abbaus erfassen und sich an deren Symptomatik ausrichten (Huppert, Tym 1986; Colsher, Wallace 1991, S. 3ff; Barrie 2002; McDowell 2006, S. 395). Inhaltlich unterscheiden die Autoren des NBA im Modul 2 zwischen Items zur Kognition und zur Kommunikation (Wingenfeld et al. 2008a, S. 39f.).

Die Berechnung eines Gesamtscores für das Modul setzt allerdings voraus, dass die Skala trotz ihrer vielfältigen Facetten eindimensional sein sollte.

Die Merkmale zur Kommunikation werden dabei bezogen auf das Konstrukt der kognitiven Items als redundant aufgefasst (Wingenfeld et al. 2008a, S. 41). Vor dem Hintergrund der Literatur ließen sich sowohl eine Differenzierung der Items zu Orientierung (Item 1-3)/Gedächtnis (Item 4) von Items zur Praxis (Item 5-8)/Sprache (Item 9-11) wie auch eine weitere Ausdifferenzierung der Items zur Kommunikation erwarten. Die Dimensionalität einer Skala hängt jedoch auch von der Komplexität der Items, ihren Schwierigkeitsgraden und der Art der Erhebung ab, so dass auch eine eindimensionale Lösung möglich ist (Franken 2010, S. 85–96). Konzeptionell fügt sich das Konstrukt des Moduls 2 (NBA) in ein nomologisches Netz ein, das den Zusammenhang zwischen kognitiven und kommunikativen Fähigkeiten und der Selbständigkeit einer Person erklären soll. Da die vorliegende Studie nicht das NBA als Ganzes zum Gegenstand hat, bleiben jedoch Fragen zur Relevanz der erhobenen kognitiven Merkmale für die Selbständigkeit einer Person der weiteren Forschung vorbehalten.

In der Überprüfung kognitiver Assessments werden als Belege für interne Strukturen hauptsächlich Zusammenhänge zwischen den Items in Form einer Faktorenanalyse untersucht (McDowell 2006). Ziel ist es, aus beobachtbaren Variablen Gruppen zu bilden, die ein gemeinsames und von anderen Gruppen unterscheidbares Thema erfassen. So kann untersucht werden, ob alle Indikatoren in angenommene Gruppen fallen bzw. welche Items ein gemeinsames Thema und damit eine spezifische Dimension erfassen, deren Merkmalsausprägungen in einem separaten Wert berechnet werden sollten.

Als Faktorenanalysen werden multivariate Analyseverfahren bezeichnet, in denen zu einer Menge manifester Variablen eine weniger umfängliche Menge latenter Variablen gesucht wird, die den Zusammenhang zwischen den manifesten Variablen erklärt. Prinzipiell lassen sich eine exploratorische, hypothesengenerierende und eine konfirmatorische, hypothesenprüfende Faktorenanalyse unterscheiden (Bühner 2006,

S. 179–298; Moosbrugger, Schermelleh-Engel 2008, S. 307–324). Eine exploratorische Faktorenanalyse (EFA) wird gewählt, wenn keine konkrete Hypothese über die Anzahl und Zuordnung der Items zu latenten Variablen, den Faktoren einer Skala, besteht. Eine konfirmatorische Faktorenanalyse (CFA) prüft dagegen ein hypothetisch unterstelltes Modell zu Anzahl und Beziehung der Faktoren und beobachteten Variablen auf seine Gültigkeit. Insofern lassen sie sich als struktur-suchende und struktur-prüfende Verfahren voneinander unterscheiden. Eine CFA lässt sich aber auch „explorativ“ anwenden, indem schrittweise alternative Modelle getestet und modifiziert werden, um sie der Struktur der Daten anzupassen. Damit unterliegt das Forschungsvorhaben methodisch einem Prozess, der theoriegeleitet wie datenorientiert eine Analyse der Daten ermöglicht (Borg, Staufenbiel 2007, S. 239, vgl. Jöreskog, Sörbom 1993).

Die klassische Faktorenanalyse setzt dabei voraus, dass die manifesten wie latenten Variablen intervallskaliert sind (Jöreskog, Moustaki 2006, S. 1). Um die Voraussetzungen einer Faktorenanalyse hinsichtlich Verteilung und Skalenniveau explizit zu gewährleisten (McDowell 2006, S. 37), werden in einzelnen Studien zu kognitiven Assessementinstrumenten polychorische Korrelationen berechnet (Braekhus et al. 1992; Abraham et al. 1994; Jones, Gallo 2000). Dabei werden ordinale Daten als ungenaue Erfassung einer zugrunde liegenden latenten kontinuierlichen Variablen verstanden und unter Annahme einer Normalverteilung Schwellenwerte als Grenze zwischen den kategorialen Ausprägungen der ordinalen Messungen berechnet. Bei der polychorischen Korrelation zu zwei ordinalen Variablen z_1 und z_2 handelt es sich um Schätzungen innerhalb der bivariaten Normalverteilung der zugrunde gelegten kontinuierlichen Variablen z_1* und z_2* (Jöreskog 2002). Zu diesen latenten kontinuierlichen Variablen lassen sich auch sogenannte Normal Scores als Mittelwerte der Integrale zwischen den Schwellenwerten einer einzelnen Variablen berechnen, die zur deskriptiven Beurteilung der Abstände zwischen Antwortkategorien verwendet werden können (Baltes-Götz 1994,. 4-2f; Schröder 2010, S. 63). Als Basis für Faktorenanalysen liefern jedoch polychorische Korrelationen genauere Schätzergebnisse (Jöreskog, Sörbom 2003).

An der Philosophisch-Theologischen Hochschule Vallendar wurden 2009 bis 2011 vier empirische Studien zur Untersuchung des NBA durchgeführt, denen eine gemeinsame Datenerhebung zugrunde liegt (Bensch 2012, in diesem Band, S. 118). Die Gesamtstichprobe wurde im Sinne einer Querschnittstudie ausgewertet. Die Daten für die vorliegende Studie wurden von Dezember 2009 bis August 2010 bei Klienten ambulanter Pflegedienste erhoben. Die Daten wurden automatisch mit der Software Remark Office OMR erfasst. Für die deskriptive Datenanalyse und die Teststatistiken

wurde SPSS 11.5 verwendet. Die Faktorenanalyse wurde mit LISREL 8.8 for Windows (Student Edition) und PRELIS 2.8 durchgeführt.

ERGEBNISSE

Stichprobe

Für die vorliegende Studie wurden 1816 vollständige Datensätze ausgewertet.[48] Die Erhebung kann im Unterschied zur Evaluationsstudie keine Repräsentativität beanspruchen. Die Altersverteilung der Untersuchungsgruppe weicht denn auch signifikant von der in der Studienpopulation der Evaluationsstudie ab ($\chi^2 = 56{,}12$; df = 7; $p < 0.001$). Keine signifikanten Unterschiede ergeben sich hinsichtlich der Verteilung der Pflegestufen ($\chi^2 = 1{,}92$, df = 3, $p > 0.20$) und Modulwertungen ($\chi^2 = 5{,}88$, df = 4, $p > 0.20$) (Franken 2010, S. 245f).

Konstruktvalidität der Subskala „Kognitive und kommunikative Fähigkeiten“

Um die Dimensionalität des Moduls 2 „Kognitive und kommunikative Fähigkeiten“ (NBA) faktorenanalytisch umfassend zu untersuchen und ein Modell für diese Subskala zu identifizieren, wurde eine mehrschrittige Vorgehensweise gewählt. Grundlage aller durchgeführten Untersuchungen ist die Berechnung polychorischer Korrelationen für die zu analysierenden Matrizen.[49]

Verbindung von exploratorischer und konfirmatorischer Faktorenanalyse

Für die Analyse der Dimensionalität der Skala und der Spezifikation eines Modells wurde zunächst ein zweistufiges Verfahren gewählt.

In einer exploratorischen Faktorenanalyse (EFA) sollte untersucht werden, wie sich die Variablen strukturieren und welche Faktoren sich dabei inhaltlich herausbilden. In einer daran anschließenden konfirmatorischen Faktorenanalyse (CFA) sollte an einer neuen Stichprobe ein entsprechend spezifiziertes Modell auf seine Güte getestet werden.

[48] Zur Beschreibung der Stichprobe siehe Franken 2010, S. 121-124; 224-243 und vgl. Bensch 2012, in diesem Band, S. 133; vgl. Bensch 2011).

[49] Bei der Schätzung der polychorischen Korrelationskoeffizienten durch den LR-χ^2 -Test wird die Annahme einer bivariaten Normalverteilung der zugrunde gelegten latenten kontinuierlichen Variablen auf einem 5%-Niveau für nahezu alle Variablenpaare zurückgewiesen (Franken 2010, S. 249–277). Da der Root Mean Square Error of Approximation (RMSEA) in der Stichprobe sowie den Teilstichproben jedoch durchgehend kleiner 0.1 ist, können die polychorischen Korrelationen trotz Verletzung der Normalverteilung für die weiteren Analysen verwendet werden (Jöreskog 2002, S. 18). Zu beachten ist dabei jedoch, dass die Variablen sehr hoch miteinander korrelieren. Bühner weist darauf hin, dass hohe Korrelationen ($r > .85$) zu Schätzproblemen führen können (Bühner 2006, S. 262).

Für die Analyse wurde die Gesamtstichprobe in zwei Zufallsstichproben mit n = 600 bzw. n = 1216 aufgeteilt. Die Größe der Stichproben erfüllen die in der Literatur genannten Ansprüche für die einzelnen Verfahren (Bühner 2006, S. 193; Schermelleh-Engel et al. 2003, S. 48–51).

Die nachfolgende Tabelle listet die Faktorladungen der einzelnen Items sowie deren Einzigartigkeiten [50] auf. Die Einzigartigkeit wird berechnet aus 1 minus der Kommunalität, also der durch den Faktor aufgeklärten Varianz des Items. Sie setzt sich zusammen aus der Spezifität[51] und dem Messfehler eines Items.

Factor Loadings

	Factor 1	**Unique Var**
PERSONEN	0.91	0.18
ORT	0.96	0.08
ZEIT	0.95	0.09
ERINNERN	0.95	0.10
HANDELN	0.91	0.17
ENTSCH	0.96	0.09
INFOS	0.95	0.10
GEFAHR	0.95	0.10
MITTEI	0.94	0.11
AUFFO	0.95	0.10
GESPR	0.94	0.12

Tab. 4.2 Faktorenanalyse in Teilstichprobe 1, Schätzverfahren: MINRES, Abbruchkriterium: Eigenwert > 1.

Im Ergebnis laden die Variablen hoch und bis auf die Items „Personen (erkennen)“ und „(Mehrschrittiges) Handeln“ gleichmäßig auf einen Faktor. Die durch den Faktor

[50] Als Einzigartigkeit wird die Varianz eines Items bezeichnet, die dieses Item mit keinem anderen Item teilt (Bühner 2006, S. 188).

[51] Als Spezifität wird die systematische Varianz eines Items bezeichnet, die nicht durch die extrahierten Faktoren erklärt wird (Bühner 2006, S. 187).

aufgeklärte Varianz liegt dabei zwischen .82 und .92. Nach diesem Ergebnis ist die Subskala „Kognitive und kommunikative Fähigkeiten“ (NBA) eindimensional.

Um die Eindimensionalität des Moduls 2 (NBA) zu testen, wurde für die konfirmatorische Faktorenanalyse ein Messmodell spezifiziert, in dem eine latente Variable „Kognition“ die Varianz der Indikatoren PERSONEN, ORT, ZEIT, ERINNERN, HANDELN, ENTSCHEIDEN, INFOS, GEFAHREN, MITTEILEN, AUFFORDERN und GESPRÄCH erklärt. Als Verfahren, um die Parameter des Modells (Ladungen, Korrelationen oder Kovarianzen, Fehlervarianzen) zu schätzen, wurde WLS (Weighted Least Squares) gewählt.[52] Die folgende Abbildung zeigt das Pfaddiagramm mit den geschätzten Ladungen und Fehlervariablen. Die Indikatoren laden in diesem Modell sehr hoch und gleichmäßig auf die latente Variable „Kognition“.

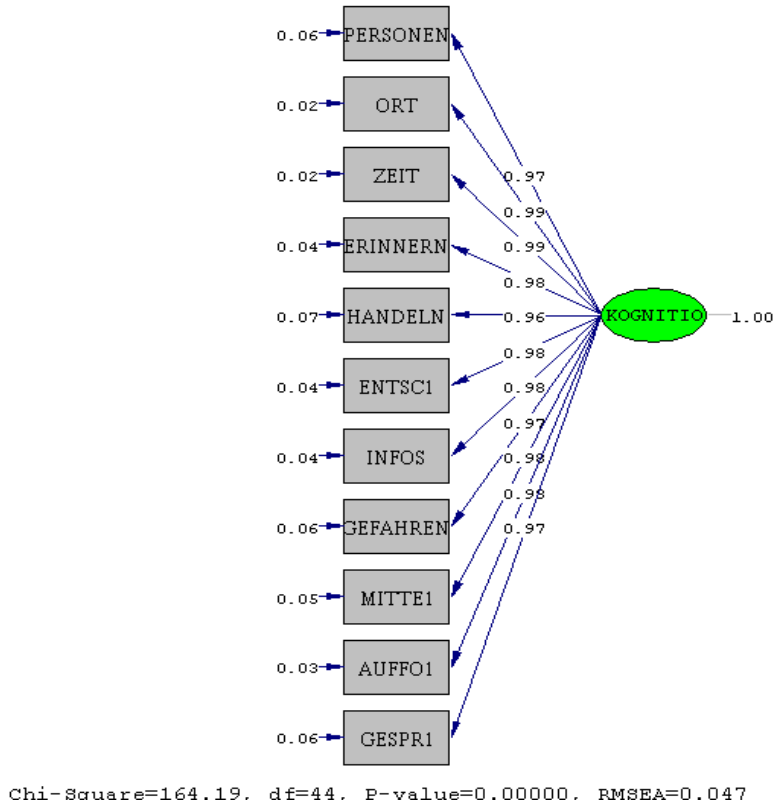


Abb. 4.1 Pfaddiagramm des eindimensionalen Modells „Kognition“ (Teilstichprobe 2)

Der χ^2 -Test weist das Modell für einen exakten Modell-Fit zurück (χ^2 (df = 44): 164.19; p = 0.00). Auch das nach Schermelleh-Engel et al., 2003 für einen approximativen

[52] In einer CFA werden ausgehend von vorläufigen Startwerten für die freien Parameter eines Modells iterativ eine damit implizierte Matrix der Kovarianzen zwischen den Items berechnet und der beobachteten Kovarianzmatrix angenähert. Es gibt dazu verschiedene Schätzmethoden. WLS setzt bei hinreichend großer Stichprobe keine Annahmen zur Verteilung der Daten voraus (Baltes-Götz 1994, S. 64f).

Modell-Fit akzeptable Verhältnis des χ^2 -Werts zur Anzahl der Freiheitsgrade ($\chi^2 \leq 3df$) wird nicht erreicht (Schermelleh-Engel et al. 2003, S. 52). Dies ließe sich allerdings mit der Verletzung der Normalverteilung und der Stichprobengröße erklären (Schermelleh-Engel et al. 2003, S. 32; Jöreskog 2002, S. 22). Weitere Tests zur Modellgüte ergeben dagegen einen akzeptablen bis guten Fit.[53]

Bei unzureichender Modellgüte sollten die standardisierten Residuen herangezogen werden, um die mögliche Ursache für einen mangelnden Modell-Fit zu suchen (Jöreskog, Sörbom 1993, S. 126; vgl. Brühl 2012, in diesem Band, S. 36). Die folgende Tabelle zeigt die standardisierten Residuen zum untersuchten Modell:

	PERS	**ORT**	**ZEIT**	**ERIN**	**HAN**	**ENTS**	**INFO**	**GEFA**	**MITT**	**AUFF**	**GESP**
PERSON	- -										
ORT	-6.48	- -									
ZEIT	-6.63	-4.54	- -								
ERINNERN	-5.17	-5.45	-4.43	- -							
HANDELN	-8.33	-8.55	-8.85	-8.59	- -						
ENTSCH	-7.45	-8.19	-8.30	-7.90	-4.79	- -					
INFOS	-7.31	-7.99	-7.74	-7.30	-7.63	-5.47	- -				
GEFAHR	-6.37	-6.51	-7.45	-7.45	-7.90	-5.28	-4,78	- -			
MITTE1	-8.21	-8.14	-8.17	-8.15	-7.87	-7.06	-6.14	-5.80	- -		
AUFFO	-7.73	-8.15	-8.09	-8.66	-7.65	-7.34	-5.73	-5.49	-4,45	- -	
GESPR	-7.20	-7.95	-7.68	-7.83	-8.17	-6.37	-6.59	-6.40	-6.19	-4.14	- -

Tab. 4.3 Standardisierte Residuen zum eindimensionalen Modell „Kognition“

In den standardisierten Residuen zeigen sich hohe negative Abweichungen insbesondere bei den Kovarianzen der Indikatoren HANDELN bis GESPRÄCH zu Indikatoren zu Orientierung und Gedächtnis sowie den Indikatoren INFOS bis

[53] Der RMSEA beträgt 0.047 (p-Wert für RMSEA < .05 = .7). Der RMSEA entspricht so nach Schermelleh-Engel et al., 2003 einem guten bis akzeptablen Fit. Der Standardized Root Mean Square Residual (SRMR) beträgt .06 und wäre nach Schermelleh-Engel et al., 2003 ebenfalls noch akzeptabel. Die Fit-Indizes Normed Fit Index (NFI), Nonnormed Fit Index (NNFI), Comparative Fit Index (CFI), Goodness-of-Fit Index (GFI) und Adjusted Goodness-of-Fit Index (AGFI) betragen 1.00 (Schermelleh-Engel et al. 2003, S. 52; zur Erläuterung der Fit-Indizes s. u. Anm. 12).

GESPRÄCH auf HANDELN. Danach überschätzt das Modell die Kovarianzen zwischen den Variablen. Dies verweist auf eine Fehlspezifikation (Jöreskog, Sörbom 1993, S. 126f.).

Konfirmatorische Faktorenanalyse zum Test alternativer Modelle

Für die weitere Analyse wurden im Rahmen einer konfirmatorischen Faktorenanalyse alternative Modelle zur Spezifikation der Skala verglichen. Die Autoren des NBA differenzieren hinsichtlich der internen Struktur des Moduls 2 (NBA) zwischen Kognition und Kommunikation (Abb. 4.3) (Wingenfeld et al. 2008a, S. 39f.). Die Berechnung des Modulwertes ausschließlich aus den Merkmalsausprägungen der ersten acht Items setzt jedoch voraus, dass die Skala eindimensional ist (Abb. 4.2) und die Merkmale zur Kommunikation bezogen auf das Konstrukt redundant sind (Wingenfeld et al. 2008a, S. 41). Vor dem Hintergrund der Literatur können zu der Skala theoriegeleitet drei hypothetische Modelle gebildet werden. Neben einem ebenfalls eindimensionalen Modell auch ein Modell mit zwei latenten Variablen („Orientierung/Gedächtnis“ und „Sprache/Praxis“) (Abb. 4.4), bzw. einem dreidimensionalen Modell, das neben „Orientierung/Gedächtnis“ die beiden Dimensionen „Praxis“ und „Sprache“ differenziert (Abb. 4.5). Die verschiedenen Modelle wurden in der Gesamtstichprobe auf ihre Modellgüte getestet. Dabei wurde rein konfirmatorisch vorgegangen. Die in LISREL vorgeschlagenen Modifikationen[54] wurden daher nicht berücksichtigt. Die folgenden Abbildungen zeigen die Pfaddiagramme der einzelnen Modelle mit den geschätzten Ladungen und den Fehlervariablen.

Die Modelle mit unterschiedlicher Dimensionalität weisen dieselbe Modellstruktur auf und unterscheiden sich nur darin, dass einzelne Parameter zusätzlich fixiert oder freigesetzt werden. Sie lassen sich auf diese Weise „hierarchisch schachteln“ [55] (Moosbrugger, Schermelleh-Engel 2008, S. 316). Modell 1 ist dabei ein Untermodell von Modell 2a wie auch von Modell 2b. So lässt sich Modell 1 aus diesen Modellen ableiten, indem ein Parameter fixiert wird, d. h. die Korrelation zwischen den Faktoren in den Modellen 2a bzw. 2b auf den Wert 1 gesetzt wird. Die Modelle 2a und 2b sind wiederum Untermodelle von Modell 3 (Abb. 4.6).

[54] Lisrel bietet für jeden fixierten oder restringierten Parameter einen Modifikationsindex an, der angibt, um welchen Wert sich das Ergebnis des $\chi 2$-Tests verringert, wenn der Parameter freigesetzt und anschließend das Modell neu geschätzt wird (Jöreskog, Sörboom 1993, S. 147).

[55] Zwei Modelle weisen dieselbe Modell- oder faktorielle Struktur auf, unterscheiden sich aber darin, dass in einem Modell beispielsweise die Korrelation zwischen latenten Variablen auf 1 fixiert wird, so dass diese Variablen zusammenfallen. Dieses Modell ist damit eine begrenzte Version oder ein Untermodell des anderen Modells, da es aus dem weniger begrenzten Modell abgeleitet werden kann.

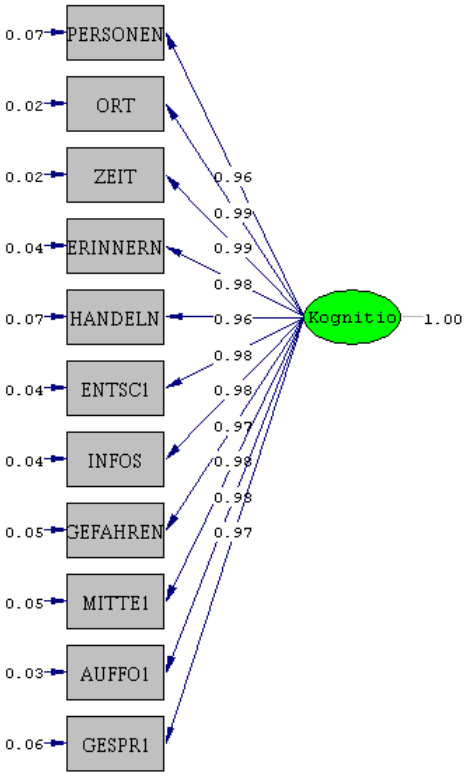


Abb. 4.2 Pfaddiagramm des Modells 1 „Kognition“ (Gesamtstichprobe)

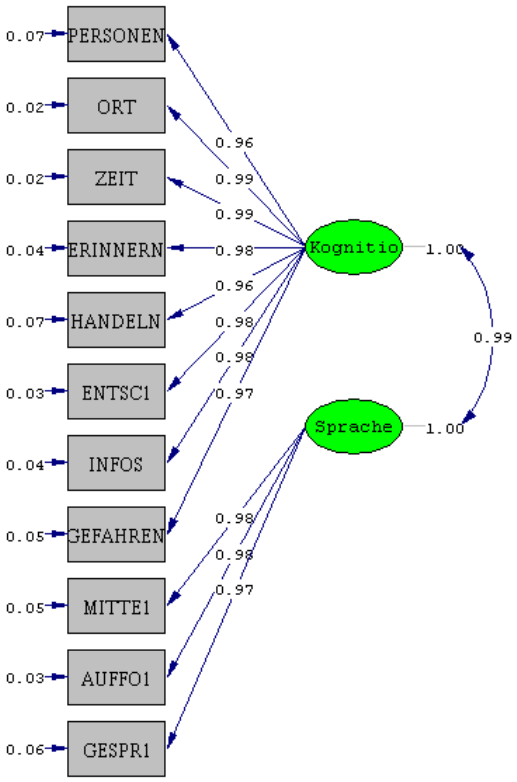


Abb. 4.3 Pfaddiagramm des Modells 2a „Kognition“, „Sprache“

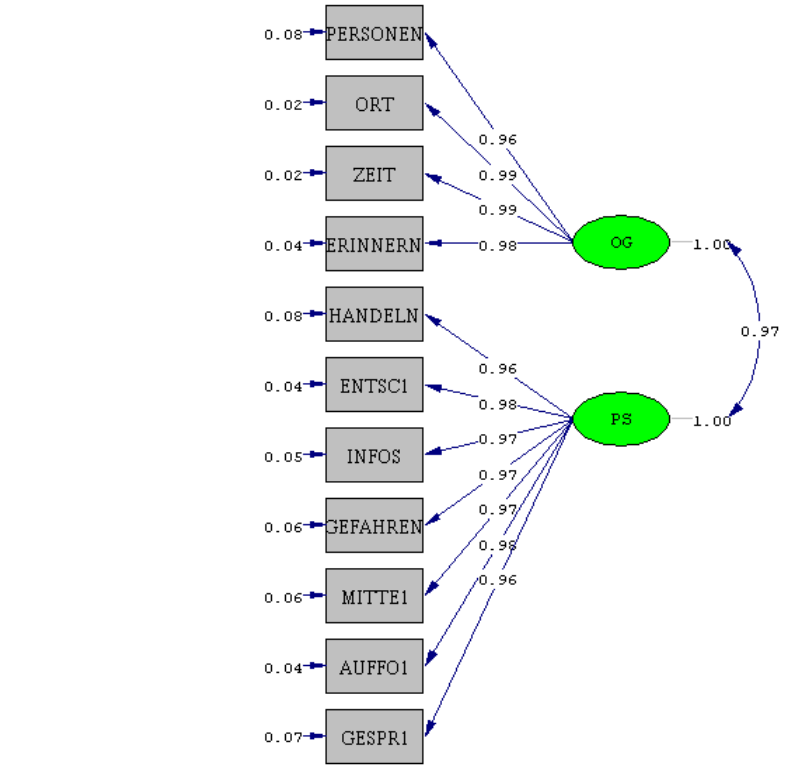


Abb. 4.4 Pfaddiagramm des Modells 2b „Orientierung/Gedächtnis“, „Praxis/Sprache“

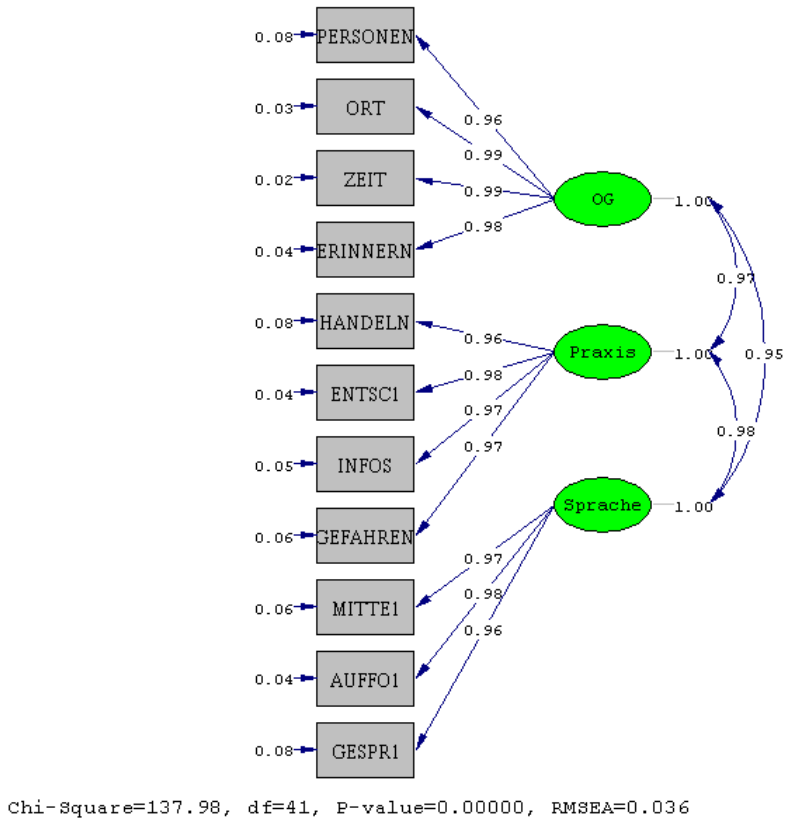


Abb. 4.5 Pfaddiagramm des Modells 3 „Orientierung/Gedächtnis“, „Praxis“, „Sprache“

Insofern die Modelle ineinander geschachtelt sind, ist die Differenz der χ^2-Werte selber wieder χ^2 verteilt mit df Freiheitsgraden (hier df = Differenz der Freiheitsgrade geschachtelter Modelle) (Schermelleh-Engel et al. 2003, S. 33f.).

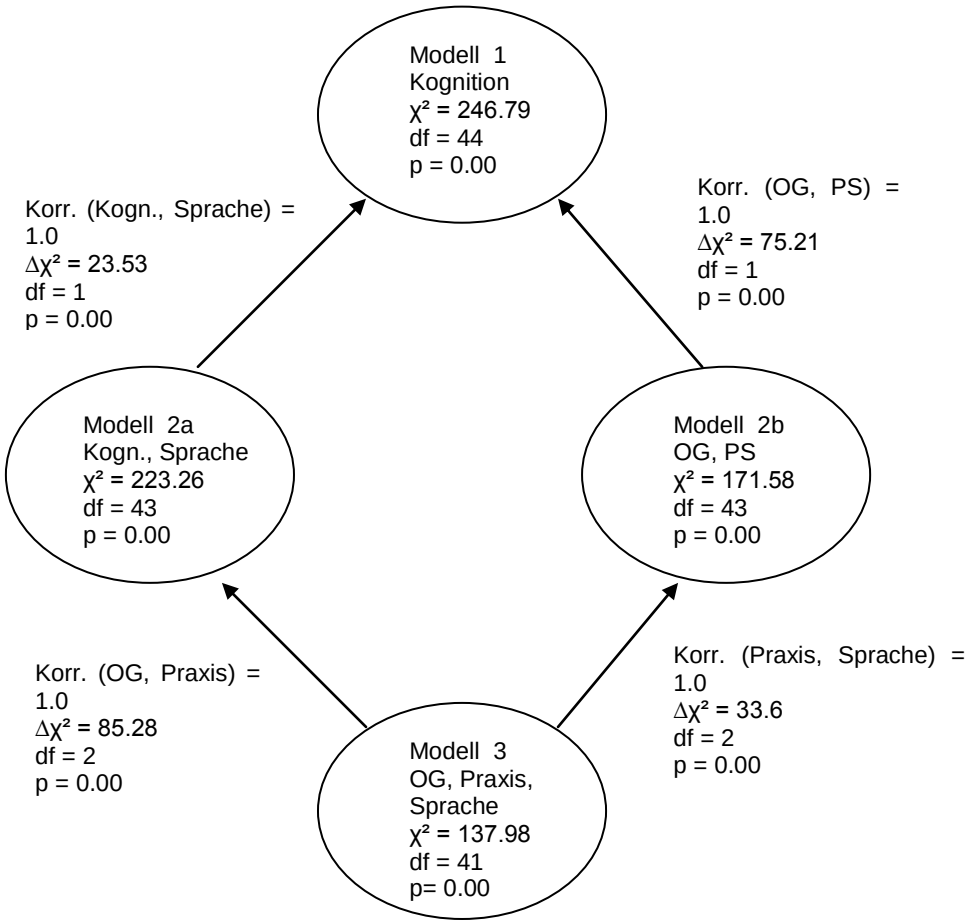


Abb. 4.6 Hierarchie der Modelle in der CFA alternativer Modelle

Das eindimensionale Modell 1 „Kognition“ (Abb. 4.2) entspricht dabei dem im zweistufigen Verfahren spezifizierten einfachen Messmodell (Abb. 4.1). Das Ergebnis der Berechnungen verändert sich in der Gesamtstichprobe nur geringfügig (Franken 2010, S. 136f, 291-297). Bei den Indizes zur Modellgüte ist jedoch aufgrund der größeren Stichprobe insbesondere der χ^2-Test für Modellabweichungen sensitiver (Tab. 4.4).

Wie in Modell 1 laden auch in den Modellen 2a (Abb. 4.3) und 2b (Abb. 4.4) die Indikatoren hoch auf die latenten Variablen. Abweichungen in der Höhe im Vergleich zu den übrigen Indikatoren der jeweiligen latenten Variablen finden sich insbesondere bei den Indikatoren PERSONEN, aber auch bei HANDELN und GESPRÄCH. Die latenten Variablen korrelieren sehr hoch. Der χ^2-Wert der Modelle 2a und 2b ist jeweils

signifikant besser als der von Modell 1 (Abb. 4.6). Trotzdem weist der χ^2-Test die Modelle immer noch zurück. Auch die übrigen Fit-Indizes verbessern sich gegenüber dem Modell 1 (Tab. 4.4). Allerdings weisen die Modellfits darauf hin, dass das Modell 2b besser zu den Daten passt als Modell 2a.[56]

Im Modell 3 verändern sich die Ladungen im Vergleich zu den vorher getesteten Modellen nicht substantiell (Abb. 4.5). Die latenten Variablen korrelieren wieder sehr hoch. Hinsichtlich der Modellgüte ist der χ^2-Wert von Modell 3 ebenfalls signifikant besser als der von Modell 2a bzw. 2b (Abb. 4.6), bleibt aber so hoch, dass auch dieses Modell nach dem χ^2-Test abgelehnt wird. Auch die übrigen Fit-Indizes verbessern sich leicht gegenüber dem Modell 2b (Tab. 4.4).

Die negativen standardisierten Residuen gleichen sich in den verschiedenen Modellen einander an (Franken 2010, S. 295, 302, 309). Sie bleiben aber hoch, wonach die modellimplizite Matrix generell höhere Kovarianzen enthält als die empirische und die Faktorladungen allgemein überschätzt werden.

	Modell 1	**Modell 2a**	**Modell 2b**	**Modell 3**
χ^2	246.79 (df =	223.26 (df =	171.58 (df = 43)	137.98 (df = 41)
p-Wert	0.00	0.00	0.00	0.00
RMSEA	.050	.048	.041	.036
CI (RMSEA)	.044; .057	.042; .054	.034; .047	.030; .043
p-Wert für	.45	.68	0.99	1.00
SRMR	.061	.055	.04	.032
NFI	1.00	1.00	1.00	1.00
NNFI	1.00	1.00	1.00	1.00
CFI	1.00	1.00	1.00	1.00
GFI	1.00	1.00	1.00	1.00
AGFI	1.00	1.00	1.00	1.00
AIC	290.79	269.26	217.58	187.98
ECVI	0.16	0.15	0.12	0.10
CI (ECVI)	0.14; 0.19	0.12; 0.18	0.100; 0.14	0.086; 0.13

Tab. 4.4 Fit-Indizes der Modelle im konfirmatorischen Modellvergleich[57]

[56] So sinken im Modell 2b RMSEA wie SRMR unter .05 und sind Akaike information criterion (AIC) bzw. Expected Cross-Validation Index (ECVI) geringer als im Modell 2a. Tatsächlich lassen sich diese Indizes in LISREL ineinander umformen, indem der ECVI mit (N – 1) multipliziert wird. Schermelleh-Engel et al. empfehlen daher eines der beiden Kriterien zu verwenden (Schermelleh-Engel et al. 2003, S. 48). Das Konfidenzintervall beim ECVI gibt dabei die Genauigkeit der Schätzung an. Erkennbar ist, dass sich im Vergleich zum Modell 1 erst beim Modell 2b die Intervalle nicht mehr überschneiden.

[57] Als Fit wird die Einschätzung verstanden, wie gut ein a-priori spezifiziertes Modell zu den Daten passt. So wird mit einem χ^2-Test geprüft, ob das postulierte Modell zur Datenstruktur passt. Da der χ^2-Test jedoch stichprobenabhängig ist, werden weitere Indizes zur Beurteilung der Modellgüte herangezogen. Absolute Fit-Indizes (GFI, AGFI, RMSEA, SRMR) vergleichen dazu das a-priori spezifizierte Modell mit einem saturierten Modell, das die Stichprobenvarianz exakt repliziert. Sie geben an, wie gut oder schlecht ein Modell die Daten beschreibt. Inkrementelle Fit-Indizes (NFI, NNFI, CFI) vergleichen das zu prüfende Modell mit einem Nullmodell, in dem alle Parameter auf null fixiert sind und die Variablen somit nicht korrelieren. Sie zeigen an, inwieweit sich ein Modell besser an die Daten anpasst als das stärker

Exploratives Vorgehen im Rahmen einer konfirmatorischen Faktorenanalyse

Jöreskog und Sörbom unterscheiden neben dem konfirmatorischen Test eines Modells und dem Vergleich alternativer Modelle ein exploratives Vorgehen im Rahmen der konfirmatorischen Faktorenanalyse. Dabei werden theoriegeleitet wie datenorientiert Modellannahmen getestet und bei mangelnder Modellgüte modifiziert (Jöreskog, Sörbom 1993, S. 114). Dieses Vorgehen wurde aufgegriffen, um die Struktur der Skala weiter aufzuklären und mögliche Ursachen für die eingeschränkte Modellgüte der bisher gefundenen Lösung zu identifizieren. Dazu wurden die Messmodelle der dreidimensionalen Lösung aus dem vorherigen Abschnitt einzeln und dann paarweise untersucht und ein daraus entwickeltes Alternativmodell geprüft.

Im Messmodell zu „Orientierung/Gedächtnis“ erklärt eine latente Variable die Varianzen der Indikatoren PERSONEN, ORT, ZEIT und ERINNERN. Alle Faktoren laden hoch auf die latente Variable. Dabei fällt die Höhe der Faktorladung zu PERSONEN jedoch hinter der der anderen Variablen zurück. In der Modellgüte weist der χ^2-Test für exakten Modell-Fit das Modell zurück. Auch der RMSEA liegt über .08. Die standardisierten Residuen weisen entsprechend hohe negative Beträge aus und verweisen damit auf eine Fehlspezifikation. Die Modifikationsindizes schlagen vor, die Fehlervarianzen aller Variablen korrelieren zu lassen. Danach enthalten die Fehlerterme ein spezifisches Konstrukt, das nicht in der latenten Variablen begründet ist (Franken 2010, S. 312–315).

Betrachtet man die Daten selbst, so bilden die Antwortmuster und deren Häufigkeit die prägnanteste Form der Darstellung (Franken 2010, S. 349f). Von 256 möglichen Antwortmustern finden sich 86 Antwortmuster in der Stichprobe. Dabei unterscheiden sich die Items in ihrem Schwierigkeitsgrad. PERSONEN zeigt sich im Schwierigkeitsgrad als leichtestes Item, gefolgt von ORT, ZEIT und ERINNERN.[58] Von 1816 Fällen verstoßen 213 gegen diese Rangfolge. 853 Fälle zeigen gleiche Ausprägungen zu allen Indikatoren, d. h. umgekehrt, dass 963 Fälle zur Varianz zwischen den Indikatoren beitragen und davon 750 Fälle unterschiedliche Schwierigkeitsgrade zwischen den Indikatoren widerspiegeln. Unterschiedliche

---

restringierte Nullmodell. Weitere Indizes wie AIC, ECVI vergleichen alternative Modell und dienen der Modellselektion (Bühner 2006, S. 252–255; Schermelleh-Engel et al. 2003)

[58] Die Befragten zeigen sich am ehesten in ihrer Fähigkeit beeinträchtigt, Ereignisse oder Beobachtungen zu erinnern, während die Fähigkeit, Personen aus dem näheren Umfeld zu erkennen, erst eingeschränkt ist, wenn die anderen Fähigkeiten zum Gedächtnis und der zeitlichen wie örtlichen Orientierung zumindest im gleichen Maße beeinträchtigt sind. Größere Einschränkungen zeigen sich dementsprechend zunächst bei Gedächtnisleistungen, wohingegen höhere Beeinträchtigungen in der zeitlichen Orientierung mit zumindest gleich hohen Einschränkungen in Gedächtnisleistungen, größere Beeinträchtigungen in der örtlichen Orientierung mit zumindest gleich großen Einschränkungen in der zeitlichen Orientierung und Gedächtnis und schließlich höhere Beeinträchtigungen in der Fähigkeit, Personen zu erkennen, mit zumindest gleich großen Einschränkungen in den übrigen Items verbunden sind.

Schwierigkeitsgrade können jedoch dazu führen, dass sich Variablen in einer Faktorenanalyse nicht nach inhaltlichen Gesichtspunkten, sondern nach Schwierigkeitsgraden gruppieren.

Im Messmodell zur „Sprache“ soll eine latente Variable die Varianzen der Indikatoren MITTEILEN, VERSTEHEN und GESPRÄCH erklären. Die Faktoren laden gleichmäßig und hoch auf die latente Variable. Das Modell ist gerade identifiziert und hat damit nur eine Lösung. Ein Modelltest ist daher nicht möglich (Franken 2010, S. 316f).

Von 64 möglichen Antwortmustern finden sich 47 Antwortmuster in der Stichprobe (Franken 2010, S. 351). Konzeptionell könnte auch in diesem Modell eine Differenzierung der Items nach Schwierigkeitsgraden erwartet werden, da Beeinträchtigungen in der Fähigkeit, sich mitzuteilen oder andere zu verstehen, die Fähigkeit einschränken könnten, an einem Gespräch teilzunehmen. Erkennbar ist jedoch eine überwiegende Tendenz zu gleichen Antworten. In 1216 Fällen haben so alle Indikatoren die gleiche Ausprägung, 600 Fälle tragen umgekehrt zur Varianz bei. In 336 Fällen lassen sich dabei die Items nach Schwierigkeitsgrad unterscheiden. MITTEILEN zeigt sich dabei als leichtestes Item, gefolgt von AUFFORDERN und GESPRÄCH. 264 Fälle verstoßen aber gegen diese Rangordnung und in 30 Fällen davon zeigen sich Muster, die mit der genannten These insofern unvereinbar sind, als der völlige Verlust, sich mitteilen und bzw. oder andere zu verstehen, nicht zu einem Verlust der Fähigkeit führt, an einem Gespräch teilzunehmen. Die Variable GESPRÄCH scheint danach in ihrer Erhebung oder dem erfassten Konstrukt etwas zu erfassen, das sie inhaltlich von den Variablen MITTEILEN und AUFFORDERN unterscheidet.

Im Messmodell „Praxis“ (Abb. 4.5 - 4.8) erklärt die latente Variable die Varianzen der Indikatoren HANDELN, ENTSCHEIDEN, INFOS und GEFAHREN. Alle Faktoren laden hoch auf die latente Variable. Dabei fällt die Höhe der Faktorladung zu HANDELN jedoch hinter die der anderen Variablen zurück. In der Modellgüte weist der χ^2-Test für exakten Modell-Fit das Modell zurück. Auch der RMSEA liegt über .08. Die standardisierten Residuen weisen zwischen den Variablen INFOS, GEFAHREN und HANDELN und ENTSCHEIDEN hohe negative Beträge; zwischen HANDELN, ENTSCHEIDEN sowie INFOS, GEFAHREN dagegen hohe positive Beträge aus. Danach werden in dem Modell die Kovarianzen zwischen den Variablen mit positiven Residuen unter- bzw. zwischen den Variablen mit negativen Residuen überschätzt. Die Modifikationsindizes schlagen denn auch vor, die Fehlervarianzen der Variablen HANDELN und ENTSCHEIDEN sowie INFOS und GEFAHREN korrelieren zu lassen. Danach enthalten die Fehlerterme der entsprechenden Variablenpaare ein spezifisches Konstrukt, das nicht in der latenten Variablen begründet ist (Franken

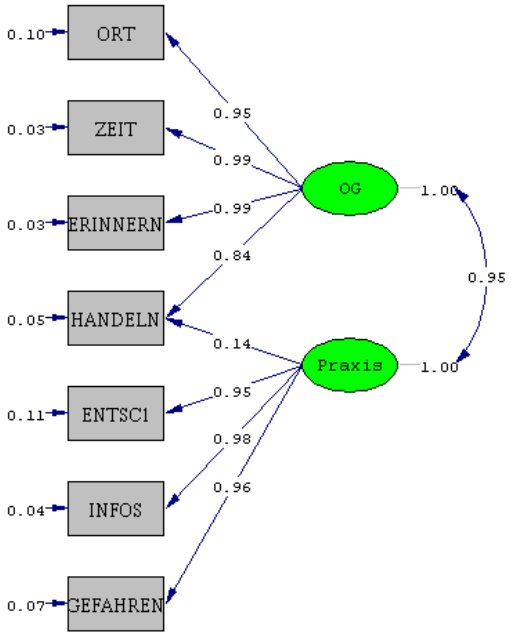
2010, S. 318–326). Da eine solche Modifikation aber nicht substantiell begründet werden kann, unterbleibt sie an dieser Stelle (Jöreskog, Sörbom 1993, S. 113).

Die Antworten weisen 118 verschiedene von 256 möglichen Mustern auf. Erkennbar ist auch hier eine Tendenz zu gleichen Antworten. So sind in den zehn häufigsten Antwortmustern alle Muster mit gleichen Antworten enthalten und variiert kein Antwortmuster in mehr als einer Variablen um eine Antwortkategorie. Auffällig aber auch der hohe Anteil an Personen, deren Fähigkeiten in praktischer Hinsicht ausschließlich beim Erkennen und der angemessenen Reaktion auf Gefahren eingeschränkt sind (Franken 2010, S. 352).

Beim Test des dreidimensionalen Modells war durch Modifikationsindizes u. a. vorgeschlagen worden, auf die Variable HANDELN Doppelladungen von den latenten Variablen „Praxis“ und „Sprache“ zuzulassen (Franken 2010, S.311). Das durch die Variable HANDELN erfasste Merkmal wäre demnach im Rahmen des untersuchten Modells eine komplexe Fähigkeit und die Varianz des Items würde durch den Einfluss praktischer wie sprachlicher Fähigkeiten erklärt. Verglichen mit dem RAI HC ließe sich HANDELN aber auch im Sinne eines „prozeduralen Gedächtnisses“ verstehen und die Varianz des Items durch die latenten Variablen „Gedächtnis“ und „Praxis“ erklären (Franken 2010, S.79). Dies sollte im Folgenden durch die paarweise Verbindung der Messmodelle untersucht werden. Um den möglichen Einfluss unterschiedlicher Schwierigkeitsgrade oder ungeklärter Inhalte zu minimieren bzw. auszuschließen, wurden dabei die entsprechenden Kombinationen auch ohne die Variablen PERSONEN bzw. GESPRÄCH berechnet.

Bei einer Verbindung der Modelle „Praxis“ und „Sprache“ mit einer Doppelladung auf HANDELN lädt der Indikator HANDELN nur gering auf die latente Variable „Sprache“ (Franken 2010, S. 322-326). Prüft man die Kombination der Modelle ohne die Variable GESPRÄCH, so sinkt die Faktorladung weiter (Franken 2010, S. 327-330). Auch bei einer Verbindung der Modelle „Orientierung/Gedächtnis“ und „Praxis“ mit einer Doppelladung auf HANDELN lädt der Indikator HANDELN nur gering auf die latente Variable „Orientierung/Praxis“ (Franken 2010, S. 331-336). Prüft man die Kombination der Modelle ohne die Variable PERSONEN, so verändert sich das Ergebnis allerdings substantiell.

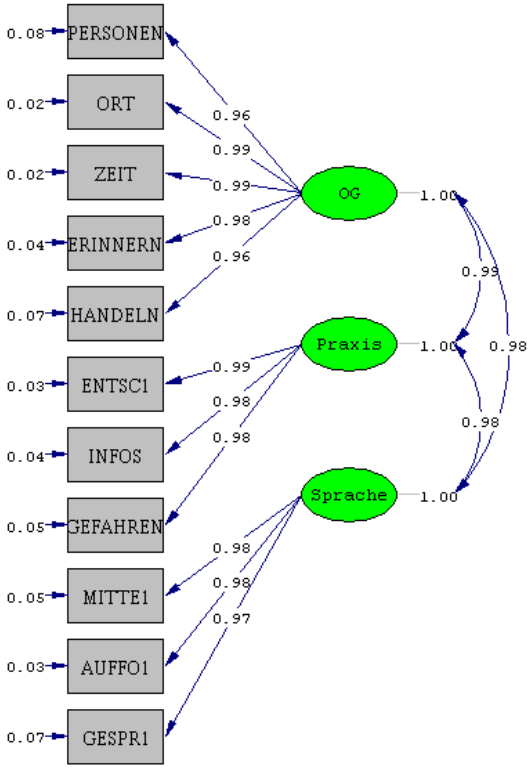
In diesem Modell lädt die Variable HANDELN primär auf die latente Variable „Orientierung/Gedächtnis“ und lediglich in einer nicht-substantiellen Nebenladung auf die Variable „Praxis“. Demnach stellt das durch die Variable HANDELN erfasste Merkmal im Rahmen des untersuchten Modells eine komplexe Fähigkeit dar, die in erster Linie durch Fähigkeiten zu Orientierung und Gedächtnis bestimmt wird (Franken 2010, S. 337-341).



Chi-Square=66.62, df=12, P-value=0.00000, RMSEA=0.050

Abb. 4.7 Modell zur Verbindung „Orientierung/Gedächtnis“ (ohne PERSONEN) und „Praxis“ mit Doppelladung auf HANDELN

Aus den Ergebnissen lässt sich zu der dreidimensionalen Lösung des vorhergehenden Abschnitts ein alternatives Modell entwickeln, in dem die Varianz der Variablen HANDELN durch die latente Variable „Orientierung/Gedächtnis“ erklärt wird.



Chi-Square=204.32, df=41, P-value=0.00000, RMSEA=0.047

Abb. 4.8 Modell „Orientierung/Gedächtnis“ (mit HANDELN), „Praxis“ (ohne HANDELN), „Sprache“

Die Indikatoren laden hoch auf die latenten Variablen. Erkennbar aber auch die im Verhältnis zu den übrigen Indikatoren hohen Fehlervarianzen der Variablen PERSONEN, HANDELN und GESPRÄCH. Zugleich korrelieren die latenten Variablen „Orientierung/Gedächtnis“ und „Praxis“ sehr hoch miteinander.

Im Vergleich zum dreidimensionalen Modell aus dem vorhergehenden Abschnitt weist in der Modellgüte der χ^2-Test für exakten Modell-Fit beide Modelle zurück. Dabei ist der χ^2-Wert des alternativen Modells jedoch erheblich größer. Während in der dreidimensionalen Lösung des vorhergehenden Abschnitts das Verhältnis des χ^2-Werts zur Anzahl der Freiheitsgrade bei 3.4 liegt, beträgt dieses Verhältnis beim alternativen Modell 5.0. Auch der RMSEA-Wert ist in jenem Modell niedriger und liegt mit seinem Konfidenzintervall unter dem Grenzwert von .05. Die bessere Modellgüte im Vergleich zeigt sich ebenso in den niedrigeren Werten der Kriterien AIC bzw. ECVI[59]. Insgesamt passt daher das dreidimensionale Modell, das die Varianz des Indikators HANDELN durch die latente Variable „Praxis“ erklärt, am besten zu den Daten der Stichprobe.

Verhältnis von numerischem und empirischem Relativ in der Modulbewertung

Für die Bewertung der im Modul 2 (NBA) erhobenen Fähigkeiten werden nach der von den Autoren des Instruments vorgeschlagenen Systematik die Merkmalsausprägungen der ersten acht Variablen addiert und die so berechneten Summenwerte einer fünfstufigen Skala zugeordnet. Zu prüfen ist jedoch, ob der im Rahmen der vorgeschlagenen Bewertungssystematik ermittelte Wert noch die empirischen Verhältnisse zwischen den kognitiven Fähigkeiten der befragten Personen abbildet, die sich bei einer Berücksichtigung der unterschiedlichen Abstände zwischen den Merkmalsausprägungen ergeben. In PRELIS können zu den Merkmalsausprägungen sogenannte Normal Scores berechnet werden. Tab. 4.5 führt die Normal Scores für die Merkmalsausprägungen der einzelnen Variablen im Modul 2 „Kognitive und kommunikative Fähigkeiten“ (NBA) an.

Nach den Normal Scores sind die Abstände zwischen den Antwortkategorien nicht gleich. Die Skala kann so interpretiert werden, dass ihre Ausprägungen nur einen Teil einer latenten normalverteilten Variablen differenzieren, so dass unter die Ausprägung 0 das Maximum sowie die überdurchschnittlichen Ausprägungen der jeweiligen Fähigkeiten fallen würde. Insofern können die Normal Scores zur Gewichtung der

[59] Der RMSEA des alternativen Modells beträgt .047 (p-Wert für RMSEA < .05 = .78), der SRMR .051. Der AICI beträgt 254, der ECVI .14 (CI .12; .17) gegenüber 187.98 bzw. .10 (CI .086; .13) beim dreidimensionalen Modell des vorhergehenden Abschnitts (Franken 2010, S. 157).

Antwortkategorien herangezogen werden, um so die unterschiedlichen Abstände zwischen den Antwortkategorien zu berücksichtigen.

		Kategorien			
		0	1	2	3
Variablen	**PERSONEN**	-0,04	1,28	1.89	2,67
	ORT	-0,06	1,38	1,94	2,75
	ZEIT	-0,08	1,29	1,94	2,82
	ERINNERN	-0,08	1,16	1,96	2,85
	HANDELN	-0,07	1,21	1,92	2,93
	ENTSCHEIDEN	-0,07	1,23	1,91	2,94
	INFOS	-0,07	1,22	1,94	2,85
	GEFAHREN	-0,09	1,18	1,94	2,92
	MITTEILEN	-0,05	1,32	1,93	2,73
	AUFFORDERN	-0,06	1,26	1,92	2,70
	GESPRÄCH	-0,06	1,28	1,93	2,85

Tab. 4.5 Normal Scores für das Modul 2 (NBA)

Durch die Verwendung unterschiedlicher Skalierungen bei der Messung an ein- und demselben Probanden lassen sich Stichprobenpaare bilden. Um zu prüfen, ob sich bei einer Summierung der ordinalen Daten des NBA im Vergleich zu einer Summierung der Normal Scores Rangunterschiede in den Bewertungen kognitiver Fähigkeiten ergeben, wurde ein Wilcoxon-Test durchgeführt. Wie in der Bewertungssystematik vorgesehen wurden dazu nur die ersten acht Variablen einbezogen.

Der Wilcoxon-Test prüft, ob die Rangunterschiede bei den Messungen an ein- und demselben Probanden größer sind als diejenigen, die zu erwarten wären, wenn die Unterschiede zufällig zustande gekommen wären (Brühl o. A., S. 15). Die Nullhypothese lautet, dass die Verwendung unterschiedlicher Skalierungen keinen Einfluss auf die Rangplätze hat. Nach der Alternativhypothese führen die unterschiedlichen Skalierungen zu Rangunterschieden. Das Signifikanzniveau ist α = 0.05%. Das Ergebnis der Berechnung des Wilcoxon-Tests ist in Tab. 4.6 dargestellt.

Es wurden 1816 Fälle einbezogen. Dabei hatten 229 Personen bei der Summierung der ordinalen Daten des NBA einen höheren Rang als bei der Summierung der Normal Scores. Umgekehrt hatten 1169 Personen bei der Summierung der ordinalen Daten einen niedrigeren Rang als bei der Summierung der Normal Scores.

Ränge

		N	Mittlerer Rang	Rangsumme
Modulwert NS 1 bis 8 absolut - Modulwert NBA 1 bis 8 absolut	Negative Ränge	229(a)	384,04	87945,50
	Positive Ränge	1169(b)	761,30	889955,50
	Bindungen	418(c)		
	Gesamt	1816		

a Modulwert NS 1 bis 8 absolut < Modulwert NBA 1 bis 8 absolut

b Modulwert NS 1 bis 8 absolut > Modulwert NBA 1 bis 8 absolut

c Modulwert NS 1 bis 8 absolut = Modulwert NBA 1 bis 8 absolut

Statistik für Test(b)

	Modulwert NS 1 bis 8 absolut - Modulwert NBA 1 bis 8 absolut
Z	-26,563(a)
Asymptotische Signifikanz (2-seitig)	,000

a Basiert auf negativen Rängen.

b Wilcoxon-Test

Tab. 4.6 Ergebnisse des Wilcoxon-Tests

Keinen Rangunterschied fand sich bei 418 Personen. Dazu gehören u. a. all diejenigen, die keine kognitiven Einschränkungen haben und daher nach beiden Skalierungen 0 Punkte erhalten. Im Ergebnis ist nicht nur die Anzahl der Personen, deren Summenwert bei Verwendung der ordinalen Daten des NBA höher als die Summe der Normal Scores ist, niedriger als die Anzahl der Personen mit höherem Summenwert bei Verwendung der Normal Scores, sondern auch deren mittlerer Rang. D. h. die Verwendung von Normal Scores führt in deutlich mehr Fällen zu deutlich höheren Differenzen. Die Rangunterschiede sind so groß, dass der Test die Nullhypothese zurückweist ($p < 0.000$).

Der Wilcoxon-Test ergibt, dass es signifikante Unterschiede in der Rangfolge der befragten Personen gibt, wenn bei den Berechnungen der Modulwerte anstelle der

vorhandenen ordinalen Merkmalsausprägungen der Subskala Normal Scores als geschätzte Werte einer zugrundeliegenden kontinuierlichen Variablen benutzt werden. Offen bleibt dabei jedoch, inwieweit die unterschiedlichen Summenwerte auch zu einer anderen Bewertung der kognitiven Fähigkeiten im Rahmen der fünfstufigen Bewertungsskala des Moduls führen und damit auch die Einstufung beeinflussen. Um diese Frage zu beantworten, wurde im Folgenden der Symmetrietest von Bowker angewendet.

Dieser Test prüft, ob bei der zweimaligen Untersuchung eines k-fach gestuften Merkmals Veränderungen von einer Kategorie *i* zu einer Kategorie *j* genauso wahrscheinlich sind wie Veränderungen von der Kategorie *j* zur Kategorie *i* (Bortz, Lienert 2008, S. 128–131). Wenn sich bei Personen die Bewertung ihrer kognitiven Fähigkeiten von einer Kategorie *i* zu einer Kategorie *j* geändert hat, sollte es nach der Nullhypothese genauso viele Personen geben, bei denen sich die Bewertung ihrer kognitiven Fähigkeiten von der Kategorie *j* zu der Kategorie *i* geändert hat. Nach der Alternativhypothese werden durch die Berücksichtigung der unterschiedlichen Abstände zwischen den Merkmalsausprägungen die kognitiven Fähigkeiten bei gleicher Einteilung der Kategorien abweichend bewertet, wobei die Änderung der Bewertung von einer Kategorie *i* zu einer Kategorie *j* wahrscheinlicher ist als eine Änderung in umgekehrter Richtung. Das Signifikanzniveau ist α = 0,05%.

Der Bowker-Test vergleicht dazu in einer k x k-Felder-Tafel alle symmetrisch zur Hauptdiagonale gelegenen Felder hinsichtlich ihrer Häufigkeiten. Die folgende Tabelle führt die Ergebnisse der unterschiedlichen Einstufungen bei gleicher Einteilung der Kategorien auf, wenn die Modulwerte durch eine Summierung der vorhandenen ordinalen Daten oder durch eine Summierung von Normal Scores als geschätzte Werte einer zugrundeliegenden kontinuierlichen Variablen berechnet werden.

		Bewertungen bei Verwendung von Normal Scores					Summe
		0	1	2	3	4	
Bewertung bei Verwendung der vorhandenen ordinalen Daten	0	410	0	0	0	0	410
	1	0	377	97	0	0	474
	2	0	0	231	89	0	320
	3	0	0	0	187	30	217
	4	0	0	0	0	395	395
Summe		410	377	328	276	425	1816

Tab. 5.7 Häufigkeiten der Modulbewertungen bei Verwendung ordinalskalierter und gewichteter Merkmalsausprägungen

Aus den Daten ergibt sich $\chi^2 = 216$ mit 10 Freiheitsgraden (Fg). Der Wert ist größer als der kritische χ^2 -Wert für 10 Freiheitsgrade und $\alpha = 0,05$. Daher wird die H_0 verworfen.

DISKUSSION

Die vorliegende Studie weist Einschränkungen auf, die bei einer Verallgemeinerung ihrer Ergebnisse und deren Interpretation berücksichtigt werden müssen. So resultiert die Untersuchungsgruppe aus Gelegenheitsstichproben unter Klienten ambulanter Pflegedienste und ähnelt damit in ihrer Zusammensetzung nicht exakt der Zielpopulation des NBA. Die Studienpopulation weicht denn auch in der Altersverteilung von der Untersuchungsgruppe der Evaluationsstudie ab.

Gegenstand der vorliegenden Studie sind Fragen zur Konstruktvalidität der Subskala „Kognitive und kommunikative Fähigkeiten“ (NBA). Fragen zur Konstruktvalidität des NBA insgesamt wie auch zur Stellung der Subskala im NBA überschreiten den Rahmen der vorliegenden Studie. Die vorliegende Studie beschränkt sich zudem auf Erwachsene und erfasst nach der Altersverteilung der Studienpopulation weit überwiegend die Beeinträchtigungen kognitiver Fähigkeiten älterer Menschen. Die Verteilung der Daten weicht dabei von einer Normalverteilung ab und die Variablen korrelieren sehr hoch, was zu Schätzproblemen geführt haben kann. Bei der Spezifikation eines Modells zeigen die standardisierten Residuen der getesteten Modelle an, dass die modellimplizierte Matrix generell höhere Kovarianzen enthält als die empirische und die Faktorladungen allgemein überschätzt werden. Hohe Korrelationen erschweren zudem die inhaltliche Differenzierung der Konstrukte. Die Ergebnisse schließen prinzipiell auch nicht aus, dass abweichende Modelle mit identischem Modell-Fits gefunden werden können.

Die Studie zeigt, dass eine EFA für sich genommen zu irreführenden Ergebnissen führen kann und das explorativ ermittelte Modell an einer neuen Stichprobe konfirmatorisch getestet werden sollte (Bühner 2006, S. 260). Sofern unterschiedliche Vorstellungen zur Struktur der Daten vorliegen, sollten dabei auch alternative Modelle miteinander verglichen und ihre Struktur schrittweise analysiert werden. Es zeigt sich aber auch, dass eine Faktorenanalyse nicht zu klären vermag, inwieweit sich ein komplexer Merkmalsbereich nach inhaltlichen Kriterien oder aufgrund unterschiedlicher Personenfähigkeiten in homogenere Teilbereiche ausdifferenziert. Dies verweist auf eine grundsätzliche Grenze des testtheoretischen Hintergrunds einer KTT, die als Theorie über Messfehler nichts darüber auszusagen vermag, wie die als "wahre" Werte verstandenen Ausprägungen einzelner Variablen zustande kommen.

Die Ergebnisse der Faktorenanalyse zeigen, dass die Subskala „Kognitive und kommunikative Fähigkeiten“ (NBA) mehrdimensional ist.

Gegenüber der von den Testautoren angenommenen Unterscheidung zwischen Indikatoren zur Kognition und Kommunikation wäre ein zweidimensionales Modell mit Indikatoren zu Orientierung/Gedächtnis sowie solchen zur Praxis vorzuziehen. Von den getesteten Modellen passt jedoch eine dreidimensionale Lösung am besten zu den Daten der Stichprobe, in der die Indikatoren zu Orientierung und Gedächtnis, die Indikatoren zur Praxis sowie die Indikatoren zur Sprache jeweils einer latenten Variablen zugeordnet sind.

Die Antwortmuster zeigen aber, dass sich die Fähigkeiten teilweise auch nach Schwierigkeitsgraden differenzieren und so bei der Faktorenanalyse neben inhaltlichen Aspekten auch unterschiedliche Anforderungsstufen den Zusammenhang zwischen den Variablen bestimmen.

Beim Versuch, den Einfluss unterschiedlicher Schwierigkeitsgrade in der Berechnung der Modelle zu minimieren, ließ sich die Variable zum „Alltagshandeln“ denn auch als komplexe Fähigkeit im Sinne eines „prozeduralen Gedächtnisses“ spezifizieren, so dass zumindest vermutet werden kann, dass unter inhaltlichen Gesichtspunkten die dimensionale Struktur von der gefundenen Lösung abweicht. Darüber hinaus zeigen die auch im Vergleich hohen Korrelationen und die erkennbare Tendenz zu gleichen Antworten, dass die Skala inhaltlich wenig differenziert.

Obwohl die Skala im Vergleich zu Referenzinstrumenten thematisch nahezu alle Aspekte kognitiver und kommunikativer Fähigkeiten erfasst, scheinen die Komplexität der Items und die Art der Erhebung die inhaltlichen Differenzierungen wieder aufzuheben. Hier wäre zu prüfen, ob man die Skala entweder verkürzt oder die Items präzisiert.

Dies betrifft auch einzelne Items, deren Inhalte vieldeutig bleiben. Hier müsste zunächst ihr Verständnis seitens der Gutachter und befragten Personen untersucht werden. Insgesamt wäre aber zu klären, welches theoretische Verständnis von Kognition inhaltlich einer allgemeinen Erfassung entsprechender Fähigkeiten zugrunde gelegt werden soll.

Für die Bewertung der im Modul 2 „Kognitive und kommunikative Fähigkeiten“ (NBA) erhobenen Fähigkeiten werden nach der von den Autoren des Instruments vorgeschlagenen Systematik die Merkmalsausprägungen der ersten acht Variablen

addiert und die so berechneten Summenwerte einer fünfstufigen Skala zugeordnet. Legt man den ordinalen Daten jedoch eine kontinuierliche, normalverteilte Variable zugrunde und berechnet für die kategorialen Ausprägungen der ordinalen Messungen Normal Scores als geschätzte Werte dieser Variable, zeigt sich, dass die Abstände zwischen den Antwortkategorien unterschiedlich groß sind und die ordinalen Daten des Moduls daher nicht als intervallskaliert interpretiert werden dürfen.

Der Vergleich der Summierung ordinaler Daten des NBA mit einer Summierung der Normal Scores und der sich daraus ergebenden Bewertung kognitiver Fähigkeiten zeigt, dass die unterschiedlichen Skalierungen nicht nur zu signifikanten Unterschieden in der Rangfolge der Summenwerte, sondern auch zu asymmetrischen Veränderungen in der Bewertung des Moduls führen.

Die Veränderungen betreffen die Bewertung geringer bis schwerer Beeinträchtigungen kognitiver Fähigkeiten. Von 474 Personen, deren kognitiven Fähigkeiten nach der derzeitigen Bewertungssystematik als gering beeinträchtigt eingeschätzt werden, haben 97 Personen oder 20,5% bei Berücksichtigung des Skalenniveaus erhebliche Beeinträchtigungen. Von 320 Personen, deren Beeinträchtigungen als erheblich eingeschätzt werden, haben 89 Personen oder 27,8% schwere Beeinträchtigungen und von 217 Personen mit schweren Beeinträchtigungen leiden 30 Personen oder 13,8% unter dem weitgehenden oder völligen Verlust ihrer kognitiven Fähigkeiten.

Eine Lösung der angezeigten Problematik wäre durch eine Gewichtung der Items für eine linear-additive Skalenbildung zu erreichen oder durch eine Ausweitung der Ratingskala und eine daran anschließende erneute Überprüfung des Skalenniveaus zu versuchen.

Um messtheoretische Annahmen nicht schadhaft zu verletzen, werden dazu mindestens fünf (Rohrmann 1978; Bagozzi 1981b, S. 380), eher sieben Skalenpunkte empfohlen (Bagozzi 1981a, S. 200).

LITERATUR

Abraham, Ivo L.; Manning, Carol A.; Snustad, Diane G.; Brashear, H. Robert; Newman, Maureen C.; Wofford, Amy B. (1994): Cognitive Screening of Nursing Home Residents. Factor structures of the Mini-Mental State Examination. In: Journal of the American Geriatric Society, Jg. 42, H. 7, S. 750–756

Abt-Zegelin, Angelika (2000): Pflegebedürftigkeit - was ist gemeint? In: Dr.med. Mabuse, Jg. 25, H. 123, S. 17–1

Bagozzi, Richard P. (1981a): Causal Modeling. A general method for developing and testing theories in consumer research. In: Advances in Consumer Research, Jg. 8, S. 195–202

Bagozzi, Richard P. (1981b): Evaluating structural equation models with unobservable variables and meas error. A comment. In: Journal or Marketing Research, Jg. 18, H. 3, S. 375–382

Baltes-Götz, Bernhard (1994): Einführung in die Analyse von Strukturgleichungsmodellen mit LISREL 7 und PRELIS unter SPSS. Universität Trier. Universitätsrechenzentrum. Online verfügbar unter http://dtserv1.compsy.uni-jena.de/ss2005/metheval_uj/sem/content.nsf/f7525b3312d0d83cc1256db0002dec75/d6fe971ffd651795c1256fd6004c53d1/Body/M2/lisrel7.pdf?OpenElement, zuletzt geprüft am 25.09.2010

Banos, James H.; Franklin, Lucy M. (2002): Factor structure of the Mini-Mental State Examination in adult psychiatric inpatients. In: Psychological Assessment, Jg. 14, H. 4, S. 397–400

Barrie, Marlene A. (2002): Objective screening tools to assess cognitive impairment and depression. In: Topics in Geriatric Rehabilitation, Jg. 18, H. 2, S. 28–46

Bartholomeyczik, Sabine (2004): Assessment als Operationalisierung von Pflegebedürftigkeit. In: Pflege Aktuell, Jg. 58, H. 1, S. 8–13

Bensch, Sandra (2011): Konstruktvalidität der Module „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessments zur Feststellung von Pflegebedürftigkeit. Inaugural-Dissertation an der Pflegewissenschaftlichen Fakultät. Philosophisch-Theologische Hochschule Vallendar

Borg, Ingwer; Staufenbiel, Thomas (2007): Lehrbuch Theorien und Methoden der Skalierung. 4. vollst. überarb. u. erw. Auflage. Bern: Hans Huber

Bortz, Jürgen; Lienert, Gustav A. (2008): Kurzgefasste Statistik für die klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben. 3. aktu. und bearb. Auflage. Heidelberg: Springer

Braatz, Frank; Gansweid, Barbara (2005): Beurteilung von Pflegebedürftigkeit aus methodologischer Sicht. In: Gaertner, Thomas; Mittelstaedt, Gert von (Hg.): Die soziale Pflegeversicherung. Erfahrungen der MDK-Gemeinschaft in der Begutachtung, Qualitätsprüfung und Beratung - Bilanz und Ausblick -. Münster: Daedalus Verlag, S. 179–187

Braekhus, A.; Laake, K.; Engedal, K. (1992): The Mini-Mental State Examination. Identifying the most efficient variables for detecting cognitive impairment in the elderly. In: Journal of the American Geriatrics Society, Jg. 40, S. 1139–1145

Brühl, Albert: Tests und Analysen. Arbeitspapier. Fünfte Fassung. Unveröffentlichtes Manuskript, o. A., SPI Köln

Bühner, Markus (2006): Einführung in die Test- und Fragebogenkonstruktion. 2. aktual. u. erw. Auflage. München, Boston, San Francisco u.a.: Pearson Studium

Colsher, Patricia L.; Wallace, Robert B. (1991): Epidemiologic considerations in studies of cognitive function in the elderly. Methodology and nondementing acquired dysfunction. In: Epidemiologic Reviews, Jg. 13, S. 1–27

Franken, Georg (2010): Konstruktvalidität der Subskala "Kognitive und kommunikative Fähigkeiten" des Neuen Begutachtungsassessments zur Feststellung von Pflegebedürftigkeit (NBA). Masterarbeit. Betreut von Prof. Dr. Albert Brühl. Vallendar. Philosophisch-Theologische Hochschule Vallendar, Pflegewissenschaftliche Fakultät. http://opus.bsz-bw.de/kidoks/volltexte/2012/66/ zuletzt geprüft am 20.08.2012

Hartig, Johannes; Frey, Andreas; Jude, Nina (2008): Validität. In: Moosbrugger, Helfried; Kelava, Augustin (Hg.): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer Medizin, S. 135–163

Hassler, Martina; Görres, Stefan (2005): Was Pflegebedürftige wirklich brauchen… Zukünftige Herausforderungen an eine bedarfsgerechte ambulante und stationäre pflegerische Versorgung. Hannover: Schlütersche

Huppert, Felicia A.; Tym, Elizabeth (1986): Clinical and neuropsychological assessment of dementia. In: British Medical Bulletin, Jg. 42, H. 1, S. 11–18

Jones, Richard N.; Gallo, Joseph J. (2000): Dimensions of the Mini-Mental State Examination among community dwelling older adults. In: Psychological Medicine, Jg. 30, H. 3, S. 605–618

Jöreskog, Karl G. (2002): Structural Equation Modeling with Ordinal Variables using LISREL. Online verfügbar unter http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf, zuletzt aktualisiert am 10.02.2005, zuletzt geprüft am 20.10.2010

Jöreskog, Karl G.; Moustaki, Irini (2006): Factor Analysis of Ordinal Variables with Full Information Maximum Likelihood. Uppsala University, Athens University of Economics and Business. Online verfügbar unter http://www.ssicentral.com/lisrel/techdocs/orfiml.pdf, zuletzt geprüft am 20.11.2010

Jöreskog, Karl G.; Sörbom, Dag (1993): LISREL 8. Structural Equation Modeling with the SIMPLIS Command Language. Lincolnwood, IL: Scientific Software International, Inc.

Jöreskog, Karl G.; Sörbom, Dag (2003): PRELIS 2. User's Reference Guide. A program for multivariate data screening and data summarization; a preprocessor for LISREL. 3rd ed., updated to PRELIS 2. Lincolnwood, IL: Scientific Software International, Inc.

McDowell, Ian (2006): Measuring Health. A Guide to Rating Scales and Questionnaires. 3rd ed. Oxford, New York, Tokyo: University Press

Menning, Sonja; Hoffmann, Elke (2009): Funktionale Gesundheit und Pflegebedürftigkeit. In: Böhm, Karin; Tesch-Römer, Clemens; Ziese, Thomas (Hg.): Gesundheit und Krankheit im Alter. Beiträge zur Gesundheitsberichterstattung des Bundes. Berlin, S. 62–78

Moosbrugger, Helfried; Schermelleh-Engel, Karin (2008): Exploratorische (EFA) und Konfirmatorische Faktorenanalyse (CFA). In: Moosbrugger, Helfried; Kelava, Augustin (Hg.): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer Medizin, S. 307–324

Murphy, Kevin R.; Davidshofer, Charles O. (2005): Psychological Testing. Principles and Applications. 6. Aufl. Upper Saddle River, New Jersey: Pearson Prentice Hall

Neuhäuser, G. (2004): Motorisches Lernen und kognitive Entwicklung. In: Schlack, H. G. (Hg.): Entwicklungspädiatrie. München

Rohrmann, B. (1978): Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. In: Zeitschrift für Sozialpsychologie, Jg. 9, S. 222–245

Schermelleh-Engel, Karin; Moosbrugger, Helfried; Müller, Hans (2003): Evaluating the Fit of Structural Equation Models. Tests of Significance and descriptive Goodness-of-Fit Measures. In: Methods of Psychological Research Online, Jg. 8, H. 2, S. 23–74. Online verfügbar unter http://www.dgps.de/fachgruppen/methoden/mpr-online/issue20/art2/mpr130_13.pdf, zuletzt geprüft am 13.10.2010

Schröder, Martina (2010): Konstruktvalidität der Subskala Mobilität des Neuen Begutachtungsassessments für Pflegebedürftigkeit (NBA). Masterarbeit. Betreut von Prof. Dr. Albert Brühl. Vallendar. Philosophisch-Theologische Hochschule Vallendar. Online verfügbar unter http://www.dip.de/datenbank-wise/detail/?no_cache=1&tx_dipwise_pi2[uid]=499, zuletzt geprüft am 09.08.2010

Werner, Burkhard (2004): Der Begriff der Pflegebedürftigkeit im Kontext der Medizin und der Pflegewissenschaft. In: Brandenburg, Hermann (Hg.): Kooperation und Kommunikation in der Pflege. Ein praktischer Ratgeber für Pflegeberufe. Hannover: Schlüter, S. 33–82

Wingenfeld, K. (2000): Pflegebedürftigkeit, Pflegebedarf und pflegerische Leistungen. In: Rennen-Allhoff, B.; Schaeffer, D. (Hg.): Handbuch Pflegewissenschaft. Weinheim: Juventa, S. 339–361

Wingenfeld, K.; Büscher, A.; Schaeffer, D. (2007): Recherche und Analyse von Pflegebedürftigkeitsbegriffen und Einschätzungsinstrumenten. Überarbeitete, korrigierte Fassung. Studie im Rahmen des Modellprogramms nach §8 Abs. 3 SGB XI. Im Auftrag der Spitzenverbände der Pflegekassen. Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/ag6/downloads/ipw_bericht_20070323.pdf, zuletzt geprüft am 16.04.2010

Wingenfeld, K.; Büscher, A.; Gansweid, B. (2008a): Das neue Begutachtungsinstrument zur Feststellung von Pflegebedürftigkeit. Überarbeitete, korrigierte Fassung. Projekt: Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Abschlussbericht zur Hauptphase 1: Entwicklung eines neuen Begutachtungsinstruments. Institut für Pflegewissenschaft an der Universität Bielefeld (IPW). Medizinischer Dienst der Krankenversicherung Westfalen-Lippe (MDK-WL). Bielefeld, Münster. Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/ag6/downloads/Abschlussbericht_IPW_MDKWL_25.03.08.pdf, zuletzt geprüft am 16.04.2010

Wingenfeld, K.; Büscher, K.; Gansweid, B. (2008b): Das neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Anlagenband. Ergänzte und korrigierte Fassung vom 25. März 2008. Institut für Pflegewissenschaft an der Universität Bielefeld (IPW). Medizinischer Dienst der Krankenversicherung Westfalen-Lippe (MDK-WL). Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/ag6/downloads/Anlagenband_IPW_MDKWL_25.03.08.pdf, zuletzt geprüft am 04.09.2010

5. Prüfung der Konstruktvalidität der Subskalen „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessments mit probabilistischen Verfahren

Sandra Bensch

Einleitung

Das Neue Begutachtungsassessment (NBA) ist ein Bestandteil des neuen Begutachtungsverfahrens zur Feststellung von Pflegebedürftigkeit. Seine Entwicklung ist im Jahre 2006 vom Bundesgesundheitsministerium in Auftrag gegeben worden und hat im Rahmen des Modellprojekts nach § 8 SGB XI „Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundeseinheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI“ stattgefunden (vgl. BMG [Hg.] 2009a, S. 16ff). Die Ergebnisse sind im Januar 2009 an die Bundesgesundheitsministerin überreicht worden und werden bis zum heutigen Tage diskutiert (vgl. BMG [Hg.] 2009b; vgl. CDU/CSU-Fraktion im Deutschen Bundestag Arbeitsgruppe Gesundheit [Hg.] 2011).

Das neue Begutachtungsverfahren dient vorrangig der Zuordnung von pflegeversicherungsrechtlichen Leistungen bei bestehender Pflegebedürftigkeit, weiterhin der Prüfung notwendiger Hilfsmittel und erforderlicher Präventions- bzw. Rehabilitationsleistungen, zur Unterstützung bei der Erstellung eines Pflege- bzw. Hilfeplans sowie zur externen und internen Qualitätssicherung der Begutachtungen (vgl. Wingenfeld et al 2008a, S. 9ff). Der Kernteil ist das bereits erwähnte Neue Begutachtungsassessment, das mit seinen acht Modulen circa 90 Items umfasst. Thematisch werden körperliche, kognitive, kommunikative und behaviourale Fähigkeiten von Personen erhoben, die einen Begutachtungsantrag gestellt haben. Dabei dienen die ersten sechs Module „Mobilität“, „Kognitive und kommunikative Fähigkeiten“, „Verhaltensweisen und psychische Problemlagen“, „Selbstversorgung“, „Umgang mit krankheits-/therapiebedingten Anforderungen und Belastungen“, „Gestaltung des Alltagslebens und soziale Kontakte“ der Erfassung von Pflegebedürftigkeit und die beiden Module „Außerhäusliche Aktivitäten“ und

„Haushaltsführung“ der Erfassung von Hilfebedürftigkeit (vgl. Wingenfeld et al 2008a, S. 22ff).

Das NBA ist so konstruiert, dass jedes Modul eine höchstmögliche Aussagekraft für sich besitzt (vgl. ebd. 2008a, S. 22). Die Testwerte werden pro Modul summiert und der Summenwert, ebenfalls pro Modul, einer fünfstufigen Bewertungsskala zugeordnet (vgl. ebd. 2008a, S. 32). Die Schwellenwerte dieser Bewertungsskalen beruhen, wie die später noch erläuterten Bedarfsgrade (aktuell: Pflegestufen), auf inhaltlichen Überlegungen (vgl. ebd. 2008a, S. 17). Dies gilt auch für die Gewichtungen der Module, so geht z. B. das Modul „Mobilität“ mit zehn Prozent und das Modul „Selbstversorgung“ mit 40 Prozent in die Bedarfsgradbestimmung ein (vgl. ebd. 2008a, S. 51). Mit einem körperlichen Anteil von somit 50 Prozent soll gesichert werden, dass die im bestehenden Pflegebegutachtungsverfahren bevorzugten körperlich beeinträchtigten Personen nunmehr nicht benachteiligt werden (vgl. Gansweid et al. 2010, S. 57). Die kognitiven Fähigkeiten werden vorrangig in den Modulen „Kognitive und kommunikative Fähigkeiten“, „Verhaltensweisen und psychische Problemlagen“ und „Gestaltung des Alltagslebens und soziale Kontakte“ mit jeweils 15 Prozent erfasst, wobei vom zweiten bzw. dritten Modul lediglich jenes mit dem höheren Wert der fünfstufigen Bewertungsskala in den Bedarfsgrad eingeht (vgl. Wingenfeld et al. 2008a, S. 85). Die Antwortkategorien sind überwiegend ordinal skaliert, jedoch auch nominal bzw. dichotom. Die Bedarfsgrade bewegen sich auf einer Skala von 0 bis 100 Punkten, wobei die ersten vier Bedarfsgrade einen Grad an Selbständigkeit angeben und damit eine quantitative Personenvariable darstellen (vgl. ebd. S. 85). Für den fünften Bedarfsgrad, der zurzeit „Personen mit besonderer Bedarfskonstellation“ heißt, ist noch nicht geklärt, ob dieser ebenfalls als quantitative oder qualitative oder sogar als gemischte Variable verwendet werden soll. Es existieren z. B. Überlegungen, dass für die Definition des fünften Bedarfsgrads eine Person in der Begutachtung mehr als 90 Punkte erreicht und gleichzeitig besondere Merkmale wie eine ausgeprägte Schmerzsymptomatik aufweist oder atmungs- bzw. zeitaufwendige ernährungsunterstützende Maßnahmen benötigt (vgl. ebd. S. 81; vgl. BMG [Hg.] 2009c, S. 22ff und 48ff).

Im Rahmen des oben genannten Modellprojekts ist die Interraterreliabilität und die interne Konsistenz des Neuen Begutachtungsassessments mit Cohens Kappa bzw. Cronbachs Alpha mit n = 1.490 (Umsetzungsstudie) untersucht worden (vgl. Windeler et al 2008, S. 45ff). Dabei handelt sich um Verfahren der klassischen Testtheorie, welche für die Testwerte stets konstante wahre Werte und konstante Fehlerwerte annehmen, welche nicht miteinander korrelieren dürfen. Tagesschwankungen der Testpersonen werden nicht berücksichtigt. Darüber hinaus setzen Verfahren der klassischen Testtheorie ein Intervallskalenniveau des Tests voraus, denn sie arbeiten

mit Mittelwerten und Standardabweichungen (vgl. Moosbrugger in: Moosbrugger/Kelava [Hg.] 2007, S. 100ff; vgl. Schermelleh-Engel et al. in: Moosbrugger/Kelava [Hg.] 2007, S. 131ff). Ein Intervallskalenniveau liegt zumindest für die ersten beiden Module des NBA nicht vor, wie später noch gezeigt wird. Die Validierungen des NBA haben sich bisher auf die Überprüfung der Inhaltsvalidität, z. B. durch Gruppendiskussionen und Befragungen von Expertinnen und Experten (vgl. Wingenfeld et al. 2008a, S. 9ff und 111) und der Kriteriumsvalidität beschränkt. Letztere ist für einen Vergleich der Demenzbestimmung u. a. im Modul „Kognitive und kommunikative Fähigkeiten“ mit dem Referenztest „Test zur Frühbestimmung von Demenzen mit Depressionsabgrenzung“ (TFDD) erfolgt (vgl. Windeler et al. 2008, S. 10 und 108). Diese Vorgehensweise überrascht, da das NBA kein Diagnostikinstrument für Demenzen darstellt und außerdem auch bei Kindern zur Feststellung von Pflegebedürftigkeit eingesetzt werden soll. Obwohl weder für das Konstrukt „Pflegebedürftigkeit“, noch für die einzelnen Subkonstrukte, wie zum Beispiel „Mobilität“ oder „Kognitive und kommunikative Fähigkeiten“, theoretische Konzeptionierungen im Sinne einer Konstruktvalidität vorliegen (vgl. Hartig et al. in: Moosbrugger/Kelava [Hg.] 2007, S. 139ff) bzw. diese mit empirischen Daten inferenzstatistisch abgesichert worden sind, ist das Neue Begutachtungsassessment im Rahmen des Modellprojekts als valide und reliabel befunden worden (vgl. Windeler et al. 2008, S. 62). Die Frage, ob das Neue Begutachtungsassessment tatsächlich misst, was es zu messen vorgibt, bleibt unbeantwortet.

METHODEN

Die Eindimensionalität, d. h. die Messung einer latenten Variablen, soll nun mit Hilfe von Testmodellen der probabilistischen Testtheorie geklärt werden. Diese Testmodelle unterliegen Annahmen, die später noch erläutert werden.

Die probabilistische Testtheorie berücksichtigt im Gegensatz zur klassischen verschiedene Personenfähigkeiten und unterschiedliche Itemschwierigkeiten. Darüber werden Wahrscheinlichkeitsaussagen getroffen, nämlich ob eine Person v mit der Personenfähigkeit θ_v ein Item i mit der Itemschwierigkeit σ_i lösen kann bzw. – bezogen auf Modelle für ordinale Daten – mit welcher Wahrscheinlichkeit eine Antwortkategorie besetzt wird (vgl. Rost 2004, S. 203 ff.; vgl. Strobl 2010, S. 7 ff.).

Die Zufallsvariable X_{vi} beachtet dabei, dass eine Person v an manchen Tagen trotz guter Fähigkeiten ein schlechtes Ergebnis im Item i erzielt bzw. trotz schlechter

Fähigkeiten ein Ergebnis über ihrem Niveau erreicht (vgl. Strobl 2010, S. 7ff). Damit erscheint gerade die Anwendung von Testmodellen der probabilistischen Testtheorie zur Konstruktvalidierung des Neuen Begutachtungsassessments als besonders geeignet: Pflegebedürftigkeit ist ein dynamisches Merkmal, d. h. die Abhängigkeit von personeller Hilfe zur Bewältigung von Alltagsaktivitäten kann z. B. zu einem Zeitpunkt t_0 gar nicht vorhanden sein, zu t_1 in bestimmten Bereichen in Form von Unterstützungen und zu t_2 in einigen dieser Bereiche oder in anderen zur vollständigen Übernahme durch eine dritte Person führen. Gerade bei Personen mit kognitiven Funktionsstörungen können diese drei Zustände innerhalb eines Tages auftreten.

Die Sammlung der empirischen Daten beschränkt sich auf die ersten beiden Module des NBA „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“. Beide sind vierstufig ordinalskaliert angelegt mit den Punktwerten 0, 1, 2 und 3 von völliger Fähigkeit bzw. Selbständigkeit zu völliger Unfähigkeit bzw. Unselbständigkeit gehend. Das Modul „Mobilität“ umfasst fünf Items: Positionswechsel im Bett, Stabile Sitzposition halten, Aufstehen aus sitzender Position/Umsetzen, Fortbewegen innerhalb eines Wohnbereichs und Treppensteigen (vgl. Brühl 2012, in diesem Band, S. 27). Das zweite Modul unterteilt sich in acht kognitions- und drei kommunikationsbezogene Items: Personen aus dem näheren Umfeld erkennen, Örtliche Orientierung, Zeitliche Orientierung, Gedächtnis, Mehrschrittige Alltagshandlungen ausführen, Entscheidungen im Alltagsleben treffen, Sachverhalte und Informationen verstehen, Risiken und Gefahren erkennen, Elementare Bedürfnisse äußern, Verstehen von Aufforderungen und Beteiligung an einem Gespräch (vgl. Wingenfeld et al 2008b, B-2 ff; vgl. Franken 2012, in diesem Band, S. 89). Da bereits von der Entwicklergruppe zwei Dimensionen innerhalb eines Moduls angenommen werden – Kognition und Kommunikation – ist es sinnvoll, die Skalen zur Untersuchung der Eindimensionalität zu trennen. Die Daten werden aus organisatorischen Gründen von Pflegenden der stationären und ambulanten Einrichtungen erhoben, so dass die Datenerfassung nicht den Umständen der Begutachtungssituation nach SGB XI entspricht. Die Testpersonen müssen im vorliegenden Fall keine Pflegestufe besitzen, jedoch 18 Jahre alt sein, um formal von Erwachsenen sprechen zu können. Die Pflegenden dürfen die Daten erst nach einer Schulung mit einem differenzierten Handbuch zu den einzelnen Antwortkategorien der Items bzw. ausschließlich erheben, wenn ihnen die Testpersonen hinreichend bekannt sind. Dadurch sollten sich die bereits beschriebenen tagesformabhängigen Fähigkeitszustände der Probandinnen und Probanden nivellieren und höchstmöglich wahre Testwerte entstehen.

Insgesamt sind für die Untersuchung probabilistischer Testmodelle 5.080 Daten im Modul „Mobilität“, 5.065 Daten im Kognitionsteil und 5.107 Daten im Kommunikationsteil des zweiten Moduls vorhanden. Diese werden mit dem 1-

parameter logistic-Rasch-Modell (1-pl-Rasch-Modell) untersucht, welches lediglich die Schwierigkeitsparameter der zur Skala gehörenden Items als Itemparameter besitzt und keine Trennschärfe- oder Rateparameter. Damit werden bereits zwei wesentliche Annahmen des Rasch-Modells deutlich: die Items erklären *ein* Konstrukt und trennen vergleichbar genau die Personen mit den höheren Fähigkeitsgraden von denen mit den niedrigeren. Zum Vergleich, welches Testmodell nun besser zu den empirischen Daten passt, können verschiedene herangezogen werden, z. B. eines zur Messung einer qualitativen Personenvariablen oder ein mehrdimensionales Messmodell. Da Items ausschließlich summiert werden dürfen, wenn sie innerhalb einer Skala *ein* latentes Konstrukt erklären können (vgl. Rost 2004, S. 122 und 253ff) – und dies im Rahmen der Entwicklung des Neuen Begutachtungsassessments bereits ohne entsprechende Untersuchungen als gültig angenommen worden ist – fällt die Entscheidung auf den Vergleich der Ergebnisse aus dem 1-pl-Rasch-Modell mit denen der latenten Klassenanalyse.

Für die Untersuchung, ob das Rasch-Modell auf die empirischen Daten passt und damit die Items innerhalb der Skalen summiert werden dürfen, werden als Erstes die ordinalen Daten in Form von zwei Varianten dichotomisiert (vgl. Tab. 5.1).

<table>
<tr><th colspan="4">Erste Dichotomisierungsform</th><th colspan="4">Zweite Dichotomisierungsform</th></tr>
<tr><th>Ordinaler Wert</th><th>Ordinale Bezeichnung</th><th>Dichotomer Wert</th><th>Dichotome Bezeichnung</th><th>Ordinaler Wert</th><th>Ordinale Bezeichnung</th><th>Dichotomer Wert</th><th>Dichotome Bezeichnung</th></tr>
<tr><td>0</td><td>selbständig</td><td rowspan="2">1</td><td rowspan="2">selbständig</td><td>0</td><td>selbständig</td><td>1</td><td>selbständig</td></tr>
<tr><td>1</td><td>überwiegend selbständig</td><td>1</td><td>überwiegend selbständig</td><td rowspan="3">0</td><td rowspan="3">unselbständig</td></tr>
<tr><td>2</td><td>überwiegend unselbständig</td><td rowspan="2">0</td><td rowspan="2">unselbständig</td><td>2</td><td>überwiegend unselbständig</td></tr>
<tr><td>3</td><td>unselbständig</td><td>3</td><td>unselbständig</td></tr>
</table>

Tab. 5.1 Dichotomisierung der ordinalen Daten

Dann erfolgt die inferenzstatistische Untersuchung in verschiedenen Statistikprogrammen. Die zweite Dichotomisierungsform produziert dabei stets so schlechte Werte in den Skalen, dass sie nicht weiter verfolgt wird. Dies ist bedauerlich, da diese Form der Datenverdichtung die inhaltlich bessere ist – sie lässt die personelle Hilfe vollständig aus dem dichotomen Wert = 1 außen vor.

Dichotome Mobilitätsskala im Rasch-Modell

In WINMIRA erreichen die globalen Modellgeltungstests zur Überprüfung der H_0: „Das Rasch-Modell passt auf die empirischen Daten.“, für die erste Dichotomisierungsform ausschließlich statistisch hoch signifikante Werte bei $p \leq 0{,}05$:

```
                                emp. value     chi-square p-value

    Cressie Read            :          667.38    p=  0.0000
    Pearson Chisquare       :          713.88    p=  0.0000


 ================================================================


    Likelihood ratio        :          623.31    p=  0.0000
    Freeman-Tukey  Chi^2    :          618.97    p=  0.0000
    Degrees of freedom      :              25
```

Tab. 5.2 Modellgeltungstests für das Modul „Mobilität“

Signifikante Prüfgrößen bedeuten bei der Untersuchung des Rasch-Modells, dass die H_0 verworfen werden muss und nicht gilt. Man könnte bereits an dieser Stelle vermuten, dass die beiden oben vorgestellten Annahmen (das Modul „Mobilität“ ist eindimensional und trennt in den Items vergleichbar genau die Personen mit verschiedenen Fähigkeiten) verletzt werden. Dazu muss man jedoch zunächst das Verhältnis der möglichen Pattern (Antwortmuster aller Mobilitäts-Items) zu den beobachteten betrachten: $2^5 = 32$ Antwortmuster sind in der Stichprobe möglich, jedoch treten lediglich 23 Pattern auf. Da weniger Antwortmuster beobachtet werden, als möglich sind, ist davon auszugehen, dass die abgebildeten Prüfgrößen des Pearson-Chi-Quadrat-, des Cressie-Read- und des Likelihoodquotiententests – die einer Chi-Quadrat-Verteilung folgen – nicht Chi-Quadrat-verteilt sind. Das bedeutet, dass die empirischen Daten zur Schätzung des Rasch-Modells im Vergleich zum saturierten Modell[60] zu einseitig sind, d. h. zu wenig Aussagekraft besitzen. Bei einer Patternrelation = 1, d. h., alle möglichen Pattern des Tests treten auf, könnten die Personen- und Itemparameter Werte aufweisen, die sogar zu einer Modellgültigkeit führen könnten. Um dem Rasch-Modell und damit einer Itemsummation noch eine

[60] Im saturierten Modell entspricht die Parameteranzahl der Anzahl der Beobachtungen, d.h. das saturierte Modell kann genau an die beobachten Daten angepasst werden, bzw. das saturierte Modell beschreibt die Daten perfekt.

Chance zu geben, wird für die Mobilitätsdaten das Bootstrapverfahren (bootstrap, engl. = Stiefelschlaufe) eingesetzt. Dabei entsteht auf der Basis des empirischen Datensatzes eine neue, künstliche Stichprobe mit neuen, zufälligen Testwerten (vgl. Rost 2004, S. 336ff; vgl. Bühner 2011, S. 537). Die Anzahl der Stichproben kann je nach Statistikprogramm beliebig hoch sein, allerdings stabilisieren sich die Prüfgrößen zwischen 1.000 und 2.000 künstlichen Datensätzen (vgl. Langeheine et al. 1995, S. 47). Die H_0 wird verworfen, wenn sich die empirischen Prüfgrößen in den fünf Prozent der größten Prüfgrößen der simulierten Verteilung befinden. Das Bootstrapping ist ein umstrittenes Verfahren, da trotz statistisch nicht signifikanter Ergebnisse andere Indikatoren mitunter dennoch für einen Modellverstoß sprechen (vgl. Heene et al. 2010 aufgeführt in Bühner 2011, S. 537). Für die Mobilitätsskala entstehen in WINMIRA nach einem Bootstrapping mit 400 Stichproben ausschließlich statistisch hoch signifikante Prüfgrößen:

Parametric Bootstrap estimates for Goodness of Fit:

No.:	Satlik	LogLik	LR	CressieRead	Pearson X^2	FT
		Z:	51.259	12.777	1.913	56.8883
		P(X>Z):	0.000	0.000	0.028	0.0000
		Mean:	35.551	53.967	115.097	36.3806
		Stdev:	11.467	48.008	313.045	10.2409
		p-values (emp. PDF):	0.000	0.000	0.015	0.0000

Tab. 5.3 Gesamtmodellgeltungstest zum Rasch-Modell für die Mobilitätsskala mit Bootstrap

Diese Ergebnisse bestätigen sich auch in anderen globalen Modellgeltungstests wie dem Likelihoodquotiententest nach Andersen im Paket „extended Rasch modeling“ (eRm) in der Statistiksoftware R und dem Total-Item-Chi-Square-Test im Statistikprogramm RUMM. Das Rasch-Modell passt nicht auf die empirischen, dichotomisierten Daten, so dass die Items nicht summiert werden dürfen.

Wegen der angezeigten Modellverletzungen wird überprüft, wie die empirischen Daten und damit letztendlich die Mobilitätsskala selbst charakterisiert ist. Dazu wird zunächst in Abb. 5.1 die Item Characteristic Curve (ICC) der fünf Items dargestellt.

Die Abszisse offenbart das latente Konstrukt, in diesem Fall „Mobilität“, und die Ordinate die Lösungswahrscheinlichkeiten der Testpersonen für die fünf Items. Die

Annahme gleicher Trennschärfe zeigt sich in den identischen Steigungen der Funktionen (vgl. Strobl 2010, S. 12) – nur eben an unterschiedlichen Stellen des latenten Konstrukts, da die Items unterschiedlich schwierig sind. Die Itemschwierigkeiten in Abbildung 5.1 entsprechen nicht der Abfolge im Fragebogen, so stellt sich „Stabile Sitzposition halten“ (rote Funktion) als einfachstes und „Treppensteigen“ (türkisfarbene Funktion) als schwierigstes Item dar.

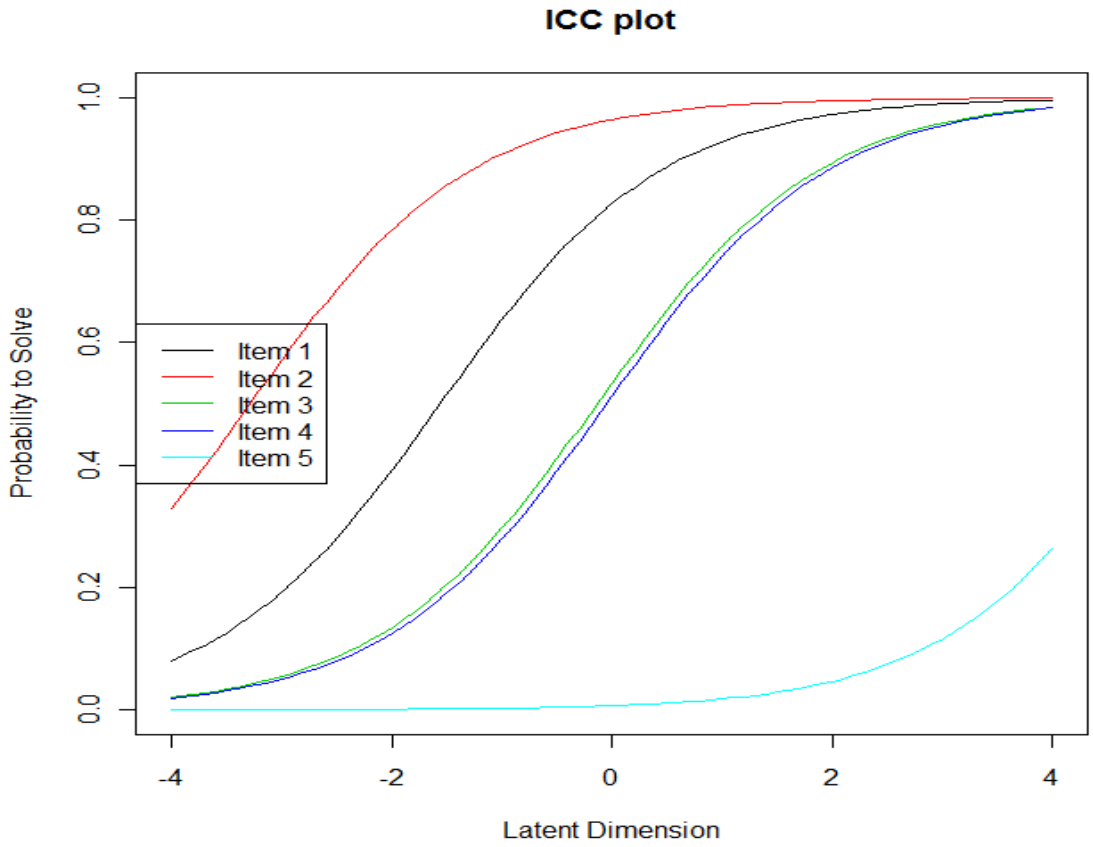


Abb. 5.1 ICC Mobilität dichotom (eRm in R)

Die Aufgaben „Aufstehen aus sitzender Position/Umsetzen“ (grüne Funktion) sowie „Fortbewegen innerhalb des Wohnbereichs“ (blaue Funktion) unterscheiden sich in ihren Schwierigkeiten nur unwesentlich voneinander, sodass ein Item entfernt werden könnte. Die beiden Aufgaben stellen sehr ähnliche Anforderungen an die Testpersonen, wodurch der Informationszugewinn zu gering ist. Die restlichen Items besitzen klar voneinander unterscheidbare Schwierigkeiten und eignen sich diesbezüglich für die Skala. An Abbildung 5.1 lässt sich die Stärke der probabilistischen Testtheorie mit logarithmierten Funktionen darstellen: bei einer Lösungswahrscheinlichkeit $p(x_{vi}) = 0{,}5$ entspricht die Personenfähigkeit θ_v der Itemschwierigkeit σ_i (vgl. Strobl 2010, S. 10; vgl. Rost 2004, S. 97ff). So lässt sich für jede Person der Stichprobe vorhersagen, ob sie die fünf Items der Mobilitätsskala

lösen kann oder nicht. Dies gilt aber immer nur solange, wie die Annahme der suffizienten Statistiken aufrecht erhalten werden kann, d. h., dass der Score einer Person r_v (Summation der Testwerte eines Person über alle Items hinweg) alles über deren Fähigkeit aussagt (vgl. Strobl 2010, S. 14ff; vgl. Brandstätter 2001, S. 29ff).

Die Items lassen sich im Verhältnis zueinander auch über den graphischen Modelltest darstellen. Um die Gesamtstichprobe in zwei Subgruppen zu teilen, werden zwei Splitkriterien verwendet: der Mittelwert und das Setting. Die Itemparameter sind für beide Gruppen gleich, wenn die Items oder deren 95-prozentige Konfidenzintervalle die Winkeldiagonale berühren – dann gilt die H_0 als verifiziert (vgl. Strobl 2010, S. 67). An dieser Stelle kann eine weitere Annahme des Rasch-Modells geprüft werden: die spezifische Objektivität. Diese besagt, dass die Lösungsfähigkeit alleine die Lösungswahrscheinlichkeit bestimmt und sich nicht weitere Faktoren wie z. B. das Setting auf die Lösungswahrscheinlichkeit auswirken. (vgl. Strobl 2010, S. 23; vgl. Brandstätter 2001, S. 27).

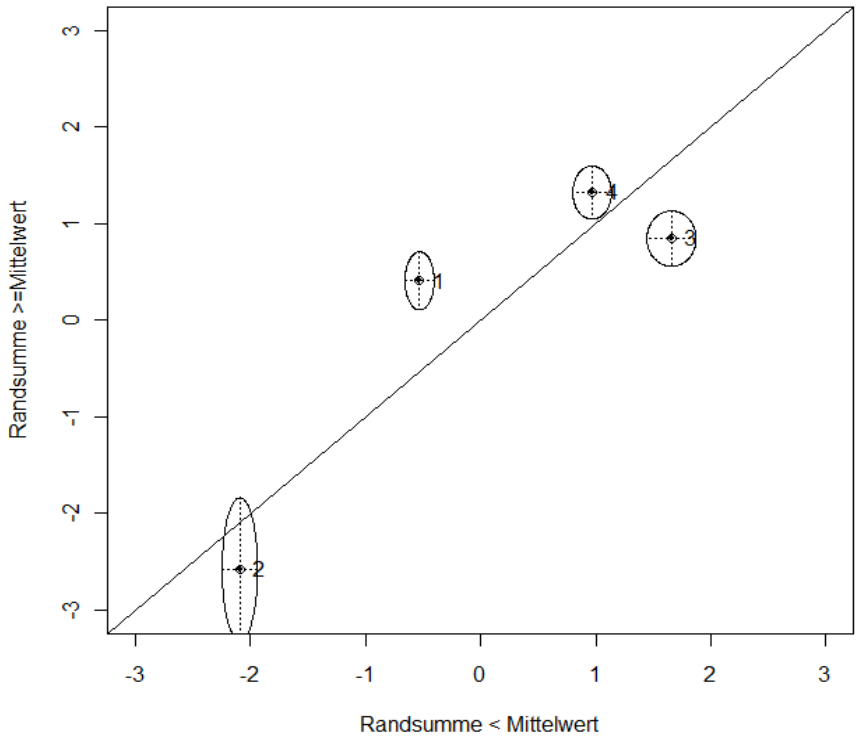


Abb. 5.2: Grafischer Modelltest Mobilität dichotom mit dem Splitkriterium „Mittelwert“ (eRm in R)

Abb. 5.2 demonstriert vier Items der Mobilitätsskala, von denen lediglich das Item „Stabile Sitzposition halten“ mit seinem Konfidenzintervall die Winkeldiagonale berührt. Für dieses Item ist also noch anzunehmen, dass die Itemparameter für die beiden

Teilgruppen $r_v \geq$ Mittelwert (MW) bzw. $r_v <$ MW gleich geschätzt werden. Daraus folgt, dass dieses Item für die beiden Teilgruppen gleich leicht (vgl. Abb. 5.2) ist. Dies kann für die drei anderen Items nicht konstatiert werden. Außerdem geht das Item „Treppensteigen" gar nicht mit in die Berechnung ein, weil bei der Splittung nach dem Mittelwert nicht genügend 0- bzw. 1-Antworten in der jeweiligen Subgruppe vorliegen. So ist davon auszugehen, dass die Gruppe mit $r_v \geq$ MW fast ausschließlich 1-Antworten in diesem Item bzw. die mit $r_v <$ MW fast ausschließlich 0-Antworten produziert. Ein Blick auf die Antwortmusterverteilung offenbart, dass auf die beiden extremen Antwortmuster 1.026 ($r_v = 0$) bzw. 1.358 Daten ($r_v = 5$) entfallen. Daraus folgt, dass der Test für einen bestimmten Prozentsatz der Testpersonen zu schwer ist bzw. für einen großen Anteil der Stichprobe zu leicht. Das Pattern 1 1 1 1 0 ist sogar 1.393-mal vergeben worden. Das Item „Treppensteigen" stellt sich somit als sehr schweres Item dar, was sich in den Abb. 5.1 und 5.3 bestätigt.

Für ein Testmodell kann auch untersucht werden, ob die einzelnen Items modellkonform sind. Dazu existieren verschiedene Itemfit-Statistiken (lokale Fitindizes), die hauptsächlich prüfen, ob Items „zu gut" für einen Test geeignet sind und damit vorrangig deterministische und wenig probabilistische Antwortmuster entstehen oder umgekehrt Items „zu schlecht" sind und überzufällige Pattern produziert werden (vgl. Bühner 2011, S. 543ff und 574). Erweisen sich Items als ungeeignet, indem sie sich nicht im Referenzbereich befinden, sollten sie eliminiert und die Modelleignung mit der verschlankten Skala erneut überprüft werden. Dies geht natürlich nur, solange eine sinnvolle Menge an Items nach der Elimination vorhanden bleibt. Die Infit-Statistiken der Software eRm in R erweisen sich beispielsweise in den Berechnungen als zuverlässige Indikatoren; modellkonforme Items sollten im Referenzbereich von $0{,}75 \leq x \leq 1{,}33$ liegen (vgl. Wilson et al. 2006, S. i16). Die Resultate der Item Infits werden im vorliegenden Fall stets durch die Q-Indizes in WINMIRA unterstützt. In der dichotomen Mobilitätsskala erreicht mit 0,903 lediglich das Item „Positionswechsel im Bett" den Referenzbereich der Item Infits. Alle anderen Aufgaben weisen Werte < 0,75 auf, z. B. „Treppensteigen" = 0,238. Dies deutet auf stark deterministische Antwortmuster hin, d. h. die Items interagieren so stark miteinander, dass die Antworten bereits vorhersagbar sind (vgl. Linacre 2009, S. 200). Dieser Hinweis hat sich in WINMIRA mit Q-Indizes < 0,10 (vgl. Bühner 2011, S. 544) bestätigt und verweist auf die Verletzung einer weiteren wichtigen Annahme im Rasch-Modell: der lokalen stochastischen Unabhängigkeit. Diese besagt, dass das Lösen des Items i nicht vom Ergebnis des Items j abhängen darf. Umgekehrt darf also der Fähigkeitsgrad für das Bewältigen des Items j nicht die Voraussetzung für das Resultat von Item i sein (vgl. Strobl 2010, S. 16ff; vgl. Fricke 1972 aufgeführt in Brandstätter 2001, S. 29).

Für die Mobilitätsskala ist diese Forderung nach lokaler stochastischer Unabhängigkeit im Grunde bereits in der Formulierung der Items an sich verletzt: Wer nicht selbständig sitzen oder sich nicht im Wohnbereich fortbewegen kann, sollte auch keine Treppen steigen können. Im gültigen Rasch-Modell müssen jedoch alle möglichen Antwortmuster vorkommen können, d. h. die lokale stochastische Unabhängigkeit muss gegeben sein. Erst dann dürfen Items summiert werden.

Auf das Auftreten von Boden- und Deckeneffekten (r_v = 0 bzw. 5) in der vorliegenden Stichprobe weist Abbildung 5.3 hin (vgl. Bortz et al. 2006, S. 558). Personen verfügen teilweise über noch höhere Fähigkeiten, als die Mobilitätsskala erfassen kann. Gleichzeitig besitzen Menschen aus der Stichprobe noch geringere Fähigkeiten, als der Test zu erkennen in der Lage ist. Dies demonstrieren auch die beiden Abszissen – auf der oberen sind die geschätzten Personenparameter und auf der unteren die geschätzten Itemparameter abgebildet. Insgesamt, so zeigen es die Säulenhöhen von r_v = 4 und 5, sind die Items der Mobilitätsskala eher leicht zu bewältigen. Dies macht eine Testüberarbeitung notwendig, da das Instrument anderenfalls unzureichend zwischen den Fähigkeitsgraden der Testpersonen differenziert.

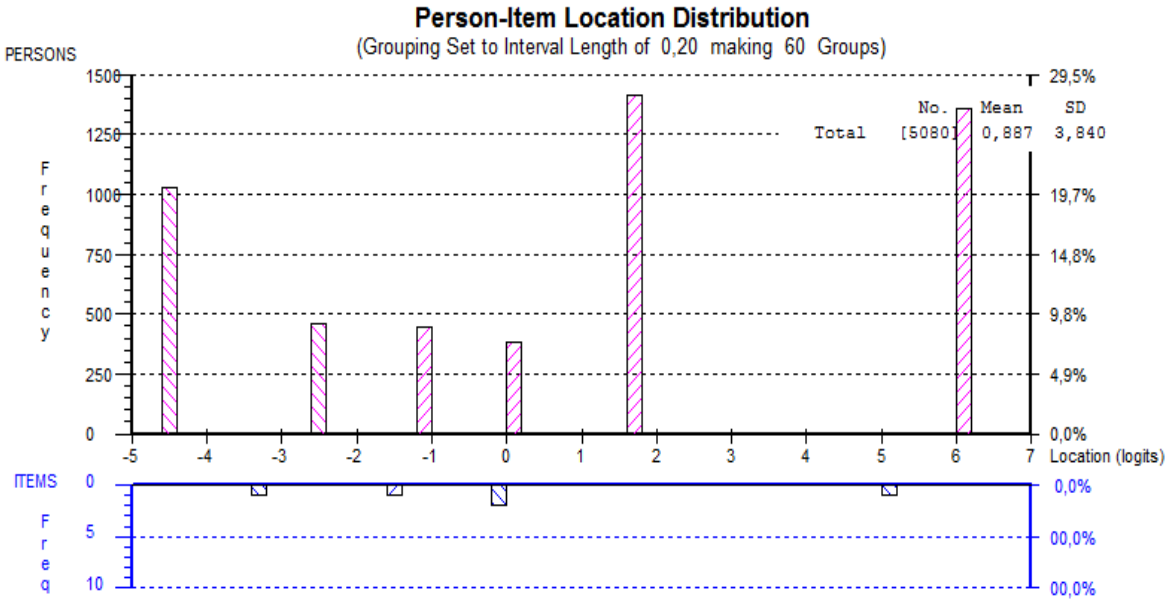


Abb. 5.3 Person-Item-Map Mobilität dichotom (RUMM)

Dichotome Kognitionsskala im Rasch-Modell

Die Ergebnisse der dichotomen Kognitionsskala ähneln stark denen der dichotomen Mobilitätsskala. Ohne und mit Bootstrapping werden alle globalen Modellgeltungstests in WINMIRA, eRm in R und RUMM für die erste Dichotomisierungsform (vgl. Tab. 5.1) statistisch hoch signifikant. Damit eignet sich das dichotome Rasch-Modell nicht für die

Erklärung der empirischen Daten und die acht kognitionsbezogenen Items des Moduls 2 dürfen nicht summiert werden. Besonders wird auf das Ergebnis des Andersen-Tests mit dem Splitkriterium „Setting“ hingewiesen. Hier entsteht eine statistisch hoch signifikante Prüfgröße, d. h., dass für die Testpersonen aus dem stationären Bereich andere Itemparameter geschätzt werden als für Testpersonen des ambulanten Bereichs. Somit sind die Items für die beiden Subgruppen unterschiedlich schwer bzw. leicht zu bewältigen (vgl. Strobl 2010, S. 41; vgl. Rost 2004, S. 348). Für die Kognitionsskala spielt dies wie für die Mobilitätsskala, die das gleiche Ergebnis aufweist, eine bedeutsame Rolle, da sie beide im Rahmen der Pflegebedürftigkeitsfeststellung der Bedarfsgradbestimmung dienen und ambulant bzw. stationär lebende Personen eigentlich gleich behandeln müssten. Der graphische Modelltest veranschaulicht die Verletzung der spezifischen Objektivität (vgl. Abb. 5.4).

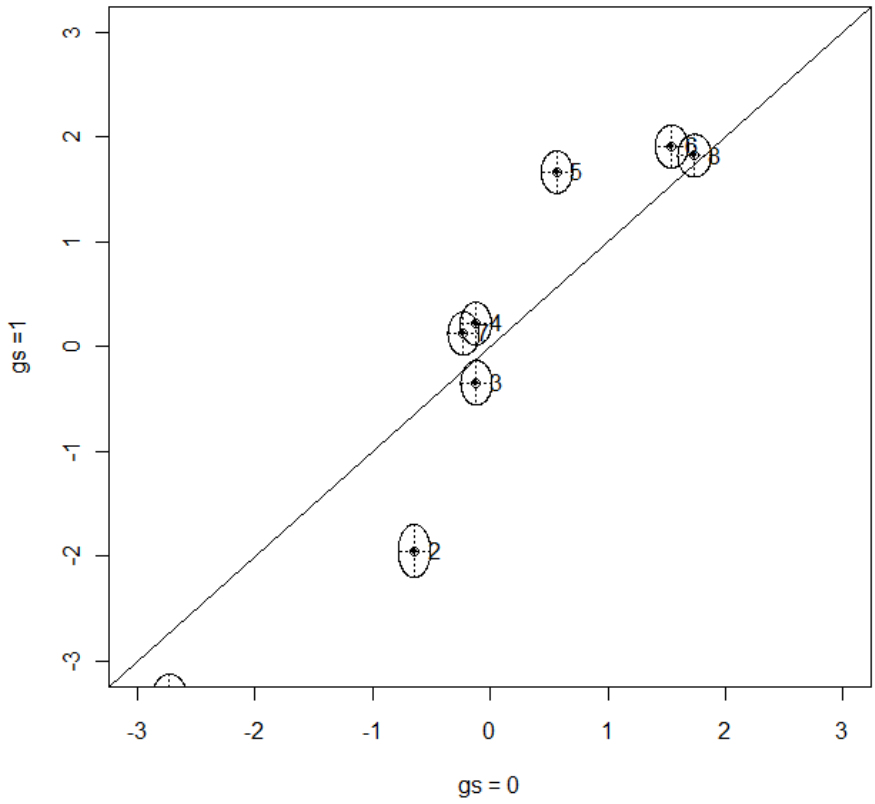


Abb. 5.4 Grafischer Modelltest Kognition dichotom mit dem Splitkriterium „Setting“ (eRm in R)

Im grafischen Modelltest in Abb. 5.4 wird deutlich, dass lediglich die Items „Risiken und Gefahren erkennen“ und „Zeitliche Orientierung“ für beide Teilstichproben gleich schwer bzw. gleich leicht zu lösen sind. Die Aufgaben „Personen aus dem näheren Umfeld erkennen“ und „Örtliche Orientierung“ sind für ambulant lebende Personen (Wert = 1) einfacher zu bewältigen als für stationär lebende (Wert = 0). Umgekehrt

haben es Menschen aus dem stationären Bereich einfacher beim Lösen der Aufgaben „Gedächtnis“, „Mehrschrittige Alltagshandlungen ausführen“, „Entscheidungen im Alltag treffen“ und „Sachverhalte und Informationen verstehen“. Diese Ergebnisse können verschiedene Ursachen haben, z. B. sind die Items „Personen aus dem näheren Umfeld erkennen“ und „Örtliche Orientierung“ im ambulanten Setting eventuell einfacher zu bewältigen, da sich die Umgebung der Pflegebedürftigen konstanter als im stationären Bereich darstellt.

Settingabhängige Unterschiede dürfen jedoch nicht in einem Test auftreten, der für alle Personen die gleichen Konsequenzen hat.

Die Infit-Statistiken präsentieren im nachfolgenden Output (Tab. 5.4) für alle Items außer „Zeitliche Orientierung“ Werte im Referenzbereich, sodass sich nach Exklusion der Items mit den statistisch signifikanten Chi-Quadrat-Statistiken folgende Aufgaben als modellkonform erweisen: „Örtliche Orientierung“, „Gedächtnis“, „Entscheidungen im Alltagsleben treffen“, „Sachverhalte und Informationen verstehen“ sowie „Risiken und Gefahren erkennen“.

Itemfit Statistics:

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
Pe.	4041.240	1845	0.000	2.189	1.020	7.97	0.76
ör.	1320.087	1845	1.000	0.715	0.782	-6.65	-9.75
ze.	1163.090	1845	1.000	0.630	0.734	-11.43	-13.03
Ge.	1272.993	1845	1.000	0.690	0.781	-9.48	-10.73
me.	1979.781	1845	0.015	1.072	1.075	1.58	3.69
En.	1421.804	1845	1.000	0.770	0.861	-3.70	-7.20

Tab. 5.4 Lokale Modellgeltungstest für sechs Items der Kognitionsskala

Diese fünf kognitionsbezogenen Items erreichen beim Trennschärfeindex in WINMIRA (Q-Index) Werte, die dem idealen p_{zq}-Wert = 0,50 am nächsten liegen:

itemlabel	Q-index	Zq	p(X>Zq)	
Personen	0.0195	1.1671	0.12158	-Q...!....+
örtl.	0.0159	-0.7068	0.76015	-....!.Q..+
zeitl.	0.0121	-1.2788	0.89951	-....!..Q.+
Gedächtn.	0.0138	-0.9773	0.83580	-....!..Q.+
mehr. All.	0.0250	1.0383	0.14957	-Q...!....+
Entsch.	0.0135	-0.4961	0.69009	-....!Q...+

Tab. 5.5 Lokale Modellgeltungstest (Q-Index) für sechs Items der Kognitionsskala

Die Q-Indizes liegen wie in der Mobilitätsskala weit unter 0,10 und sprechen pro Item für ein mehrdimensionales Konstrukt bzw. für Itemdopplungen (vgl. Bühner 2011, S. 544). Ein Blick in die häufigsten Antwortmuster verrät, dass die extremen Antwortmuster, r_v = 0 und 8, 1.200-mal bzw. 2.019-mal gewählt worden sind. Davon abgesehen, dass damit Personen mit noch schlechteren bzw. besseren Fähigkeiten existieren als mit diesem Kognitionstest erfasst werden können, werden von 256 möglichen Antwortmustern lediglich 162 beobachtet. Die lokale stochastische Unabhängigkeit scheint offenbar verletzt, das zeigt sich auch in der inhaltlichen Formulierung der Items: Wer Sachverhalte und Informationen versteht, z. B. akustisch oder visuell „Halt, hier frisch gewischt!“, kann auch höchstwahrscheinlich Risiken und Gefahren erkennen.

Itemeliminationen führen in der dichotomen Kognitionsskala nicht zur Eignung des Rasch-Modells für die empirischen Daten.

So präsentieren zwar die Item Infit-Statistiken, die Q-Indizes und die Fit Residuals – das sind weitere lokale Fitindizes im Statistikprogramm RUMM – für die verschlankte Kognitionsskala mit den Aufgaben „Örtliche Orientierung“, „Gedächtnis“, „Entscheidungen im Alltagsleben treffen“, „Sachverhalte und Informationen verstehen“ und „Risiken und Gefahren erkennen“ durchweg modellkonforme Werte, jedoch nicht der Andersen-Test und der itemspezifische Wald-Test, einem weiteren lokalen Fit-Index, mit dem Splitkriterium „Setting“. Hier bleiben die Prüfgrößen wie beim Total-Item-Chi-Square-Test in RUMM statistisch signifikant, sodass auch die Kognitionsskala mit fünf Items nicht Rasch-valide ist.

Interessanterweise entstehen bei der noch mehr verschlankten Kognitionsskala mit den Items „Entscheidungen im Alltagsleben treffen“, „Sachverhalte und Informationen verstehen“ und „Risiken und Gefahren erkennen“ statistisch nicht signifikante Prüfgrößen für die globalen Modellgeltungstests in WINMIRA und dem Andersen-Test mit den Splitkriterien „Mittelwert“ (vgl. linken Output) und „Setting“ (vgl. rechten Output):

```
Andersen LR-test:
LR-value: 0.28
Chi-square df: 2
p-value:  0.869
```

```
Andersen LR-test:
LR-value: 3.237
Chi-square df: 2
p-value:  0.198
```

Diese Resultate entstehen vor dem Hintergrund des gesamten Datensatzes. In RUMM erscheint jedoch nach Exklusion der beiden extremen Antwortmuster (alle Personen lösen alle Items bzw. ein Item wird von keiner Person gelöst) mit lediglich 952 Daten eine statistisch hoch signifikante globale Prüfgröße und weist auf eine Modellverletzung hin.

Dichotome Kommunikationsskala im Rasch-Modell

Drei Items im zweiten Modul des Neuen Begutachtungsassessments beziehen sich auf die Kommunikation. Es gibt Hinweise darauf, dass die kommunikationsbezogenen Items zukünftig gemeinsam mit den kognitionsbezogenen im zweiten NBA-Modul zu einem Totalscore summiert werden sollen (vgl. BMG [Hg.] 2009c, S. 19). Innerhalb einer Skala dürfen jedoch lediglich eindimensionale Items addiert werden, d. h., es müsste erst der Nachweis erbracht werden, dass die Kommunikation ein Subkonstrukt des latenten Merkmals „Kognition" darstellt. An verschiedenen Stellen des NBA-Handbuchs zeigt sich jedoch, dass der Kommunikationsfaktor unabhängig von der Kognition verstanden wird, so wird z. B. ausdrücklich darauf hingewiesen, dass im Item „Verstehen von Aufforderungen" Hörstörungen und in der Aufgabe „Beteiligung an einem Gespräch" Hör- und Sprechstörungen zu berücksichtigen sind (vgl. Wingenfeld et al. 2008b, C-19).

Die Kommunikationsskala erreicht in den globalen Fitstatistiken von WINMIRA statistisch nicht signifikante Prüfgrößen, für die außerdem eine Chi-Quadrat-Verteilung anzunehmen ist, da die acht möglichen Antwortmuster beobachtet werden. Die Extremscores $r_v = 0$ bzw. 3 treten jedoch 1.232- bzw. 3.086-mal auf, sodass dieser Test nicht genügend zwischen den Fähigkeitsgraden der Personen differenziert. Abbildung 5.5 legt offen, dass die Items nicht sehr unterschiedlich schwierig und, da die Werte um null liegen, eher einfach zu bewältigen sind.

Mit dem Splitkriterium „Setting" ergeben sich im Andersen- und im itemspezifischen Wald-Test statistisch signifikante p-Werte. Der grafische Modelltest bestätigt, dass Personen aus dem stationären Bereich die Items „Mitteilung elementarer Bedürfnisse" und „Beteiligung an einem Gespräch" einfacher bewältigen als jene, die ambulant leben. Umgekehrt lösen Personen, die in den eigenen vier Wänden wohnen, die Aufgabe „Verstehen von Aufforderungen" leichter. Inhaltlich lassen sich die unterschiedlichen Parameterschätzungen in den Items „Mitteilung elementarer Bedürfnisse" und „Beteiligung an einem Gespräch" für die beiden Teilstichproben damit erklären, dass in den stationären Einrichtungen eher rund um die Uhr eine Ansprechperson vorhanden ist als in der eigenen Wohnung. Das könnte dazu führen, dass Menschen im Heim länger fähig bleiben, ihre Bedürfnisse mitzuteilen bzw. ein Gespräch zu führen – es ist stets jemand da, mit dem geredet werden kann. Die Aufgabe „Verstehen von Aufforderungen" ist für ambulant lebende Probandinnen und Probanden eventuell einfacher zu lösen, weil sie offenbar über bessere kognitive Fähigkeiten verfügen als Personen im stationären Bereich. Dies würde allerdings bedeuten, dass neben Kommunikation auch Kognition als latente Variable gemessen

wird, was gegen die Eindimensionalität als Hypothese im Rasch-Modell verstoßen würde.

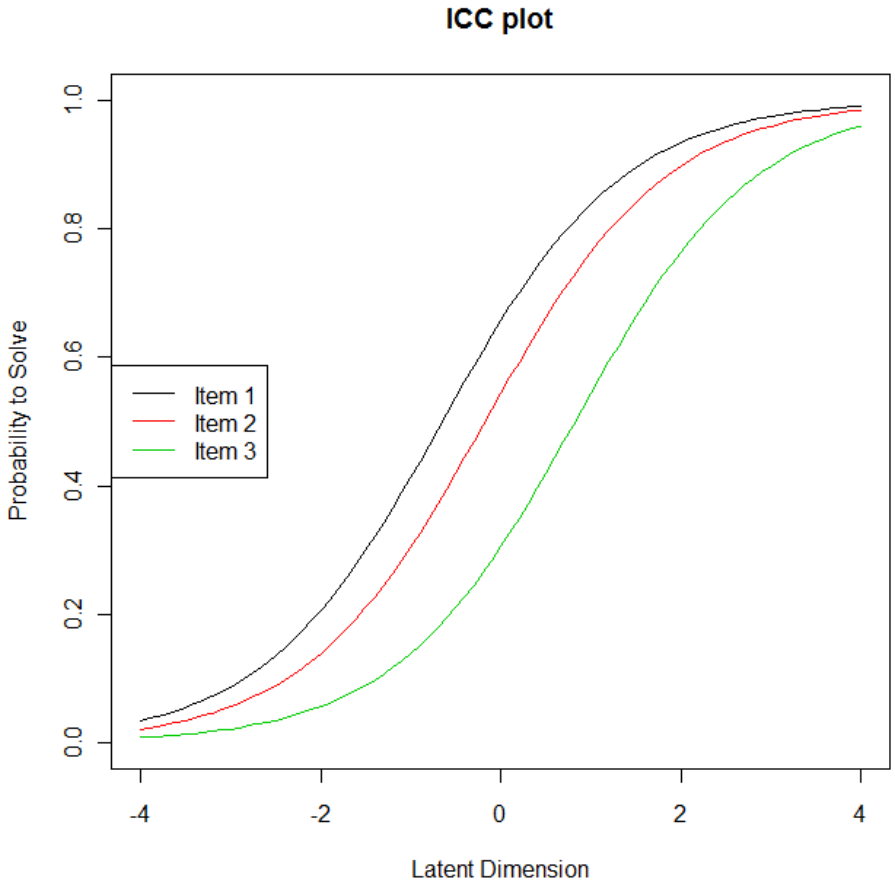


Abb. 5.5 ICC Kommunikation dichotom (eRm in R)

Schlussendlich entsteht in RUMM nach der Exklusion von 85 Prozent der Daten (exkludiert werden die extremen Antwortmuster) ein statistisch hoch signifikantes Ergebnis, sodass die Kommunikationsskala sich als nicht Rasch-valide erweist und die drei Items nicht summiert werden dürfen.

Ordinale Mobilitätsskala im Rasch-Modell

Für die ordinalen Daten der Mobilitätsskala wird nun neben den zentralen Annahmen des Rasch-Modells geprüft, ob die Antwortkategorien Ordinalskalenqualität besitzen. Trifft dies zu, besitzt jede Antwortkategorie einen eigenen Abschnitt auf dem latenten Kontinuum und man kann davon ausgehen, dass mit steigender *Schwellenschwierigkeit* τ_{ix} – eine Schwelle ist der Schnittpunkt z. B. zwischen den Antwortkategorien „Überwiegend unselbständig“ und „Überwiegend selbständig“ in einem Item – die Schwellenwahrscheinlichkeit abnimmt (vgl. Rost 2004, S. 212).

Daraus folgt, dass eine eher unfähige Person im schwierigen Item „Treppensteigen“ mit einer äußerst geringen Wahrscheinlichkeit die 0-Antwort, d. h. die Antwortkategorie „Selbständig“ erreicht. Zudem existieren im ordinalen Rasch-Modell im Gegensatz zum dichotomen variierende Trennschärfen für die Items einer Skala – nicht jedoch für die einzelnen Schwellen der Items (vgl. Rost 2000, S. 37; vgl. Rost 2004, S. 217; vgl. Bühner 2011, S. 577). Da man eine Ordinalskalierung nutzt, um die Zwischentöne der Fähigkeitsgrade zu treffen, eignen sich Items, in denen die mittleren Antwortkategorien breite Abschnitte auf dem latenten Kontinuum haben (vgl. Rost 2004, S. 218). Nachfolgend sind die vierstufigen Antwortkategorien der fünf Mobilitätsitems abgebildet, für die eine Ordinalskalenqualität bestätigt werden kann.

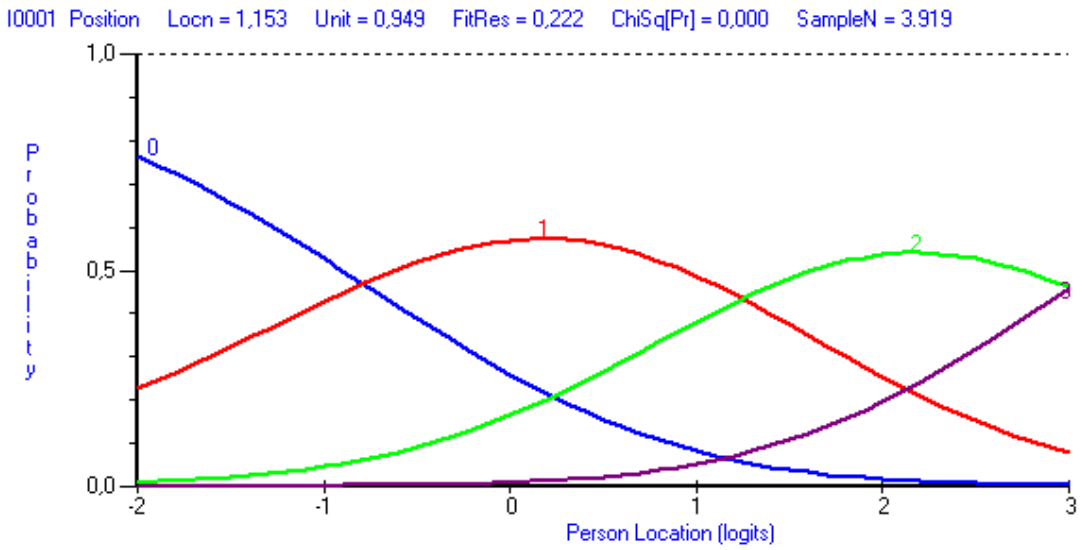


Abb. 5.6 Antwortkategorien im Item „Positionswechsel im Bett“ (RUMM)

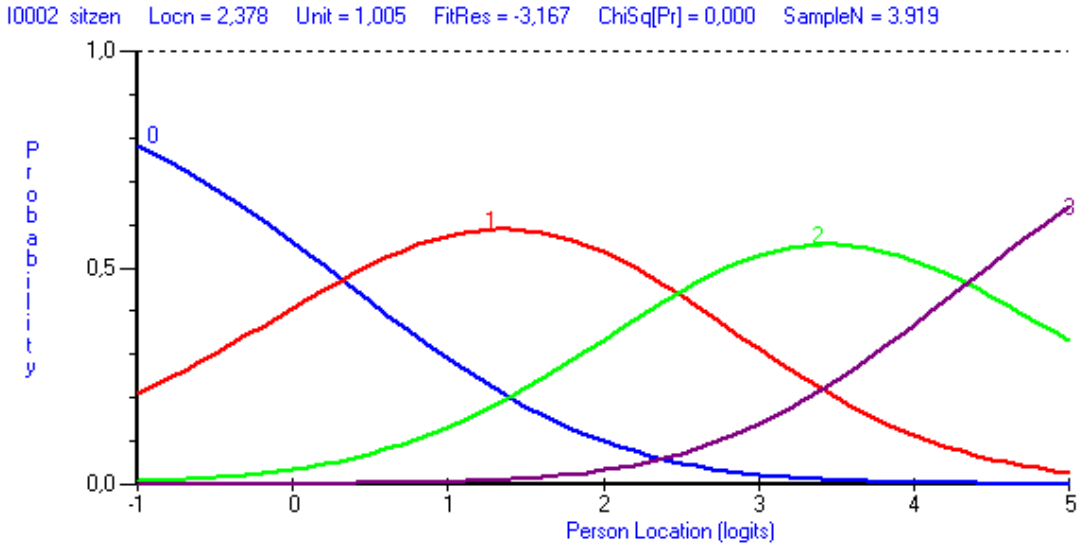


Abb. 5.7 Antwortkategorien im Item „Stabile Sitzposition halten“ (RUMM)

Die monotone blaue Funktion der 0-Antwort kreuzt die nicht-monotone rote Funktion der 1-Antwort bevor diese auf die grüne Funktion der 2-Antwort trifft. Die violette monotone Funktion hat den höchsten Punktwert und darf erst zum Schluss, nachdem sich die anderen Funktionen der Reihenfolge entsprechend einander getroffen haben, die 2-Antwort kreuzen. Die Ordinate zeigt die Wahrscheinlichkeiten der Antwortkategorienbesetzungen entsprechend der Personenfähigkeiten bzw. der Kategorienschwierigkeiten an. Beim Lesen der Abbildungen ist zu beachten, dass die 3-Antwort „Unselbständig“ rechts zu finden ist und damit zum Besetzen höhere Personenfähigkeiten suggeriert. Da Unselbständigkeit gemessen wird, bedeutet ein Wert -2 eine hohe Selbständigkeit, ein Wert von 3 eine hohe Unselbständigkeit.

Die drei letzten Items haben kleine Abschnitte für die Antwort „Überwiegend unselbständig“ (2), wodurch entsprechende Fähigkeitsnuancen eher unzureichend erfasst werden können. Über die Gesamtlängen der beiden mittleren Antwortkategorien hinweg existieren jedoch keine so großen Unterschiede zwischen den fünf Aufgaben, sie umfassen stets mehrere Logits (Maßeinheiten auf der Abszisse). Damit kann insgesamt von eher trennschwachen Items gesprochen werden. Dies ist bei ordinalskalierten Items durchaus wünschenswert, weil sich dann die Antworten der Testpersonen gleichmäßig auf die Kategorien verteilen (vgl. Rost 2004, S. 217ff). Analog zu den dichotomisierten Items ist „Stabile Sitzposition halten“ mit der Lokation = –2,378 das leichteste Item (vgl. Abb. 5.7) und „Treppensteigen“ mit der Lokation = 3,415 die schwierigste Aufgabe (vgl. Abb. 5.10). Die Schwierigkeiten der Items „Aufstehen aus sitzender Position/Umsetzen“ und „Fortbewegen innerhalb des Wohnbereichs“ unterscheiden sich unwesentlich voneinander – ein Item könnte entfernt werden.

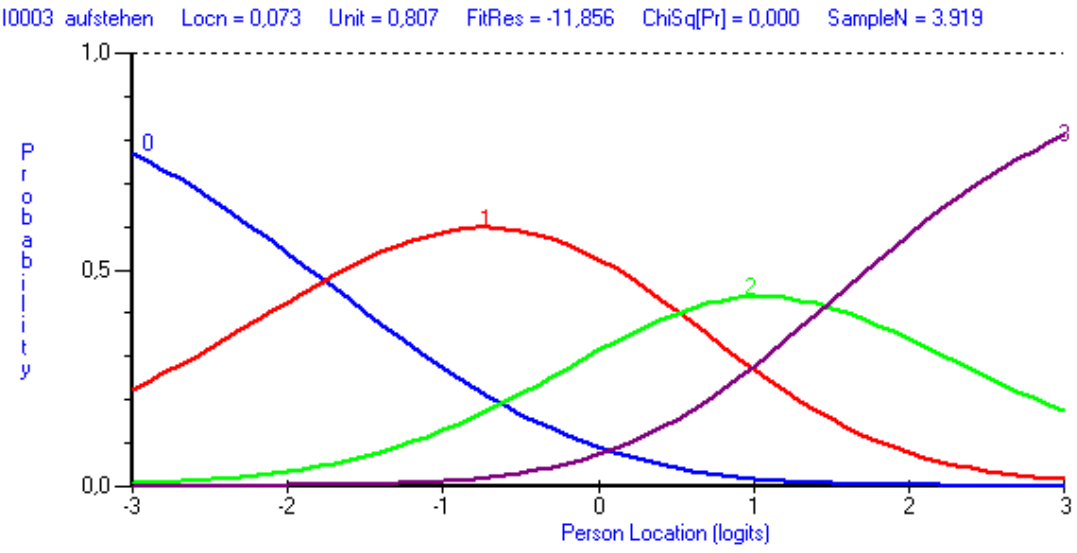


Abb. 5.8 Antwortkategorien im Item „Aufstehen aus sitzender Position/Umsetzen“ (RUMM)

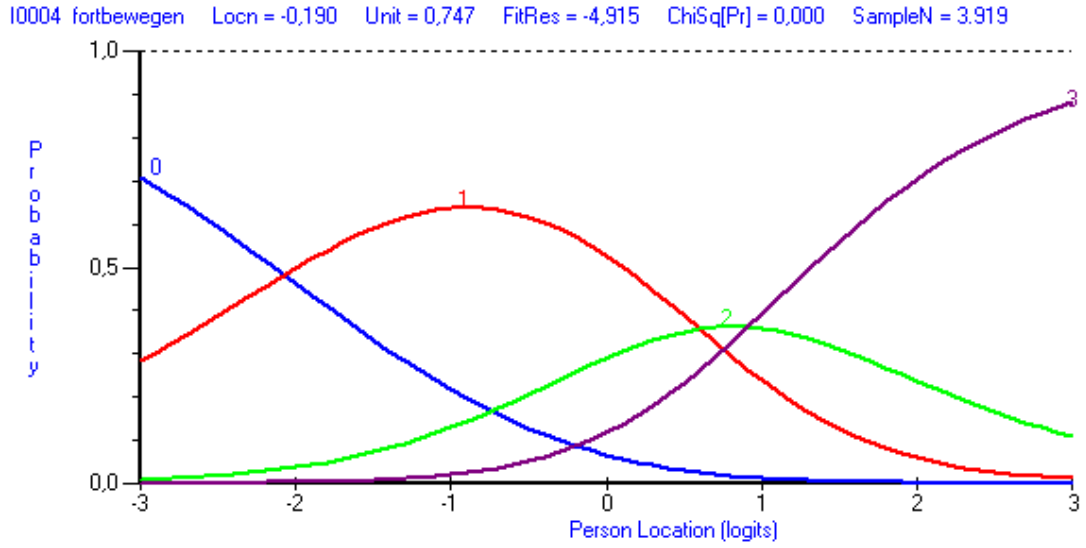


Abb. 5.9 Antwortkategorien im Item „Fortbewegen innerhalb des Wohnbereichs“ (RUMM)

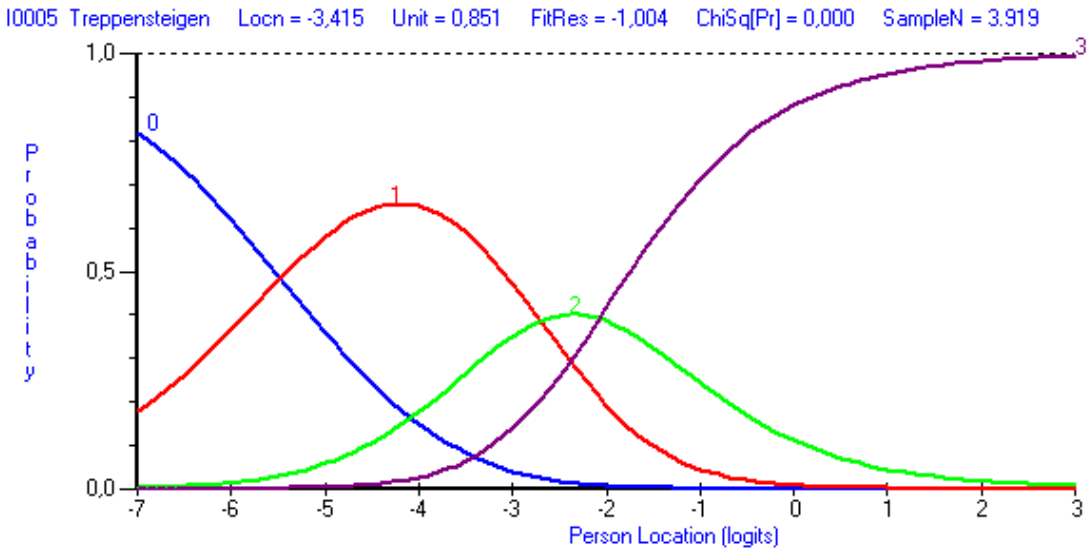


Abb. 5.10 Antwortkategorien im Item „Treppensteigen“ (RUMM)

Die in RUMM beachtete Stichprobengröße beträgt aufgrund des hohen Anteils extremer Antwortmuster lediglich n = 3.919 (vgl. Abb. 5.6 bis 5.10). Die Fit Residuals befinden sich lediglich für die Items „Positionswechsel im Bett“ und „Treppensteigen“ im Referenzbereich ±2 (vgl. Andrich 2004 aufgeführt in Mai 2010, S. 81; vgl. Bond et al. 2001 aufgeführt in Wong et al. 2007, S. 830), wobei für alle Aufgaben statistisch hoch signifikante Chi-Quadrat-Statistiken vorliegen und damit keines der Items für das Modell geeignet ist. Die negativen Werte der Fit Residuals weisen auf „zu gute“ Items hin, die deterministische Antworten produzieren (vgl. RUMM Laboratory [Hg.], S. 27). Schlussendlich besteht ein Total-Item-Chi-Square-Wert, der statistisch hoch signifikant ist und deswegen, wie die Prüfgrößen der globalen Chi-Quadrat-Statistiken nach einem Bootstrapping in WINMIRA sowie dem Andersen-Test mit den Splitkriterien „Mittelwert“ und „Setting“, auf Modellverstöße hinweist. Das Rasch-Modell passt nicht auf die

empirischen Daten, die Items des Moduls „Mobilität“ dürfen auch in ordinaler Form nicht summiert werden.

Ordinale Kognitionsskala im Rasch-Modell

Für die ordinale Kognitionsskala kann ebenfalls ein Ordinalskalenniveau verifiziert werden: Die Schwellenparameter sind für jedes Item geordnet, damit verfügen auch die Antwortkategorien über ein Ordnung bzw. über geordnete Abschnitte entlang des zu messenden Kontinuums. Der nachfolgende Output aus WINMIRA zeigt in Tab. 5.6 die Schwellenparameter τ_{ix} der Antwortkategorien in der Kognitionsskala:

item label	item location	threshold parameters 1	2	3
Pers. erk.	2.49613	-0.951	2.188	6.251
Örtl. Orien.	0.73605	-1.668	0.792	3.083
zeitl. Orie.	-0.00855	-2.728	0.273	2.430
Gedächtnis	0.02605	-3.866	0.108	3.836
mehr. Alltg.	-0.72655	-3.697	-0.587	2.105
Entscheidg.	-1.29056	-3.993	-1.112	1.234

Tab. 5.6 Schwellenparameter der sechs Items der Kognitionsskala

Der Abstand zwischen dem ersten und zweiten Schwellenparameter ergibt die Länge der Antwortkategorie „Überwiegend selbständig“. Gleiches gilt für die Antwortkategorie „Überwiegend unselbständig“ als Distanz zwischen dem zweiten und dritten Schwellenparameter. Ein Rating-Skalen-Modell würde existieren, wenn z. B. über alle Items hinweg eine gleich große 1-Antwort und eine gleich kleine 2-Antwort vorhanden wären. Diese dürften sogar an unterschiedlichen Punkten des latenten Kontinuums pro Item liegen, lediglich die Distanzen müssten stets gleich groß bzw. gleich klein sein (vgl. Brühl 2012, in diesem Band, S. 41). Die höchste Form des Rating-Skalen-Modells wäre das Äquidistanzmodell, in dem alle Items gleich große Antwortkategorien besitzen, welche lediglich verschiedene Lokationen pro Item aufweisen (vgl. Rost 2004, S. 215ff). Das Vorhandensein einer dieser Formen von Schwellenabständen würde die Verwendung der festgelegten Rohwerte 0, 1, 2 und 3 als Punkte für die Antwortkategorien legitimieren (vgl. Tab. 5.1). Jedoch besitzen weder die Punktwerte 1 noch die Punktwerte 2 die gleichen Größen über alle Items hinweg. Auch von einer Äquidistanz der Antwortkategoriegrößen kann nicht gesprochen werden, so dass die

Rohwerte 0, 1, 2 und 3 keine Legitimation für die vier Antwortkategorien aufweisen und bei Eignung des Rasch-Modells nicht verwendet werden dürfen. Dies gilt ebenso für die zuvor vorgestellte ordinale Mobilitätsskala und die noch folgende ordinale Kommunikationsskala.

Alle relevanten globalen Prüfstatistiken erreichen ohne und mit Bootstrapping – es werden lediglich 1.117 Pattern von 65.536 möglichen beobachtet – statistisch hoch signifikante p-Werte.

Die kognitionsbezogenen Items des zweiten NBA-Moduls dürfen nicht summiert werden, da sich ihre Skala als nicht Rasch-valide erweist. Es liegen Verletzungen der zentralen Annahmen des Rasch-Modells vor, wobei besonders auf die settingabhängigen Itemparameterschätzungen im itemspezifischen Wald-Test verwiesen wird: Das Überwinden der Schwellen von einer Antwortkategorie zur nächsten ist je nachdem, ob eine Testperson ambulant oder stationär lebt, ungleich schwieriger bzw. einfacher.

Ähnlichkeiten bestehen zwischen den Ergebnissen lokaler Fitstatistiken, obwohl diese auf sehr unterschiedlichen Rechengrundlagen basieren. So zeigen die nachfolgenden Z_q-Werte in WINMIRA einen Overfit für das Item „Zeitliche Orientierung“ und damit deterministische Antworten an, während sie die Aufgaben „Personen aus dem näheren Umfeld erkennen“ und „Mehrschrittige Alltagshandlungen ausführen“ als Underfits identifizieren und für diese überzufällige Itemantworten konstatieren:

itemlabel	Q-index	Zq	p(X>Zq)	
Pers. erk.	0.0296	2.4477	0.00719-!	Q....!....+
örtl. Orie.	0.0154	-1.0451	0.85202	-....!..Q.+
zeitl. Ori.	0.0131	-1.8770	0.96974+?	-....!...Q+
Gedächtnis	0.0179	-1.0807	0.86008	-....!..Q.+
mehr. Allt.	0.0277	2.0257	0.02140-?	Q....!....+
Entscheidg.	0.0168	-0.7060	0.75990	-....!.Q..+
Info. ver.	0.0217	0.0229	0.49088	-...Q!....+
Risiken er.	0.0214	0.4023	0.34371	-..Q.!....+

Tab. 5.7 Lokale Modellgeltungstests (Q-Index) für acht Items der Kognitionsskala

Dies findet sich in den Infit-Statistiken in eRm wieder, da sich dieselben Items in die entsprechende Randrichtung des Referenzbereichs bewegen:

Itemfit Statistics:

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
Pe.	4270.459	3918	0	1.090	1.131	2.50	7.68
ör.	2771.522	3918	1	0.707	0.772	-11.53	-14.54
ze.	2515.357	3918	1	0.642	0.685	-17.22	-21.43
Ge.	3096.452	3918	1	0.790	0.799	-14.11	-14.45
me.	4230.765	3918	0	1.080	1.102	3.38	6.39
En.	2825.370	3918	1	0.721	0.812	-9.21	-12.28
In.	3422.043	3918	1	0.873	0.902	-7.81	-6.72
Ri.	3447.092	3918	1	0.880	0.938	-4.07	-3.98

Tab. 5.8 Lokale Modellgeltungstests für acht Items der Kognitionsskala

In RUMM verifizieren die Fit Residuals die Hinweise der Z_q-Werte: Die Items „Personen aus dem näheren Umfeld erkennen“ und „Mehrschrittige Alltagshandlungen ausführen“ befinden sich mit positiven Werten außerhalb des Referenzbereichs und deuten auf Underfits hin, während „Zeitliche Orientierung“ mit dem in Abb. 5.11 kleinsten negativen Wert einen Item-Overfit aufweist.

INDIVIDUAL ITEM-FIT for Analysis Name KOGORDV1 - Serial Order

	Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob
1	1	I0001	Poly	2,350	0,034	2,942	3426,25	73,342	8	0,000000
2	2	I0002	Poly	0,757	0,031	-5,978	3426,25	75,457	8	0,000000
3	3	I0003	Poly	0,001	0,030	-10,180	3426,25	100,467	8	0,000000
4	4	I0004	Poly	0,036	0,032	-5,268	3426,25	70,269	8	0,000000
5	5	I0005	Poly	-0,699	0,030	4,307	3426,25	15,201	8	0,055355
6	6	I0006	Poly	-1,268	0,030	-4,846	3426,25	52,742	8	0,000000
7	7	I0007	Poly	0,245	0,031	-1,902	3426,25	39,973	8	0,000003
8	8	I0008	Poly	-1,422	0,031	-1,240	3426,25	44,983	8	0,000000

Abb. 5.11 Fit Residuals Kognition ordinal (RUMM)

Itemeliminationen führen nicht zur Modellgeltung der verschlankten Kognitionsskalen, sodass die H_0-Hypothese, dass das Rasch-Modell auf die empirischen Daten passt, als falsifiziert gilt.

Ordinale Kommunikationsskala im Rasch-Modell

Die ordinale Kommunikationsskala erweist sich ebenfalls als nicht Rasch-valide, denn die globalen Prüfstatistiken weisen allesamt statistisch hoch signifikante Werte auf.

Damit dürfen auch die drei kommunikationsbezogenen Items des zweiten NBA-Moduls nicht summiert werden. Franken stellt in diesem Band ein- bzw. mehrfaktorielle Untersuchungen mit allen elf Items des zweiten NBA-Moduls auf der Basis der klassischen Testtheorie vor (vgl. Franken 2012 in diesem Band, S. 93ff).

Obwohl die Annahmen des Rasch-Modells verletzt werden, lässt sich für die Kommunikationsskala eine Ordinalskalenqualität konstatieren. In den Abbildungen 5.12 bis 5.14 ist deutlich zu erkennen, dass es sich bei Kategoriewahrscheinlichkeiten im Gegensatz zu Schwellenwahrscheinlichkeiten um unbedingte Wahrscheinlichkeiten handelt (vgl. Bühner 2011, S. 578): Teilweise werden an den Horizontalwendepunkten der mittleren Antwortkategorien p-Werte von über 70 Prozent erreicht.

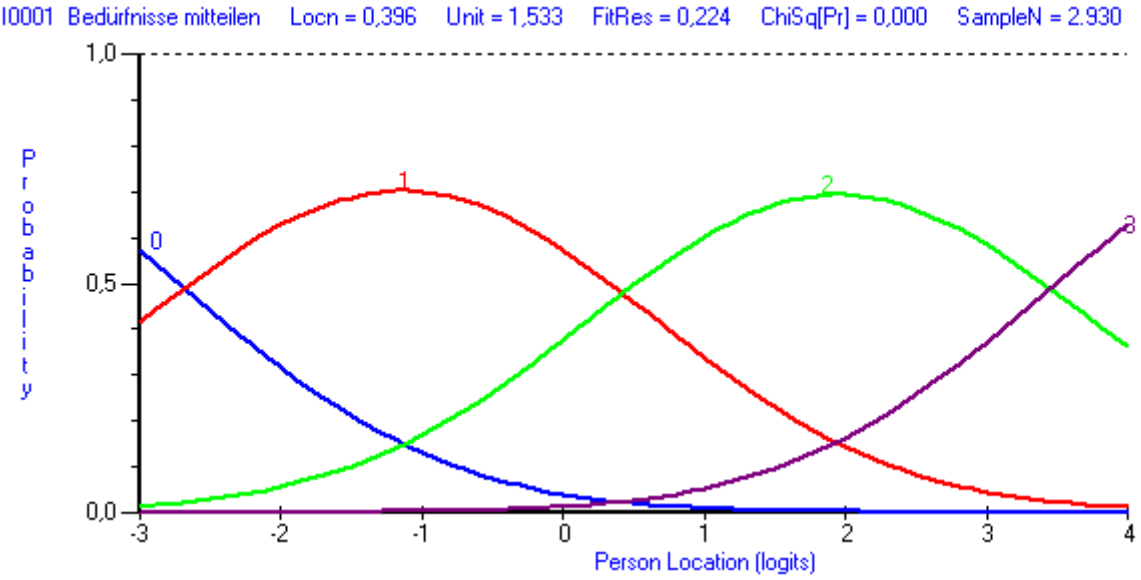


Abb. 5.12 Antwortkategorien im Item „Elementare Bedürfnisse mitteilen“ (RUMM)

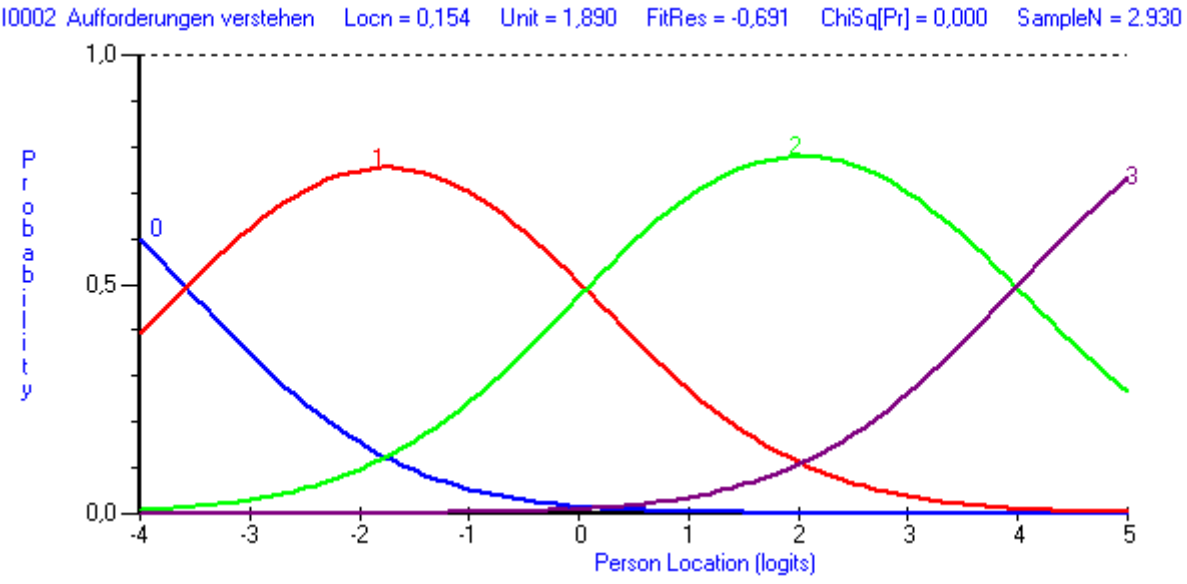


Abb. 5.13 Antwortkategorien im Item „Verstehen von Aufforderungen“ (RUMM)

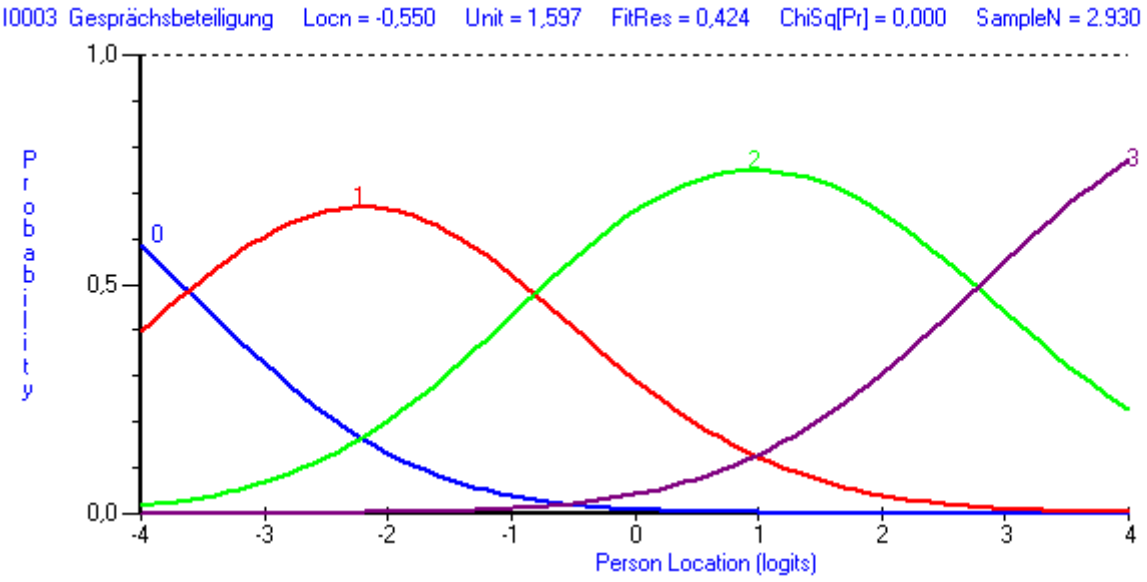


Abb. 5.14 Antwortkategorien im Item „Beteiligung an einem Gespräch“ (RUMM)

ZWISCHENFAZIT

Für das Modul „Mobilität“ sowie die kognitionsbezogenen Items in Abgrenzung von den kommunikationsbezogenen des zweiten NBA-Moduls erweisen sich die Summenwerte als nicht Rasch-valide. Damit gelten die zentralen Annahmen des Rasch-Modells – suffiziente Statistiken, lokale stochastische Unabhängigkeit, spezifische Objektivität, gleiche Trennschärfen und Eindimensionalität – als verletzt. Die Items der jeweiligen Skalen dürfen nicht summiert werden.

Die empirischen Daten weisen für jede Skala einen sehr hohen Anteil an extremen Antwortmustern auf. In den ordinalen Datensätzen besteht stets eine Patternrelation < 1, denn nicht alle möglichen Antwortmuster sind in den Stichproben aufgetreten. Das führt zu einer eventuell fehlenden Chi-Quadrat-Verteilung globaler Prüfgrößen, sodass die Falsifikation einer H_0 durchaus falsch sein kann. Bootstrapverfahren und inhaltliche Analysen der Itemformulierungen können dieses Risiko vermindert haben. Setzt sich ein Test, wie die dichotomisierte Kommunikationsskala zu 85 Prozent aus extremen Personenscores zusammen, so besteht ein dringender Überarbeitungsbedarf der Skala. Testpersonen mit sehr guten bzw. sehr schlechten Fähigkeiten können sonst gar nicht differenziert werden können.

In der Pflege ist es nicht immer vorteilhaft, quantitative Personenvariablen zu erfassen, da pflegerische Phänomene vielfältig bzw. facettenreich sind. Pflegerische Interventionen erreichen ihre Ergebnisse, das zeigt die Empirie, auch nicht in der Quantität der ausgeführten Maßnahmen sondern in ihrer Qualität. Dementsprechend ist die Anwendung eines klassifizierenden Modells, das Testpersonen mit vergleichbaren Merkmalen kategorisiert und Gruppen zuordnet, durchaus sinnvoll.

Die Merkmale „Mobilität“, „Kognition“ und „Kommunikation“ werden deswegen als qualitative Personenvariablen im Rahmen der latenten Klassenanalyse untersucht.

Dichotome Mobilitätskala in der latenten Klassenanalyse

Zur Erklärung der Mobilitätsdaten nach der ersten Dichotomisierungsform (vgl. Tab. 5.1) mit der latenten Klassenanalyse (LCA) werden Personen mit vergleichbaren Mobilitätsmerkmalen in Gruppen bzw. Klassen zusammengefasst. Itemprofile (Pattern) präsentieren dazu sehr aussagekräftige Ergebnisse (vgl. Abb. 5.15). Dabei handelt es sich um die favorisierte Dreiklassenlösung mit vier Items, da sich für die vollständige Mobilitätsskala kein geeignetes klassifizierendes Modell finden lässt.

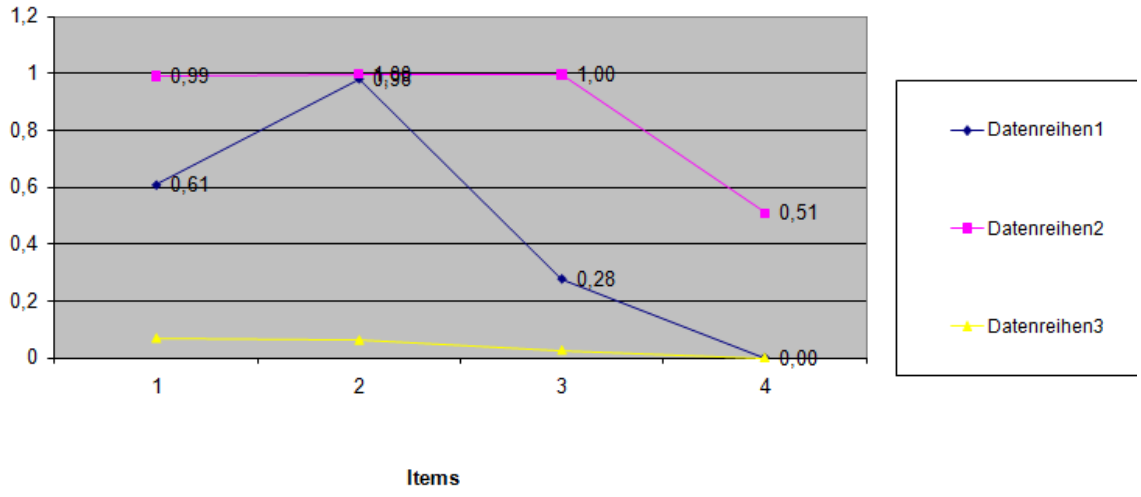


Abb. 5.15 Itemprofile der Dreiklassenlösung mit vier Items Mobilität dichotom (LCA-Analyzer) (ohne das Item „Aufstehen aus sitzender Position/Umsetzen)

In Abb. 5.15 ist ein klassifizierendes Modell mit drei Klassen abgebildet. Auf der Abszisse sind die vier Items „Positionswechsel im Bett“, „Stabile Sitzposition halten“, „Fortbewegen innerhalb des Wohnbereichs“ und „Treppensteigen“ der Klassenlösung dargestellt. Die Ordinate listet die Lösungswahrscheinlichkeiten der Personen einer Klasse pro Item, π_{ig}, auf, denn in der LCA besteht die Hypothese, dass diese stets für alle Personen einer Klasse gelten (vgl. Rost 2004, S. 158; vgl. Gollwitzer in: Moosbrugger, Kelava [Hg.] 2007, S. 285). Bei drei überschneidungsfreien Itemprofilen

existieren drei geordnete Klassen (vgl. Rost 2004, S. 149), wobei die Klasse 2 (pinkfarbene Funktion) die größten Lösungswahrscheinlichkeiten und die Klasse 3 (gelbe Funktion) die niedrigsten π_{ig}-Werte aufweist. Insgesamt verfügen die drei Klassen mit 22,80 Prozent (Klasse 1), 53,58 Prozent (Klasse 2) sowie 23,62 Prozent (Klasse 3) über ungleichmäßige Größen als Folge des zu leichten Tests für die empirische Stichprobe. Daraus erklärt sich weiterhin die geringe Trennschärfe des – wie aus dem Rasch-Modell bekannt ist – sehr leichten Items „Stabile Sitzposition halten“ und sehr schweren Items „Treppensteigen“ über die Klassen hinweg (vgl. Abb. 5.15) (vgl. ebd., S. 375).

Grundsätzlich gelten in der latenten Klassenanalyse die gleichen Annahmen wie im quantitativen Testmodell, z. B. die Itemhomogenität und die lokale stochastische Unabhängigkeit (vgl. ebd., S. 154ff; vgl. Gollwitzer in: Moosbrugger, Kelava [Hg.] 2007, S. 285). Der größte Unterschied zum Rasch-Modell ist, dass nicht mehr vorrangig der Personenscore r_v als suffiziente Statistik interessiert sondern die aufgetretenen Antwortpattern $\underline{x}$ (vgl. Rost 2004, S. 12ff). Die latente Variable konstruiert sich dabei erst durch die Berechnung des qualitativen Modells (vgl. ebd., S. 155). Es folgt eine Patternanalyse der den einzelnen Klassen zugeordneten Antwortmuster, denn in der LCA gilt, dass jede Person einer Klasse zugeordnet wird (disjunkte Klassen) und auch zugeordnet werden kann (exhaustive Klassen) (vgl. Gollwitzer in: Moosbrugger, Kelava [Hg.] 2007, S. 286). Dabei erfolgt die Klassenzuordnung stets mit einer bestimmten Wahrscheinlichkeit $p(g|\underline{x})$ (vgl. Tab. 5.9 und 5.10) und einer Treffsicherheit T als durchschnittliche Höhe der maximalen Zuordnungswahrscheinlichkeiten z. B. über alle Personen einer Klasse hinweg (vgl. Rost 2004, S. 160ff).

Klasse 1		Klasse 2		Klasse 3	
Pattern	$p(g\|\underline{x})$	Pattern	$p(g\|\underline{x})$	Pattern	$p(g\|\underline{x})$
0 1 0 0	0,8281	0 1 1 1	1	0 0 0 0	0,9944
0 1 1 0	0,9083	1 0 1 1	1	0 0 1 0	0,9247
1 0 1 0	0,4130	1 1 0 1	1	1 0 0 0	0,8933
1 1 0 0	0,9809	1 1 1 0	0,8727		
		1 1 1 1	1		

Tab. 5.9 Klassenzuordnungswahrscheinlichkeiten in der Dreiklassenlösung Mobilität dichotom (LCA-Analyzer) ohne das Item „Aufstehen aus sitzender Position/Umsetzen“

Die Pattern in Tab. 5.9 gehören mit einer Treffsicherheit von 78 Prozent zur Klasse 1, 97 Prozent zur Klasse 2 sowie 94 Prozent zur Klasse 3. Das Absinken einer Rate der Treffsicherheit T ist möglich, wenn zwei benachbarte sehr hohe Treffsicherheitsraten existieren (vgl. ebd., S. 161). Für das Antwortmuster 1 0 1 0 besteht mit geringen 41,3 Prozent maximaler Zuordnungswahrscheinlichkeit die Gefahr, dass es möglicherweise doch zu einer anderen Klasse gehört. Würde dieses Pattern der Klasse 3 angehören, könnten die Mitglieder der Klasse 1 sitzen bei gleichzeitiger Unfähigkeit Treppen zu steigen. So jedoch werden sie lediglich durch das letztere Merkmal charakterisiert und weisen mit hoher Wahrscheinlichkeit zwar statische Fähigkeiten (sitzen), jedoch kaum dynamische Fähigkeiten auf (horizontal drehen, aufstehen, steigen) (vgl. Abb. 5.15). Die Klasse 2 ist bezüglich ihrer Pattern und damit ihrer Fähigkeiten eine sehr heterogene, aber sehr leistungsstarke Klasse, da ihre Mitglieder mindestens $r_v = 3$ erreichen und höchstwahrscheinlich sowohl statische als auch dynamische Talente besitzen (vgl. Tab. 5.9 und Abb. 5.15). Hierbei handelt es sich um die *körperlich intakten* Personen, während die Klasse 3 die *körperlich gebrechlichen* Personen umfasst, die weder sitzen noch Treppen steigen können und damit offenbar keine dynamischen und statischen Fähigkeiten aufweisen. Das Pattern 0 0 1 0 in der Klasse 3 ist durchaus realistisch, da ein Mensch mit hoher Querschnittlähmung, gesichert durch einen Rumpfgurt, einen elektrischen Rollstuhl bedienen und sich damit selbständig innerhalb des Wohnbereichs fortbewegen kann.

Diese Dreiklassenlösung mit 69 Prozent Erklärungsrate der empirischen Daten (LR^2 im LCA-Analyzer) ist die derzeit beste Lösung für die dichotomisierten Mobilitätsdaten, alle anderen Klassenlösungen und/oder Itemkombinationen weisen entweder statistisch signifikante Prüfgrößen in den globalen Modellgeltungstests, Überschneidungen in den Itemprofilen, noch ungleichmäßigere Klassengrößen oder geringere Klassenzuordnungswahrscheinlichkeiten bzw. Treffsicherheiten auf. Die oben dargestellte Klassenlösung kann durchaus zukünftig verwendet werden, allerdings sollte sie an einer erneuten und möglichst noch größeren Stichprobe getestet werden. Dies liegt v. a. darin begründet, dass von 16 möglichen Antwortpattern lediglich zwölf aufgetreten sind und somit der Erwartungswert < 1 ist, d.h., die erwarteten Patternhäufigkeiten unterscheiden sich von den beobachteten. Allerdings bekräftigen die Bootstrappings im LCA-Analyzer bzw. den Statistikprogrammen WINMIRA und Latent GOLD die Modellannahme. Hinzu kommt, dass die spezifischen Modellgeltungstests der latenten Klassenanalyse (vgl. Mai 2010, S. 74), dazu gehören der Dissimilarity Index und die bivariaten Residuen (BVR), ebenfalls sehr geeignete Werte produzieren: Während der Dissimilarity Index wiedergibt, wie viel Anteil der erwarteten Patternhäufigkeiten reklassifiziert werden müsste, damit die beobachteten Patternhäufigkeiten das Testmodell perfekt erklären

(vgl. Vermunt et al. 2005a, S. 61; vgl. Langeheine et al. 1995, S. 46), untersuchen die bivariaten Residuen die Verhältnisse der Items paarweise zueinander und treffen somit eine Aussage über das Vorhandensein lokaler stochastischer Unabhängigkeit (vgl. Vermunt et al. 2005a, S. 24 und 72).

Dichotome Kognitionsskala in der latenten Klassenanalyse

Die beste Klassenlösung mit den acht kognitionsbezogenen Items des zweiten NBA-Moduls besitzt eine 30-prozentige Wahrscheinlichkeit (LR^2 im LCA-Analyzer), die empirischen Daten zu erklären. Sie verfügt über drei geordnete Klassen ohne Rangänderungsraten (überschneidungsfreie Itemprofile, vgl. Mai 2010, S. 75ff), eine Patternrelation = 1 und einen Erwartungswert > 1. Die globalen Prüfgrößen sind damit vermutlich Chi-Quadrat-verteilt und bekräftigen damit die Annahme der H_0. Die Dreiklassenlösung enthält die Items „Personen aus dem näheren Umfeld erkennen“, „Zeitliche Orientierung“, „Gedächtnis“ und „Risiken und Gefahren erkennen“.

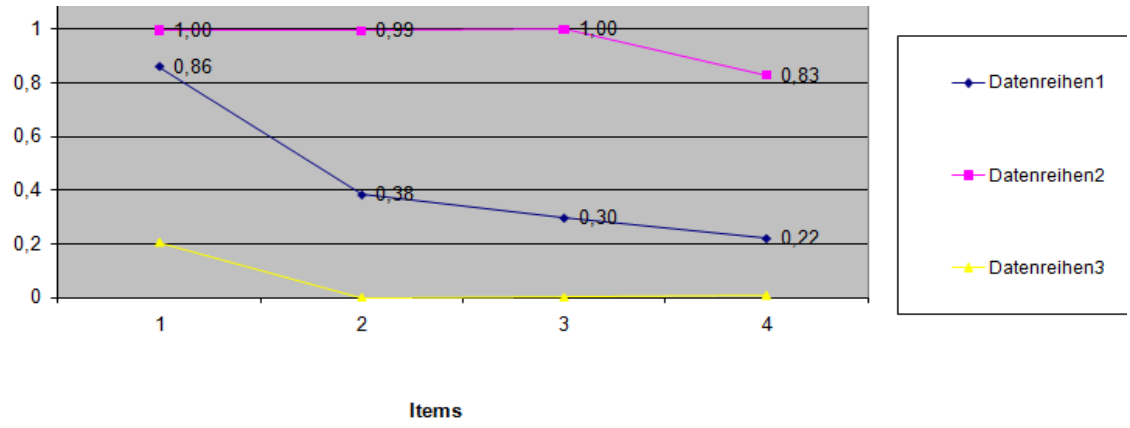


Abb. 5.16 Itemprofile der Dreiklassenlösung mit vier Items Kognition („Personen aus dem näheren Umfeld erkennen“, „Zeitliche Orientierung“, „Gedächtnis“ und „Risiken und Gefahren erkennen“) dichotom (LCA-Analyzer)

Die Klassengrößen sind jedoch sehr unterschiedlich, sodass die Testpersonen ungleichmäßig verteilt sind und umfassen 14,40 Prozent (Klasse 1), 54,67 Prozent (Klasse 2) und 30,93 Prozent (Klasse 3). Die Trennschärfe der vier Items ist über die drei Klassen hinweg durchaus zufriedenstellend, da die π_{ig}-Werte der einzelnen Items überwiegend deutliche Distanzen voneinander aufweisen. Eine Patternanalyse dient der Charakterisierung der Mitglieder dieser Klassen (vgl. Tab. 5.10).

Klasse 1		Klasse 2		Klasse 3	
Pattern	p(g | x)	Pattern	p(g | x)	Pattern	p(g | x)
0 0 1 0	0,9551	0 1 1 1	0,7706	0 0 0 0	0,9729
0 0 0 1	0,5067	1 1 1 0	0,8938	1 0 0 0	0,6011
0 0 1 1	0,9878	1 1 1 1	0,9931		
0 1 0 0	0,9997				
0 1 0 1	0,9999				
0 1 1 0	0,8356				
1 0 0 1	0,9608				
1 0 1 0	0,9710				
1 0 1 1	0,6775				
1 1 0 0	0,9998				
1 1 0 1	0,9966				

Tab. 5.10 Klassenzuordnungswahrscheinlichkeiten in der Dreiklassenlösung Kognition dichotom (LCA-Analyzer)

Die Antwortpattern in Tabelle 5.10 gehören mit einer Treffsicherheit von 90 Prozent zur Klasse 1, 89 Prozent zur Klasse 2 und 79 Prozent zur Klasse 3. Es existieren vier Antwortpattern, nämlich 0 0 0 1, 1 0 1 1, 0 1 1 1 und 1 0 0 0, die keine hohen Klassenzuordnungswahrscheinlichkeiten erreichen. So könnte z. B. das Pattern 0 0 0 1 mit einer Wahrscheinlichkeit von 49 Prozent auch zur Klasse 3 gehören. Die Gefahr, dass das Antwortmuster fälschlicherweise der Klasse 1 zugeordnet worden ist, erscheint sehr groß. Sowohl in Abb. 5.16 als auch in der Tab. 5.10 wird offensichtlich, dass die Fähigkeit, Personen aus dem näheren Umfeld zu erkennen (erstes Item), in allen drei Klassen vorhanden ist und somit offenbar sehr lange erhalten bleibt. Die Klasse 1 erweist sich wie in der Mobilitätsskala als sehr heterogene Gruppe mit $r_v = 1$ bis 3 und umfasst unglücklicherweise elf der 16 Pattern bei einem 14-prozentigen Anteil am Gesamtmodell (vgl. Tab. 5.10). Angehörige dieser Klasse besitzen zwar Fähigkeiten der Personenerkennung, weisen jedoch zeitliche Defizite, Erinnerungslücken und Störungen in der Gefahrenerkennung auf (vgl. Abb. 5.16). Mitglieder der Klasse 2 sind die kognitiv Agilen und mit hoher Wahrscheinlichkeit zur Personenerkennung fähig. Die der Klasse zugeordneten Pattern legen offen, dass die Mitglieder zeitlich orientiert sind und ein funktionierendes Gedächtnis besitzen (vgl.

Tab. 5.10). Der leichte Abfall der Lösungswahrscheinlichkeit im Item „Risiken und Gefahren erkennen“ deutet auf unerwartete Gefahren hin, z. B. im Umgang mit elektrischen Geräten (vgl. Abb. 5.16). Personen der Klasse 3 verfügen über geringe bis keine Fähigkeiten in den vorgenannten Bereichen und gelten als kognitiv gestört. Gemäß der Patternanalyse sind sie von zeitlicher Desorientierung und Gedächtnisverlusten gekennzeichnet (vgl. Tab. 5.10).

Dennoch ist diese Klassenlösung die vorerst geeignetste, um die empirischen Daten zu erklären, das demonstrieren auch die Werte des Dissimilarity Indexes und der bivariaten Residuen. Der dichotome Datensatz resultiert dabei lediglich aus der Verdichtung eines ordinalen Datensatzes. Für das Rasch-Modell ist bereits aufgeführt worden, dass die erste Dichotomisierungsform stets bessere Ergebnisse hervorgebracht hat als die zweite (vgl. Tab. 5.1). Dies lässt sich für die latente Klassenanalyse erneut festhalten. Da der im Modellprojekt definierte Pflegebedürftigkeitsbegriff jedoch keine personelle Hilfe – auch keine leichte – als Selbständigkeit vorsieht (vgl. Wingenfeld et al. 2007, S. 43; vgl. Wingenfeld et al. 2008a, S. 28), müssten erneut bzw. echte dichotome Daten mit den kognitionsbezogenen Items des zweiten NBA-Moduls erfasst werden, um die empfohlene Dreiklassenlösung der Kognitionsskala und auch die der Mobilitätsskala weiterhin als valide erklären zu können.

Dichotome Kommunikationsskala in der latenten Klassenanalyse

Eine Berechnung von Mehrklassenlösungen ist für einen Test mit drei Items nicht möglich, da bereits bei einer Zweiklassenlösung die unabhängigen Parameter des getesteten Modells denen des saturierten Modells entsprechen. Man spricht dann vom gerade identifizierten Modell, das die Daten perfekt erklärt. Da dieses keinen theoretischen Mehrwert besitzt, wird darauf verzichtet (vgl. Bühner 2011, S. 403).

Ordinale Mobilitätsskala in der latenten Klassenanalyse

Im qualitativen Testmodell ordinaler Daten geht es darum, die Klassen qualitativ über Erwartungswertprofile (Wahrscheinlichkeit, mit der die Personen einer Klasse g im Item i die Antwortkategorie x besetzt) zu unterscheiden (vgl. Rost 2004, S. 226). Die Hauptfrage ist weiterhin, ob eine Klassenlösung auf die empirischen Daten passt oder nicht.

Die Berechnungen verschiedener Mehrklassenlösungen der vollständigen Mobilitätsskala sowie mehrerer Itemkombinationen präsentieren jedoch stets statistisch hoch signifikante Prüfgrößen. Die Testpersonen lassen sich nicht sinnvoll

kategorisieren, sodass die H_0, dass ein klassifizierendes Modell auf die ordinalen Daten des Moduls „Mobilität“ passt, verworfen werden muss. Die Verletzung der lokalen stochastischen Unabhängigkeit zwischen den Items zeigt sich beispielsweise deutlich an den Ergebnissen der bivariaten Residuen (BVR) in der Fünfklassenlösung der Mobilitätsskala. Diese sollten eigentlich Werte < 1 aufweisen (vgl. Vermunt et al. 2005b, S. 180).

Indicators	q6	q7	q8	q9	q10
q6	.				
q7	8,5102	.			
q8	0,0690	0,0421	.		
q9	0,6322	0,2225	1,2452	.	
q10	0,1452	1,8046	1,4352	2,9907	.

Abb. 5.17 BVR der Fünfklassenlösung Mobilität ordinal (Latent GOLD)

Gemäß Abb. 5.17 ist jedes der Items in eine paarweise Verbindung involviert, die einen hohen Residualwert produziert. Besonders starke Zusammenhänge existieren zwischen den Items „Positionswechsel im Bett“ und „Stabile Sitzposition halten“ (BVR = 8,5102), gefolgt von „Fortbewegen innerhalb des Wohnbereichs“ und „Treppensteigen“ (BVR = 2,9907).

Ordinale Kognitionsskala in der latenten Klassenanalyse

Keine der untersuchten Mehrklassenlösungen kann die Daten der acht kognitionsbezogenen Items erklären. Somit lassen sich die Testpersonen mit ihren diesbezüglichen Fähigkeiten nicht sinnvoll kategorisieren. Alle Itemkombinationen müssen wegen Modellverletzungen ebenfalls verworfen werden. Die Verbindung der Aufgaben „Entscheidungen im Alltagsleben treffen“, „Sachverhalte und Informationen verstehen“ sowie „Risiken und Gefahren erkennen“ erreicht im Fünfklassenmodell in WINMIRA statistisch nicht signifikante Ergebnisse (vgl. Tab. 5.11), welche jedoch ohne und mit Bootstrapping in Latent GOLD nicht bestätigt werden können.

In Tab. 5.11 weisen sogenannte Informationskriterien, das Bayes Information Criterion (BIC) und das Consistent Akaikes Information Criterion (CAIC), auf eine Nichtpassung der Fünfklassenlösung hin.

Modell	1-Klassenmodell	2-Klassenmodell	3-Klassenmodell	4-Klassenmodell	5-Klassenmodell
t	9	19	29	39	49
log(L)	-20797,46	-16337,76	-14629,80	-14052,99	-14019,51
LR^2	13570,91 p = 0,0000	4651,50 p = 0,0000	1235,58 p = 0,000	81,96 p = 0,0000	15,00 p = 0,3778
Pearson-Chi-Quadrat	23198,06 p = 0,0000	5565,51 p = 0,0000	1327,86 p = 0,0000	76,49 p = 0,0000	13,82 p = 0,4632
df	54	44	34	24	14
Bootstrap (999) Pearson-Chi-Quadrat	p = 0,000	p = 0,000	p = 0,000	p = 0,001	p = 0,065
BIC	41671,70	32837,59	29506,98	28438,65	28457,00
CAIC	41680,70	32856,59	29535,98	28477,65	28506,00

Tab. 5.11 Mehrklassenlösungen mit drei Items Kognition ordinal (WINMIRA)

Beide sind Indikatoren für zu komplexe Modelle insbesondere bei großen Stichproben. Für die probabilistischen Modelle gilt nämlich das Sparsamkeits- bzw. Einfachheitsprinzip, sodass die Informationskriterien „Strafpunkte“ für eine zu hohe Anzahl an Parametern vergeben. Dabei gilt: Je kleiner der Wert, umso besser ist das Modell (vgl. Rost 2004, S. 220; vgl. Borg et al. 2007, S. 364). Tab. 5.11 offenbart BIC- bzw. CAIC-Werte, die von der Vier- zur Fünfklassenlösung bereits wieder steigen, sodass letzteres Modell gegen das oben genannte Prinzip verstößt.

Ordinale Kommunikationsskala in der latenten Klassenanalyse

Die Statistikprogramme WINMIRA und Latent GOLD offerieren stets statistisch signifikante globale Prüfstatistiken in den Mehrklassenlösungen, sodass von Modellverletzungen auszugehen ist. Die Daten der Kommunikationsskala lassen sich nicht mit einer qualitativen Personenvariablen bestimmen.

FAZIT

Weder das verwendete quantitative Testmodell noch das qualitative können die ordinalskalierten empirischen Daten der Module „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessments erklären. Die

Likelihoodwerte der jeweiligen Modelle, die zur Bestimmung des „besten“ Modells eine zentrale Rolle spielen – eine Likelihood gibt an, wie plausibel ein Modell auf die Daten passt unter der Annahme, dass das Modell gilt (vgl. Rost 2004, S. 112) – müssen erst gar nicht miteinander verglichen werden.

Für die dichotomisierten Mobilitäts- bzw. Kognitionsskalen kann jeweils eine Dreiklassenlösung mit vier Items empfohlen werden. Dabei hätten jedoch die Resultate noch besser sein können, außerdem handelt es sich um künstlich verdichtete Daten.

Es wird dringend empfohlen, die beiden Module auf eine theoretisch fundierte Grundlage zu stellen und mit geeigneten Messverfahren die offenbar den Auftraggeberinnen und Auftraggebern bzw. Entwicklerinnen und Entwicklern sehr wichtige Eindimensionalität zu überprüfen.

Die vorhandenen Datensätze könnten anderenfalls mit mehrdimensionalen Messmodellen untersucht und die Skalen bezüglich ihrer Konstruktvalidität überprüft werden. Jedoch würden die Personenscores und Antwortmuster der vorliegenden Stichprobe, die darauf hinweisen, dass die Tests für eine erhebliche Anzahl von Personen zu leicht bzw. zu schwer gewesen sind, diese Ergebnisse beeinflussen. Der sicherste Weg wäre die Überarbeitung der Module bzw. des Neuen Begutachtungsassessments, insbesondere wenn sich die hier vorgestellten Ergebnisse mit anderen Stichproben bestätigen.

LITERATUR

BMG (Hg.) (2009a): Bericht des Beirats zur Überprüfung des Pflegebedürftigkeitsbegriffs. Online verfügbar unter http://www.bmg.bund.de/uploads/publications/Neuer-Pflegebeduertigkeitsbegr.pdf, download am 11.07.2009, zuletzt geprüft am 01.03.2011.

BMG (Hg.) (2009b): Neuer Pflegebedürftigkeitsbegriff. Schluss mit der „Minutenpflege“. In: Heilberufe, H. 03, S. 9. Online verfügbar unter http://www.heilberufe-online.de/pdf.php?url=/archiv/2009/03/8.pdf&nl, zuletzt geprüft am 02.03.2011.

BMG (Hg.) (2009c): Umsetzungsbericht des Beirats zur Überprüfung des Pflegebedürftigkeitsbegriffs. Online verfügbar unter http:// www.bmg.bund.de/uploads/publications/Umsetzungsbericht-Pflegebeduerftigkeitsbegriff_200905.pdf, download am 25.11.2010, zuletzt geprüft am 01.03.2011.

Borg, I.; Staufenbiel, T. (2007): Lehrbuch Theorien und Methoden der Skalierung. 4. Aufl. Bern: Huber.

Bortz, J.; Döring, N. (2006): Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler. 4. Aufl., Nachdr. Heidelberg: Springer-Medizin-Verl.

Brandstätter, E. (2001): Faktorenanalyse oder Rasch-Modell? Eine Kreuzvalidierung am Beispiel des Leistungs-Motivations-Tests. Frankfurt am Main: Lang (Europäische Hochschulschriften Reihe 6, Psychologie, Bd. 686).

Bühner, M. (2011): Einführung in die Test- und Fragebogenkonstruktion. 3. Aufl. München: Pearson Studium.

CDU/CSU-Fraktion im Deutschen Bundestag Arbeitsgruppe Gesundheit (Hg.) 2011: Eckpunkte für eine Pflegereform 2011: Menschlich, bedarfsgerecht, zukunftsfest. Entwurf. Online verfügbar unter: http://www.pro-pflege-selbsthilfenetzwerk.de/Aktuelles/Eckpunkte_fuer_eine_Pflegereform_2011.pdf, download am 20.04.2011, zuletzt geprüft am 02.06.2011.

Gansweid, B.; Wingenfeld, K.; Büscher, A. (2010): Definition der Pflegebedürftigkeit: Konzepte und Verfahren zur Neudefinition des Pflegebedürftigkeitsbegriffs im SGB XI und zur Entwicklung des neuen Begutachtungsverfahrens. In: Sozialer Fortschritt, Jg. 59, H. 2, S. 53–60.

Gollwitzer, M. (2007): Latent-Class-Analysis. In: Moosbrugger, H.; Kelava, A. (Hg.): Testtheorie und Fragebogenkonstruktion. Berlin: Springer, S. 279–306.

Hartig, J.; Frey, A.; Jude, N. (2007): Validität. In: Moosbrugger, H.; Kelava, A. (Hg.): Testtheorie und Fragebogenkonstruktion. Berlin: Springer, S. 135–136.

Langeheine, R.; van Pol, F. de; Pannekoek, J. (1995): Neue Ergebnisse zur Jagodzinski-Langeheine-Debatte in der ZA-Information 1987. In: ZA-Information, Jg. 37, S. 38–50. Online verfügbar unter http://www.ssoar.info/ssoar/files/2010/2186/ za-information_1995_37_38-50.pdf, zuletzt geprüft am 14.03.2011.

Linacre, J. M. (2009): A User’s Guide to WINSTEPS. MINISTEP. Rasch-Model Computer Program. http://ifile.hkedcity.net/1/001/950/public/Secondary/EI0020070012/winsteps.pdf, download am 03.07.2011, zuletzt geprüft am 03.07.2011.

Mai, M. (2010): Das Sturzrisiko von Patienten im Krankenhaus. Entwicklung eines konstruktvaliden Sturzrisikoeinschätzungsinstruments unter dem Einsatz von Modellen aus dem Bereich der probabilistischen Testtheorie. Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors der Pflegewissenschaft (Dr. rer. cur.). 1. Aufl. München: Dr. Hut.

Moosbrugger, H. (2007): Klassische Testtheorie (KTT). In: Moosbrugger, H.; Kelava, A. (Hg.): Testtheorie und Fragebogenkonstruktion. Berlin: Springer, S. 99–112.

Moosbrugger, H.; Kelava, A. (Hg.) (2007): Testtheorie und Fragebogenkonstruktion. Berlin: Springer.

Rost, J. (2000): Haben ordinale Rasch-Modelle variierende Trennschärfen? Eine Antwort auf die Wiener Repliken. Kommentare. In: Psychologische Rundschau, Jg. 51, H. 1, S. 36–37.

Rost, J. (2004): Lehrbuch Testtheorie – Testkonstruktion. 2. Aufl. Bern: Huber.

RUMM Laboratory (Hg.) (2004): Rasch Unidimensional Measurement Models. RUMM 2020. Displaying the RUMM 2020 Analysis. Edith Cowan University.

Schermelleh-Engel, K.; Werner, C. (2007): Methoden der Reliabilitätsbestimmung. In: Moosbrugger, H.; Kelava, A. (Hg.): Testtheorie und Fragebogenkonstruktion. Berlin: Springer, S. 113–133

Strobl, C. (2010): Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis. 1. Aufl. München: Hampp.

Vermunt, J. K.; Magidson, J. (2005a): Technical Guide for Latent GOLD 4.0: Basic and Advanced. Online verfügbar unter http://www.statisticalinnovations.com/products/LGtechnical.pdf, zuletzt geprüft am 01.03.2011.

Vermunt, J. K.; Magidson, J. (2005b): Latent GOLD 4.0. User's Guide. Online verfügbar unter http://www.statisticalinnovations.com/products/latentgold_v4.html, zuletzt geprüft am 01.03.2011.

Wilson, M.; Allen, D. D.; Corser Li, J. (2006): Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. Herausgegeben von Oxford Journals. Medicine Health Education Research. Vol. 21, Suppl 1, S. i4-i18. Online verfügbar unter http://her.oxfordjournals.org/contenS/21/suppl_1/i4.full#sec-4, zuletzt geprüft am 14.03.2011.

Windeler, J.; Görres, S.; Thomas, S.; Kimmel, A.; Langner, I.; Reif, K.; Wagner, A. (2008): Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Abschlussbericht. Hauptphase 2. Endfassung. Institut für Public Health und Pflegeforschung an der der Universität Bremen; Medizinischer Dienst des Spitzenverbandes Bund der Krankenkassen e. V. Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/neueseiten/ag6/downloads/Anhang_zum_Pflegebedurftigkeitsbegriff_SGB_Bericht-Hauptphase-2XI.pdf, download am 27.06.2009, zuletzt geprüft am 01.03.2011.

Wingenfeld, K.; Büscher, A.; Gansweid, B. (2008a): Das neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Projekt: Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Abschlussbericht zur Hauptphase 1: Entwicklung eines neuen Begutachtungsinstruments. Überarbeitete, korrigierte Fassung. Unter Mitarbeit von C. Büker, V. Meintrup und P. U. Menz et al. Institut für Pflegewissenschaft an der Universität Bielefeld; Medizinischer Dienst der Krankenversicherung Westfalen-Lippe. Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/neueseiten/ag6/downloads/Abschlussbericht_IPW_MDKWL_25.03.08.pdf, download am 27.06.2009, zuletzt geprüft am 01.03.2011.

Wingenfeld, K.; Büscher, A.; Gansweid, B. (2008b): Das neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Anlagenband. Projekt: Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Ergänzte und korrigierte Fassung. Institut für Pflegewissenschaft an der Universität Bielefeld; Medizinischer Dienst der Krankenversicherung Westfalen-Lippe. Online verfügbar unter http://www. uni-bielefeld.de/gesundhw/neueseiten/ag6/downloads/Anlagenband_IPW_MDKWL_25.03.08.pdf, download am 27.06.2009, zuletzt geprüft am 01.03.2011.

Wingenfeld, K.; Büscher, A.; Schaeffer, D. (2007): Recherche und Analyse von Pflegebedürftigkeitsbegriffen und Einschätzungsinstrumenten. Überarbeitete, korrigierte Fassung. Unter Mitarbeit von C. Büker, D. Heitmann und T. Seidl et al. Institut für Pflegewissenschaft an der Universität Bielefeld. Online verfügbar unter http://www.uni-bielefeld.de/gesundhw/ag6/downloads/ipw_bericht_pflegebedurftigkeit.pdf, download am 27.06.2009, zuletzt geprüft am 01.03.2011.

Wong, M.; Evans, D. (2007): Students' Conceptual Understanding of Equivalent Fractions. Mathematics: Essential Research, Essential Practice — Volume 2. Proceedings of the 30th annual conference of the Mathematics Education Research Group of Australasia. Online verfügbar unter: http://www.merga.net.au/documents/RP782007.pdf, download am 03.07.2011, zuletzt geprüft am 01.08.2011.

6. Wie lassen sich bessere standardisierte Messinstrumente der Pflege entwickeln?

Neue methodologische Ansätze zur Messung von Pflegebedürftigkeit

Albert Brühl, Katarina Planer, Christian Grebe

Zusammenfassung

Festhalten können wir, dass Pflegebedürftigkeit aktuell nicht valide messbar ist. Das hängt einerseits mit der Abhängigkeit von Pflegebedürftigkeit vom jeweiligen Pflegesetting (ambulant/stationär) ab. Andererseits bereiten vorschnelle Quantifizierungen innerhalb eines solchen komplexen Konstrukts Probleme.

Ein zentraler Gegenstand der Pflegewissenschaft muss es daher zukünftig sein, weitere Schritte hin zu einer validen Messung von Pflegebedürftigkeit zu unternehmen. Um ein besseres Messverfahren für Pflegebedürftigkeit zu entwickeln bestehen unserer Einschätzung nach zwei Ansatzpunkte:

1. Die Entwicklung eines Instruments zur Messung von Pflegebedürftigkeit sollte besser an die aktuelle Arbeitspraxis angebunden werden können, um nicht ein viel zu komplexes (und dennoch invalides) Instrument in die Welt zu setzen. Instrumente müssen praxisrelevant, valide *und* praktikabel sein.
2. Eine Theorie zur Pflegedürftigkeit muss besser begründen können, wodurch sich Pflegebedürftige unterscheiden. Hierzu müssen die bereits vorliegenden sinnvollen theoretischen Konstrukte besser strukturiert und operationalisiert werden.

Im Abschlusskapitel sollen deshalb Verfahren vorgestellt werden, die uns im konkreten Fall der Messung von Pflegebedürftigkeit aber auch allgemein geeignet erscheinen, standardisierte Verfahren zur Messung in der Pflege zu entwickeln und bei

1. der empirischen Fundierung der Instrumentenentwicklung und
2. der Strukturierung und Operationalisierung von Theorien

zu helfen.

Die bislang am Lehrstuhl für Statistik und standardisierte Verfahren der Pflegeforschung der PTHV durchgeführten Studien (Bensch 2011; Franken 2010; Schröder 2010; Hüngsberg 2011) liefern zusammengefasst folgenden Erkenntnisgewinn:

1. *Das ordinale Skalenniveau der Subskalen des NBA weist keine Äquidistanz auf und damit liefern Verfahren der KTT keine validen Ergebnisse bei der Messung von Pflegebedürftigkeit mit diesem Instrument (Franken 2010; Schröder 2010; Bensch 2011). Verfahren der KTT beruhen auf Varianzen und Korrelationen intervallskalierter Daten, linearen, bzw. monotonen Zusammenhängen der Variablen, wahren Werten und Messfehlern sowie der Annahme von Merkmalskonstanz. Diese Grundannahmen der KTT werden bei der Anwendung für pflegebezogene Daten vorausgesetzt, es wird aber nicht geprüft, ob diese Annahmen berechtigt sind.*
2. *Weder bei „Mobilität" noch bei „Kognitiven und kommunikativen Fähigkeiten" handelt es sich um eindimensionale in der vorliegenden Form quantifizierbare Konstrukte. Verfahren der KTT, wie die konfirmatorische Faktorenanalyse, die bivariate Kovarianzmatrizes nutzen, reichen nicht aus, um ein inhaltlich nicht dimensioniertes Instrument zu dimensionieren, weil sich eine „höhere Ordnung" der Komplexität der Wechselwirkungen aller Merkmale (Carstensen 2000, S. 22) mit diesen Methoden nicht erfassen lässt.*
3. *Die Probabilistische Testtheorie alleine stellt keine Verfahren zur theoretischen Dimensionierung des Konstrukts bereit.*
4. *Probabilistische Verfahren ermöglichen die Transformation kategorialer Daten in intervallskalierte Daten. Damit können komplexe Merkmalskombinationen auf der abstrakteren Dimension „Schwierigkeit" quantifiziert werden und damit zu Messungen verwendet werden.*

Die Beziehung zwischen Theorie- und Instrumentenentwicklung

Ein valides Messinstrument ist bestenfalls eine gelungene Operationalisierung der Theorie über ein latentes Konstrukt. Mit Hilfe eines validen Messinstruments oder Testverfahrens lassen sich empirische Daten erheben, mit denen sich die Struktur der Theorie über das latente Konstrukt prüfen lässt. Damit wird die Gültigkeit der Operationalisierung und damit auch der Theorie falsifiziert bzw. bestätigt. Spiegelt sich die Struktur der Theorie nicht in den empirischen Daten wider, kann dies an einer unzulänglichen Operationalisierung der Theorie in das Messinstrument liegen oder aber in theoretischen Annahmen, die sich in der Praxis nicht bestätigen lassen.

Theorie- und Instrumentenentwicklung ohne Nutzung empirischer Daten führt zu Theorien, die ohne empirische Prüfungen auf einem hypothetischen Status verbleiben und zu Messinstrumenten, denen zwar implizit ein theoretischer Hintergrund zugrunde liegt, die sich aber mit empirischen Tests nicht validieren lassen, weil die Theorie unspezifisch operationalisiert wurde. Die Komplementarität von Theorie- und Instrumentenentwicklung unter Nutzung empirischer Ergebnisse wird in einem „Ping-pong"-Prozess deutlich (Abb. 6.1). Die Analyseergebnisse der empirischen Daten, die mit Hilfe des Instruments erhoben wurden dienen der Konkretisierung des Strukturmodells (und damit der Theorie).

Den Kreislauf der Instrumentenentwicklung, in dem man z.B. empirische Ergebnisse zu Mess- und Strukturmodell immer wieder bei der Instrumentenentwicklung berücksichtigt, kann man wie folgt darstellen:

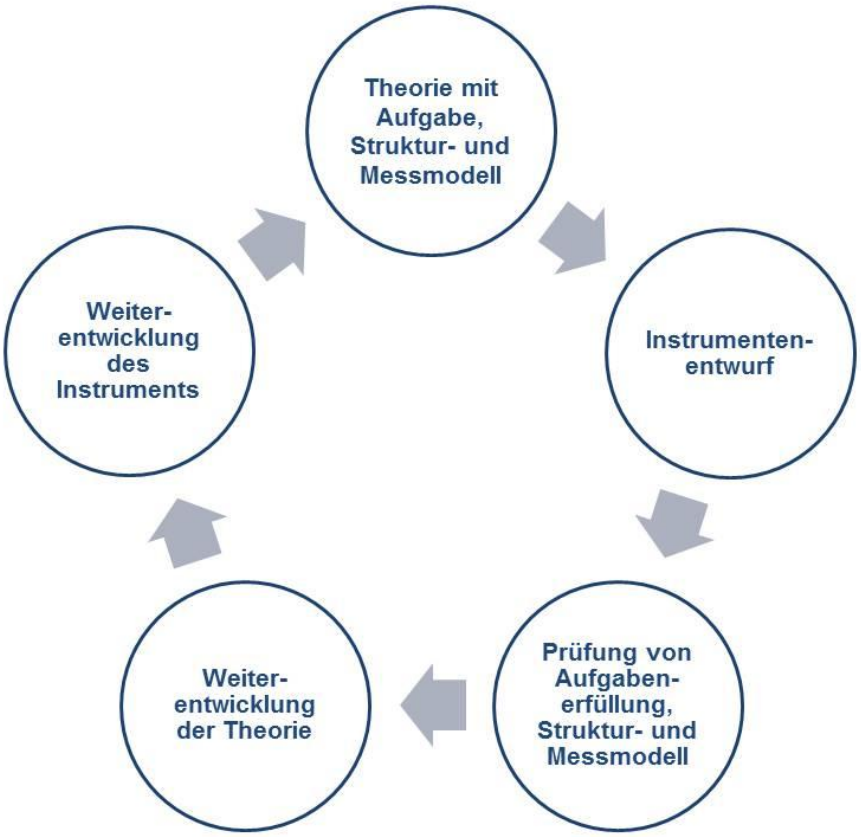


Abb. 6.1 Prozess der Instrumentenentwicklung

Theorie- und Instrumentenentwicklung sind damit als zirkulärer, hermeneutischer Prozess zu verstehen.

Ausgangspunkt ist die gedanklich entwickelte Struktur einer Theorie, die inhaltlich logisch in Dimensionen, Facetten und Komponenten differenziert und strukturiert

werden kann. Diese Strukturtheorie wiederum kann in Items operationalisiert werden, die es ermöglichen, die Theorie mittels statistischer Verfahren auf ihre Gültigkeit hin zu überprüfen. Die Ergebnisse dieser Verfahren liefern Hinweise auf die Richtigkeit der Strukturtheorie und damit der Ausgangsüberlegungen über das Phänomen. Unter der Voraussetzung mit dem Phänomen vertraut zu sein wird es möglich, über die Weiterentwicklung der Theorie die Instrumente zu verbessern, die das theoretische Konstrukt messen und vorhersagen sollen.

Die verfeinerte Theorie wiederum bildet die Grundlage für die Spezifikation des Messinstruments, mit dem erneut empirische Daten erzeugt werden können, die wiederum zur Theorieschärfung genutzt werden können. Dieser Prozess ist solange zu wiederholen, bis ein für seinen Zweck und im entsprechenden Kontext valides Instrument, bzw. eine nützliche Theorie entstanden ist.

Für die Messbarkeit des latenten Konstrukts der Pflegebedürftigkeit stellt sich die Frage, welches Strukturmodell geeignet ist, um validen Messungen des Grads von Pflegebedürftigkeit zu Grunde gelegt zu werden. Die dem NBA impliziten theoretischen Annahmen bieten hierbei hilfreiche Hinweise, die als Ausgangslage für weitere Entwicklungen genutzt werden können.

Die Hypothese, dass Pflegebedürftigkeit in ihrem Ausmaß u.a. auch durch das soziale System des Pflegebedürftigen determiniert ist, ergibt sich aus der Beobachtung der Pflegepraxis und theoretischen Überlegungen. Diese Hypothese kann ebenfalls geprüft werden.

Ausgehend von vorliegenden Forschungsergebnissen und Kritik an bisherigen Forschungsdesigns werden zwei für die Pflegewissenschaft neue Herangehensweisen vorgeschlagen, die unterschiedliche methodologische Ansätze mit dem Ziel der parallelen Theorie- und Instrumentenkonstruktion kombinieren.

Für die Theorie- und Instrumentenentwicklung wird mit beiden methodologischen Ansätzen folgender Erkenntniszuwachs erwartet:

- Konkretisierung der Dimensionalität und Struktur des latenten Konstrukts der Pflegebedürftigkeit
- Beschreibung der Voraussetzungen zur Indexbildung zur Messung von Pflegebedürftigkeit
- Bedeutung von Kognition und Motorik für Pflegebedürftigkeit
- Bedeutung der Interaktionen des sozialen Systems bei der Erfassung von Pflegebedürftigkeit
- Implikationen für die Themenbereiche: Pflegepersonalbemessung, Messung von Pflegequalität und Prävention von Pflegebedürftigkeit.

Erster Ansatz:

a) Empirische Fundierung der Instrumentenentwicklung

Nonparametrische Regression mit Methoden des maschinellen Lernens

Wenn man die Aufgabe Pflegebedürftigkeit valide zu quantifizieren aktuell nicht erfolgreich erledigen kann, so kann es sinnvoll sein, zuerst einmal zu versuchen, Pflegeaufwand zu erklären. Natürlich ist dies dann keine Messung von Pflegebedürftigkeit, aber in der Erklärung von aktuell in der Pflege produziertem Aufwand liegt die Chance, den Anteil der Varianz der Leistungszeit darin zu identifizieren, der mit Unterschieden in der Pflegebedürftigkeit zusammenhängt. Davon getrennt können die Einflussgrößen identifiziert werden, die nicht direkt mit Pflegebedürftigkeit in Verbindung stehen, trotzdem aber Pflegeaufwand erklären.

Das NBA erhebt den Anspruch, den Grad an Selbstständigkeit eines Versicherten zu messen und nicht, wie im bisherigen Verfahren, Zeitwerte. Ein Rückbezug auf den Zeitaufwand oder die Kosten der Pflege wird allerdings spätestens bei einer Anwendung des Instruments im Kontext der Personalbemessung notwendig, da dies die Kalkulationsgrößen sind, die zur Personalbemessung benötigt werden und die ein entsprechendes Instrument folglich vorherzusagen hat.

Für die im Folgenden vorzustellenden Methoden ist eine solche Zeit- oder Kostenmessung die Voraussetzung, der Übersichtlichkeit halber ist von hier an nur noch von einer Zeitmessung die Rede. Diese Messgröße wird nicht erst verwendet, um ein fertig gestelltes Modell zu validieren. Vielmehr erfolgt die Nutzung der Zeitwerte bereits bei der Modellierung mit dem Ziel, das Modell von vornherein so zu entwickeln, dass die Zeiten optimal erklärt werden. Methodisch bedeutet dies, dass die Zeitmessung als abhängige, zu erklärende Variable fungiert, die durch Prädiktorvariablen, nämlich zustandsbezogene Assessmentdaten der Klienten, erklärt werden soll. Es handelt sich folglich um ein multivariates Regressionsproblem.

Als weit verbreitetes und viel genutztes Standardverfahren der klassischen Testtheorie würde sich also prinzipiell die multivariate lineare Regression anbieten. Als parametrisches Verfahren ist der Einsatz dieser Methode allerdings an eine Reihe von Modellannahmen über das zu entwickelnde lineare Modell geknüpft. Diese Annahmen beziehen sich auf die funktionelle Spezifikation des Modells, auf die Störgröße und auf

die Prädiktorvariablen (vgl. von Auer 2003). So muss es sich insbesondere um einen linearen Zusammenhang zwischen Prädiktorvariablen und der abhängigen Variable handeln und die Prädiktorvariablen müssen frei von perfekter Multikollinearität[61] sein. Zudem müssen die Störgrößen und damit die Residuen homoskedastisch[62] sein, dürfen nicht autokorreliert sein und müssen einen Erwartungswert von 0 haben. Auch sollten sie normalverteilt sein.

Diese Modellannahmen werden in der Pflege häufig verletzt.

Zusammenhänge von Zustandsvariablen der Klienten und aufwandsbezogenen Messgrößen sind häufig nicht linear, was Smith et al. beispielsweise für die ADLs zeigen (vgl. Smith 1987).

Auch sind unserer Erfahrung nach die Residuen nur selten normal verteilt und homoskedastisch.

Eine Alternative zur linearen Regression stellen parameterfreie Verfahren dar, bei denen die genannten Modellannahmen nicht erfüllt sein müssen. Für Regressionsprobleme sind insbesondere drei Ansätze interessant: Regressionsbäume, Multivariate Adaptive Regression Splines (MARS) und Ensemblemethoden. Alle genannten Ansätze sind dem maschinellen Lernen zuzurechnen, es handelt sich also um explorative strukturentdeckende Ansätze.

Bei Methoden des maschinellen Lernens wird ein Modell mittels einer Lernstichprobe angelernt, es entsteht also auf Basis der empirischen Daten der Lernstichprobe. Die Validierung erfolgt dann entweder durch die Anwendung des Modells auf eine Teststichprobe oder aber durch Kreuzvalidierung. In diesem Fall wird keine separate Teststichprobe benötigt.

Der weitaus größte Teil der international für die Langzeitpflege entwickelten Fallgruppensysteme basiert methodisch auf Entscheidungsbäumen, genauer: auf Regressionsbäumen[63]. Die international in den 1980er und 1990er Jahren entwickelten Fallgruppensysteme setzen dabei zumeist auf Algorithmen, die auf dem Automatic Interaction Detector (AID) von Sonquist basieren (vgl. Sonquist et al. 1964), vor allem auf AUTOGRP (vgl. Mills et al. 1976). Dieses wurde zudem beispielsweise auch bei der Entwicklung der ersten Diagnosis Related Groups (DRG) genutzt. Heute stehen modernere Algorithmen für Regressionsbäume zur Verfügung, insbesondere CART (vgl. Breiman et al. 1998) und Conditional Inference Trees (vgl. Hothorn et al. 2006).

[61] Perfekte Multikollinearität würde bedeuten, dass zwischen mehreren der Prädiktorvariablen ein linearer Zusammenhang bestünde, der über alle Fälle gültig ist.

[62] Homoskedastizität bedeutet, dass die Störgröße für alle Beobachtungen eine konstante Varianz aufweist, die Residuen also konstant um die Regressionsgerade streuen

[63] beispielsweise alle drei Generationen der Resource Utilization Groups (RUGs)

Allen Regressionsbaumalgorithmen gemein ist das generelle Vorgehen, das als rekursives Partitionieren bezeichnet wird (vgl. Strobl et al. 2009). Dabei wird eine Gruppierung der Stichprobe dergestalt vorgenommen, dass die entstehenden Subgruppen hinsichtlich der Ausprägung der abhängigen Variable in sich möglichst homogen und untereinander möglichst unterschiedlich sind. Mit den entstehenden Subgruppen wird ebenso verfahren, bis ein Stopp-Kriterium erreicht wird[64]. Der Vorhersagewert einer Gruppe entspricht dem arithmetischen Mittel der abhängigen Variable (also der Zeitwerte) aller Fälle der betreffenden Gruppe.

Der CART- Algorithmus, in der Programmiersprache R als rpart umgesetzt, partitioniert mittels der Kleinste-Quadrate-Methode. Es wird also genau der Split ausgewählt, der die niedrigste Residuenquadratsumme aufweist. Dazu wird eine Suche über alle Möglichkeiten durchgeführt, also über alle Prädiktorvariablen und alle möglichen Splitpunkte ihrer Werteausprägungen. Ein solches Vorgehen wird im maschinellen Lernen als exhaustive search bezeichnet.

Conditional Inference Trees (cTree) verwendet anstelle der Methode der kleinsten Quadrate Signifikanztests auf der Basis von Permutationen. Bei der Partitionierung wird ebenfalls eine exhaustive search durchgeführt, die Auswahl für den nächsten Split fällt auf jenen mit der höchsten Signifikanz.

Der Algorithmus von cTree vermeidet durch sein Signifikanzkonzept eine Überanpassung des Modells an die Daten der Lernstichprobe (overfitting). CART, respektive rpart, führt je nach Wahl des Stoppkriteriums dagegen fast zwangsläufig zu einer solchen Überanpassung, im Extremfall könnte jeder Fall eine eigene Fallgruppe bilden. Um dem zu begegnen, wird in einem zweiten Schritt der Baum zurückgeschnitten (pruning). Als Kriterium dienen hier Kreuzvalidierungen, die Wahl fällt auf den Baum mit dem niedrigsten Kreuzvalidierungsfehler[65].

Unsere ersten Erfahrungen mit rpart und cTree zeigen, dass eine Varianzaufklärung der Leistungszeit pro Bewohner pro Tag von ca. r^2=>0,55 mit reinen Zustandsvariablen der Bewohner als Prädiktoren möglich ist.

Der große Vorteil der Regressionsbäume liegt darin, dass die entstehenden Bäume intuitiv verständlich sind. Anhand des Baumes kann sehr einfach die Entscheidungsregel nachvollzogen werden, wie und warum ein Fall welcher Fallgruppe zugeordnet wurde, also auf Basis welcher Fähigkeiten, Einschränkungen und sonstiger Charakteristika des Falles.

[64] Das kann unter anderem die Anzahl der Fälle in einer Subgruppe sein oder auch die Unterschreitung eines Mindestmaßes an zusätzlich gewonnener Varianzaufklärung

[65] bzw. auf den 1- SE- Baum, also jenen Baum, mit dem geringsten Kreuzvalidierungsfehler, der innerhalb einer Standardabweichung der Minimums liegt

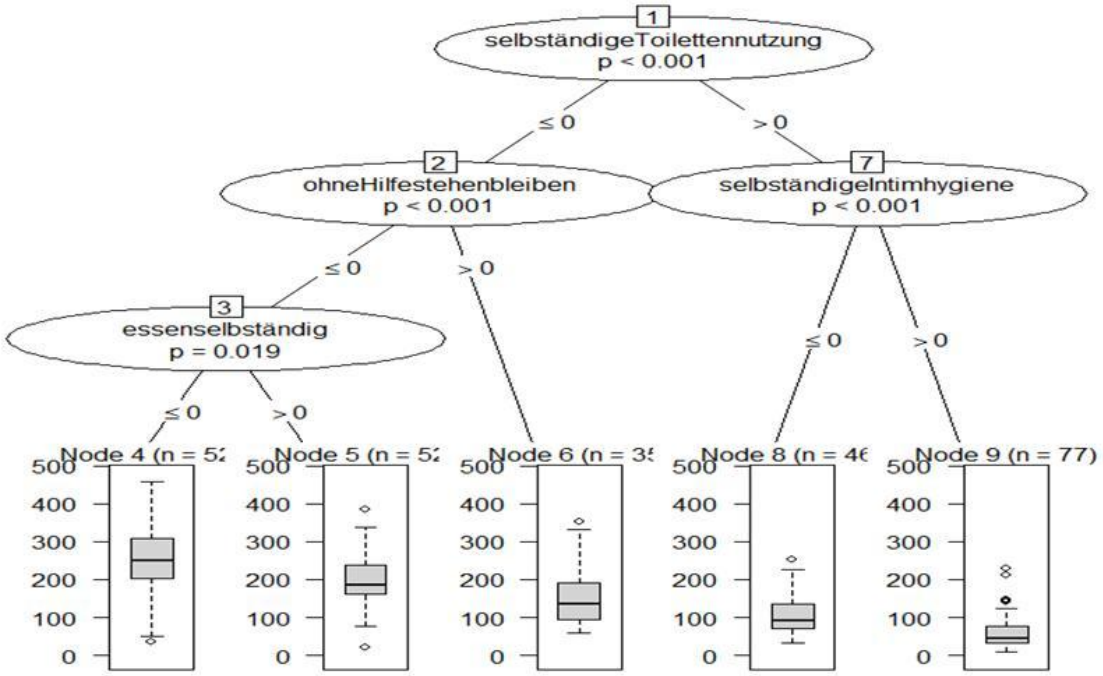


Abb. 6.2 cTree Regressionsbaum

Abb. 6.2 zeigt ein Beispiel für einen Regressionsbaum auf der Basis von cTree. Die Fälle werden fünf Gruppen zugeordnet, die sich hinsichtlich ihres durchschnittlichen zeitlichen Pflegeaufwands und in Bezug auf die Charakteristika der Fälle unterscheiden. Der linken Gruppe wird zugeordnet, wer Toilettengänge nicht mehr selbstständig bewältigt, nicht mehr selbständig stehen und nicht selbstständig essen kann. Es ist ersichtlich, dass dies die Gruppe mit dem höchsten Fallgewicht ist, sie hat den höchsten durchschnittlichen zeitlichen Pflegeaufwand der ihr zugeordneten Fälle. Demgegenüber wird der rechten Gruppe zugeordnet, wer Toilettengänge selbstständig bewältigt und seine Intimhygiene selbstständig übernehmen kann. Diese Gruppe weist in diesem Modell das niedrigste Fallgewicht auf. An diesem Beispiel wird deutlich, wie einfach und transparent die Gruppierungsregeln von Regressionsbäumen sind.

Die Schwäche der Regressionsbäume liegt in der eher geringen Stabilität der Modelle. Bereits kleine Änderungen in der Lernstichprobe können die berechneten Bäume verändern. Dies zeigt sich auch daran, dass in unseren Untersuchungen auf Einrichtungsebene für verschiedene Pflegeeinrichtungen stets auch unterschiedliche Bäume entstehen. Dieser Effekt dürfte sich aber in größeren Stichproben weitaus geringer auswirken, die Modelle sind dann aufgrund häufigerer Messwiederholungen deutlich robuster.

Um diesem Problem zu begegnen, wurden so genannte Ensemblemethoden entwickelt, insbesondere Bagging (vgl. Breiman 1996), Boosting (vgl. für L2- Boosting

Bühlmann, Yu 2003) und die Random Forests (vgl. Breiman 2001). Diesen Ansätzen gemein ist das Vorgehen, nicht nur ein Modell (beispielsweise, aber nicht zwangsläufig, einen Regressionsbaum) zu berechnen, sondern sehr viele. Dabei wird durch verschiedene Strategien dafür gesorgt, dass sich die Einzelmodelle unterscheiden. Wurden alle Einzelmodelle berechnet, wird der betreffende Fall von jedem Einzelmodell separat einer Fallgruppe zugeordnet. Das arithmetische Mittel aller Einzelmodelle ist dann der Vorhersagewert des gesamten Ensembles.

Die einfachste dieser Methoden ist das Bagging (kurz für bootstrap aggregating). Dabei werden bei jeder Iteration, also für jedes Einzelmodell, randomisierte Teilstichproben aus der Lernstichprobe mit Zurücklegen gezogen. Die Fälle, die jedem Einzelmodell zugrunde liegen, unterscheiden sich also und damit unterscheiden sich auch die Einzelmodelle.

Random Forests gehen noch einen Schritt weiter: sie kombinieren Bagging mit einer ebenfalls randomisierten Auswahl der Prädiktorvariablen. Die Einzelmodelle unterscheiden sich dadurch stärker voneinander. Wie der Name der Methode nahelegt, nutzt Random Forest für die Einzelmodelle Entscheidungsbäume, nämlich CART[66]. Bagging dagegen ist mit beliebigen Methoden für die Einzelmodelle kombinierbar. Es ist allerdings nur für solche Methoden effektiv, deren Modelle instabil sind (vgl. Breiman 1996, S. 124). Daher erzielt Bagging beispielsweise mit Bäumen gute Ergebnisse, nicht aber mit MARS- oder in Kombination mit linearer Regression (vgl. Vogt 2000).

Boosting stellt einen Oberbegriff für verschiedene Algorithmen dar, deren gemeinsame Grundidee es ist, bei den Iterationen jeweils stärkeres Gewicht auf jene Fälle zu legen, die vom entstehenden Modell noch nicht zufriedenstellend erklärt werden. Für Regressionsprobleme steht das L2-Boosting (vgl. Bühlmann, Yu 2003) zur Verfügung, bei dem mit jeder Iteration jeweils die momentanen Residuen des entstehenden Modells erklärt werden sollen. Die Modellfunktion jeder Iteration wird dem entstehenden Modell additiv hinzugefügt. L2-Boosting kann u.a. in Verbindung mit Regressionsbäumen eingesetzt werden.

Allen Ensemblemethoden gemein ist, dass sie kaum bis gar nicht transparent darstellbar sind und damit dem Anwender gegenüber als “black box“ erscheinen. Es ist aus diesem Grund auch schwierig, diese Methoden sinnvoll in ein Instrument zur Einschätzung von Pflegebedürftigkeit oder zur Personalbemessung einfließen zu lassen.

Sie haben aber dennoch ihren Wert. Zum einen zeigen sie auf, welche Performance der Modelle im Vergleich zu den jeweiligen Methoden der Einzelmodelle erreichbar ist.

[66] Auch für die Conditional Inference Trees existiert eine entsprechende Ensemblemethode, die Conditional Inference Forests (cForest)

Dies erleichtert die Beurteilung der Performance etwa von Regressionsbäumen relativ zum Ensemble. Zum anderen bieten Random Forests (und auch cForest) eine Bewertung der Prädiktorvariablen hinsichtlich ihrer Wichtigkeit im Ensemble an, was eine Möglichkeit zur Variablenselektion eröffnet.

Einen anderen Ansatz nonparametrischer Regression stellen die Multivariate Adaptive Regression Splines (MARS) dar (vgl. Friedman 1991). Diese sind auf den ersten Blick ein anderer Ansatz als das bis hierhin thematisierte rekursive Partitionieren. Auf den zweiten Blick stellt sich MARS aber als eine Verallgemeinerung der schrittweisen linearen Regression sowie als eine Modifikation von CART dar (vgl. Hastie et al. 2009, 321).

MARS führt wie die oben bereits genannten Methoden ebenfalls eine exhaustive search durch, prüft also die Eignung für jeden Prädiktor und für jede gemessene Ausprägung. Die Basisfunktionen sind bei MARS linear, MARS-Modelle sind in der Folge schrittweise lineare Modelle. Es wird jeweils die Basisfunktion mit den geringsten Residuenquadraten ausgewählt.

MARS-Modelle lassen sich auf die gleiche Weise interpretieren wie multivariate lineare Regressionsmodelle. Auch MARS-Modelle setzen sich additiv aus Produkten von Merkmalsausprägungen mit einem Koeffizienten β zusammen. Die β-Koeffizienten von MARS-Modellen sind ebenfalls analog zu den Steigungsparametern der linearen Regressionsmodelle interpretierbar, aus ihnen lässt sich ablesen, wie sich eine Veränderung der Prädiktorvariablen auf den Vorhersagewert für die abhängige Variable auswirkt.

Unsere ersten Erfahrungen mit dem MARS- Algorithmus zeigen, dass sich gemessene Pflegezeiten besser als mit Regressionsbäumen erklären lassen, es wird eine Varianzaufklärung r^2 >0,55 erreicht. MARS bietet aber zudem die Möglichkeit, Interaktionen zwischen je zwei (oder mehreren) der Prädiktorvariablen zuzulassen. In dieser Variante sind unserer Erfahrung nach sogar Varianzaufklärungen von r^2 >0,70 erreichbar.

Unsere momentane Einschätzung geht in die Richtung, dass MARS mit zugelassenen Interaktionen zwischen bis zu drei Prädiktorvariablen unter den Methoden des maschinellen Lernens am vielversprechendsten ist. Die Vorteile von MARS sind zum einen die enorme Flexibilität und zum anderen die vergleichsweise leichte Interpretierbarkeit analog zu multivariaten linearen Regressionsmodellen. Regressionsbäume sind in dieser Hinsicht zwar noch einmal deutlich intuitiver und einfacher, aber sie sind eben auch instabiler und erreichen keine so hohe Varianzaufklärung.

Über nicht-parametrische Verfahren hinaus werden wir in weiteren Projekten so genannte Mehr-Ebenen-Analysen einsetzen, die berücksichtigen, dass z.B. Bewohner in der stationären Langzeitpflege nicht als Individuen in unsere Untersuchungen eingehen, sondern hierarchisch in Wohnbereichen und in Einrichtungen zusammengefasst sind. Mehr-Ebenen-Verfahren beinhalten im Kern lineare Regressionen, die diese auf verschiedenen Ebenen einsetzen. Wir werden diese Methode nutzen, um die Ebenen der Heime und der Wohnbereiche zu modellieren und sehen, ob sich ein solches im Kern parametrisches Verfahren im Vergleich zu den nicht-parametrischen besser in der Vorhersage von Fallzeiten behauptet. Interessant sind Mehr-Ebenen-Analysen eventuell auch, wenn es gilt Veränderungsmessungen zu modellieren, in denen die verschiedenen Messzeitpunkte als Variable in ein solches Modell eingehen.

Zu beachten ist bei der Erklärung von Pflegeaufwand stets, dass Zeiten oder Kosten empirisch unter realen Bedingungen gemessen werden. Die Werte bilden somit das Leistungsgeschehen in Hier und Jetzt ab, das durch die aktuellen gesellschaftlichen Umfeldfaktoren beeinflusst wird. Die gemessenen Zeiten oder Kosten können somit auch ein nicht wünschenswertes Leistungsgeschehen auf der Basis nicht adäquater Pflegequalität widerspiegeln.

Zweiter Ansatz:

b) Weiterentwicklung der theoretischen Definition von Pflegebedürftigkeit: Die Kombination von Facettentheorie und probabilistischer Testtheorie

Die bislang undefinierten Relationen der Subkonstrukte des Konstrukts der Pflegebedürftigkeit erschweren die Entwicklung eines Instruments zur Messung von Pflegebedürftigkeit. Die Möglichkeit durch Versuch und Irrtum vielfältige theoriebasierte Itempools mittels faktorenanalytischer Verfahren (vgl. Brühl 2012, in diesem Band, S. 34) auf Dimensionalität und Homogenität zu testen, scheiden neben dem fragwürdigen Verhältnis von Aufwand und Erfolg aufgrund der fehlenden Voraussetzungen[67] der empirischen Daten aus.

Beide Probleme ließen sich aus heutiger Sicht mit einem facettentheoretischen Design umgehen. Die Facettentheorie wurde von Louis Guttman in den 1950er Jahren als

[67] Pflegerisch relevante Sachverhalte können so gut wie nie intervallskaliert erfasst werden, darüber hinaus sind personenspezifische Merkmale naturgemäß nicht konstant (was aber nicht als Messfehler interpretiert werden kann) sowie meist nicht normalverteilt. (Bensch, 2011; Reuschenbach, 2011, 59; Franken, 2010)

Methodologie/Metatheorie für komplexe verhaltens- und sozialwissenschaftliche Fragestellungen erstmals beschrieben (Guttman 1954; Borg 1996, S. 237). Sie zielt darauf ab, Zusammenhänge zwischen statistischen Unterschieden in den Daten und der Definition des theoretischen Konstrukts zu finden und grafisch abzubilden. Die Facettentheorie lässt sich in drei mit einander in Verbindung stehende Komponenten gliedern (siehe Abb. 6.3).

Das Facetten-Design umfasst

1. die Theoriedefinition und -konstruktion mit,
   - (Struktur)Hypothesen/regionale Hypothesen zur Theorie
   - Operationalisierung der Theorie in ein Erhebungsinstrument innerhalb des Abbildungssatzes
   - Korrespondenzhypothesen, die klären, ob vermeintlich theoretisch relevante Facetten die Struktur der Beobachtungen (empirische Daten) erklären.
2. die Datenanalyse, für die aufgrund der jeweiligen Korrespondenzhypothesen neben der meistverwendeten Multidimensionalen Skalierung (MDS) vielfältige weitere statistische Verfahren eingesetzt werden können.

Die Facettentheorie dient der Differenzierung inhaltlicher Fragestellungen mit dem Ziel der Theoriebildung. Vorteil dieses problemunabhängigen, formalen, integrativen und damit systematischen Forschungsansatzes ist es, dass alle erforderlichen Arbeitsschritte der Theoriekonstruktion und ihrer Validierung aufeinander abgestimmt sind.

Facetten dienen der kategorialen Differenzierung eines empirischen Sachverhalts (hier Pflegebedürftigkeit) aus einer bestimmten Perspektive heraus. In diesem Kontext wird die konkrete Fragestellung verfolgt, welche Aspekte (Facetten) den Umfang von Pflegebedürftigkeit unterscheiden und damit als inhaltlich relevant bei der Theoriekonstruktion zu berücksichtigen sind. Zentrale Hypothese ist es, dass die zu entwickelnden Kategorien (Facetten) Pflegebedürftigkeit in ihrem Ausmaß determinieren.

Pflegebedürftigkeit wird als latentes, mehrdimensionales Konstrukt verstanden, das mutmaßlich von zahlreichen Facetten bestimmt wird. Die spezifischen Zusammenhänge der an Pflegebedürftigkeit beteiligten Faktoren (Facetten, Elemente und Dimensionen) sind bislang diffus und spekulativ. Belegt werden konnte, dass es

sich nicht um ein eindimensionales, additives[68] Modell handeln kann (Franken 2010; Schröder 2010; Bensch 2012 in diesem Band, S. 138).

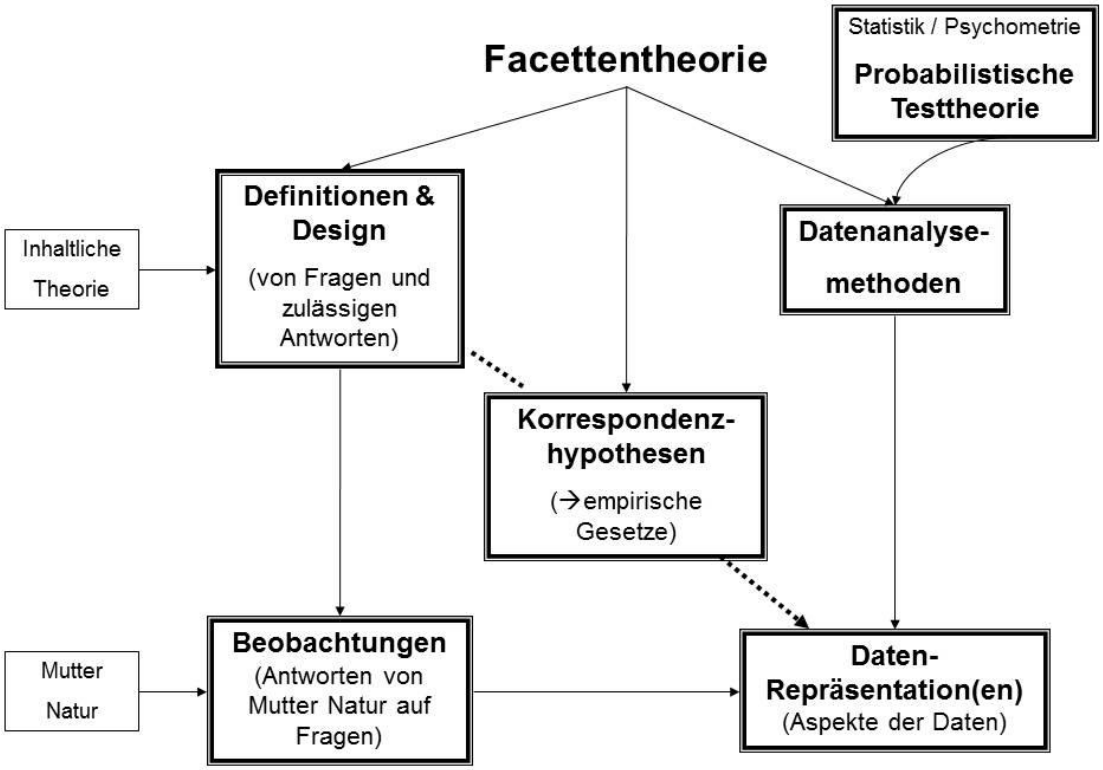


Abb. 6.3 Modifizierte schematische Übersicht der Elemente der Facettentheorie und ihrer Beziehung untereinander, bzw. zu anderen Mitspielern empirischer Forschung. (Quelle: Borg 1992, 13)

Die Chancen für eine valide Messung des Ausmaßes von Pflegebedürftigkeit sind abhängig von Kenntnissen über die Strukturen der beteiligten Faktoren, die das Konstrukt konstituieren. Die detaillierte und spezifische Überprüfung der hypothetischen Strukturen des Konstrukts der Pflegebedürftigkeit ist das erste Ziel, das zur Konkretisierung der Theorie verfolgt wird. Die vier Verhaltensdimensionen[69] der Theorie der umwelt- und familienbezogenen Pflege (Friedemann 2003, S. 31ff) werden als pflegetheoretische Ausgangsstruktur eines mehrdimensionalen Pflegebedürftigkeits-Modells als geeignet erachtet, weil sich alle Lebensvollzüge, deren Einschränkungen Pflegebedürftigkeit mutmaßlich begründen, mit den vier Verhaltensdimensionen erfassen lassen. Sie werden mit einzelnen Facetten der

[68] Bei einem additiven Modell wird angenommen, dass z.B. Einschränkungen der Mobilität und Kognition eine höhere Pflegebedürftigkeit mit sich bringen als eine Einschränkung ausschließlich in einem der beiden Bereiche. Die Praxis zeigt aber, dass ein mobiler, kognitiv eingeschränkter Mensch einen höheren Pflegebedarf haben kann als ein Mensch mit Einschränkungen der Kognition und Mobilität.

[69] Systemerhalt, Systemänderung, Kohärenz und Spiritualität

Pflegebedürftigkeits-Definition[70] von Wingenfeld et al kombiniert (Wingenfeld et al 2008, S. 28). Der Vorteil Friedemanns Theorie liegt in ihrem lebensweltorientierten Bezug, der Pflegebedürftigkeit im Kontext sozialer Systeme versteht.

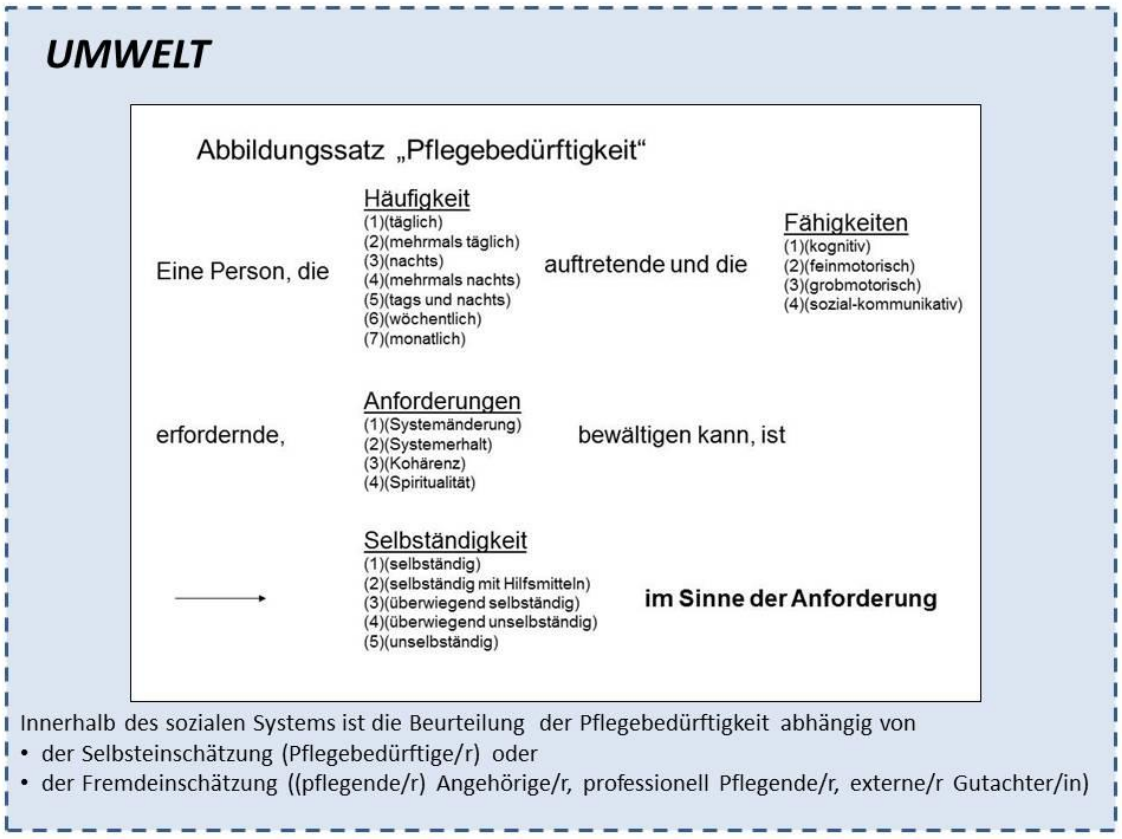


Abb. 6.4 Abbildungssatz „Pflegebedürftigkeit“ zur Operationalisierung des Konstrukts im Rahmen der Facettentheorie

Obwohl wir aktuell mit bedürfnisorientierten Pflegetheorien arbeiten, die auch wir momentan zur Erklärung von Pflegeaufwand einsetzen, wird dieser Rahmen unseres Erachtens nicht Bestand haben können. Erklärungsansätze für die Komplexität der Pflege chronisch Pflegebedürftiger bieten bedürfnisorientierte Pflegetheorien aber nicht, weil sie in tendenziell monokausalen Zusammenhängen kompensatorische Pflegeinterventionen auf einer eher körperbezogenen Ebene (Ziel ist Gesundheit im medizinischen Verständnis) fördern. Sie sind auch deshalb so beliebt, weil ihre Kompatibilität mit dem derzeit gültigen Pflegebedürftigkeitsbegriffs des SGB XI ein schlüssiges Verständnis von Pflege innerhalb des sozialrechtlichen Rahmens erzeugt. Um ein umfassenderes, lebensweltorientiertes Verständnis von Pflege, das auch die

[70] „Danach ist eine Person „als pflegebedürftig zu bezeichnen, wenn sie infolge fehlender personaler Ressourcen, mit denen körperliche oder psychische Schädigungen, die Beeinträchtigung körperlicher oder kognitiver/psychischer Funktionen, gesundheitlich bedingte Belastungen oder Anforderungen kompensiert oder bewältigt werden könnten, dauerhaft oder vorübergehend zu selbständigen Aktivitäten im Lebensalltag, selbständiger Krankheitsbewältigung oder selbständiger Gestaltung von Lebensbereichen und sozialer Teilhabe nicht in der Lage und daher auf personelle Hilfe angewiesen ist“ (Wingenfeld et al. 2007, 43).“

Umwelt (wie es auch der neue Pflegebedürftigkeitsbegriff vorsieht) mit zu berücksichtigen wird unseres Erachtens ein anders strukturierter theoretischer Ansatz erforderlich sein.

Die Facettentheorie „zwingt“ den Forscher, implizite Annahmen über theoretische Strukturen aufzuschlüsseln, um sie einer empirischen Prüfung unterziehen zu können. D. h. detaillierte Verästelungen komplexer Theorien lassen sich in Abbildungssätze überführen. Im Rahmen eines ersten struktionierten Abbildungssatzes (siehe Abb. 6.4) werden die Facetten mit ihren entsprechenden Kategorien in einen Bedeutungs- und Beziehungszusammenhang gebracht.

Die zu konstruierenden Items bilden die Kombinationen der Facetten-Elemente ab. Damit dient der Abbildungssatz der systematisierten Itementwicklung und verhindert, dass wichtige inhaltliche Bereiche mit dem Itempool nicht erfasst werden. Mögliche relevante personenbezogene Facetten wie z.B. das pflegerische Versorgungssetting (Häuslichkeit, mit/ohne ambulanten Pflegedienst, [teil]stationäre Versorgungsform), Alter, medizinische Diagnosen und die Rolle des Auskunftgebenden (Pflegebedürftiger, Angehöriger, Professioneller) werden als definitionsunabhängige Items erfasst und sind somit nicht Bestandteil des Abbildungssatzes. Sie dienen zur Überprüfung der Hypothesen der Zusammenhänge zwischen Ausprägung der Pflegebedürftigkeit und sozialen Aspekten und Kontexten. Die definitionsgemäße Grundgesamtheit aller möglichen Items des Erhebungsinstruments wird durch das kartesische Produkt der Elemente aller Facetten gebildet. Alle Kombinationsmöglichkeiten können als Struktupel [71] dargestellt werden und dienen der Operationalisierung des Abbildungssatzes in einzelne Items. Struktupel „1321“ beschreibt eine (1) tägliche, (3) grobmotorische Fähigkeiten erfordernde, (2) systemerhaltende Anforderung, die (1) selbständig zu erfüllen ist. Ein entsprechendes Item könnte z.B. danach fragen, ob ein Pflegebedürftiger in der Lage ist, eine Jacke anzuziehen. Genau so könnten Items für ein neues Messinstrument entwickelt werden.

Ob die theoretischen Annahmen mit Hilfe des entwickelten Instruments beobachtbar sind und sich in der Realität/Praxis wiederfinden lassen, wird durch die konfirmatorische Analyse der empirischen Daten überprüft. Entspricht die Partitionierung der empirischen Daten im mehrdimensionalen Raum den Strukturhypothesen des Facettendesigns, bestätigt sich die Hypothese über die Bedeutung einer Facette und ihrer Elemente für die Theorie. Für eine explorative Fragestellung ist es von Interesse, welche Facetten in welcher Verknüpfung

[71] Zusammengesetzt aus Strukt und n-tupel: jeweils eine Merkmalskombination aus dem kartesischen Produkts der Elemente der Facetten des Abbildungssatzes: die Struktur (meist eine Zahlen oder Buchstaben-Zahlen-Kombination) der hintereinander angegebenen Merkmale (Elemente) der Facetten, dem ein Item eines Instruments zugeordnet werden kann.

miteinander das Konstrukt strukturieren. Die Möglichkeit anhand der Verteilungsmuster der Daten im mehrdimensionalen Raum modulierende Facetten identifizieren zu können, bietet u.a. die Chance, z. B. das Verhältnis von kognitiven und motorischen Fähigkeiten innerhalb des Konstrukts Pflegebedürftigkeit konzeptionell genauer beschreiben zu können.

Facetten können sowohl qualitativ (wie z.B. die Facette „Fähigkeit“ in Abb. 6.4) im Sinne unterschiedlicher Kategorien als auch quantitativ (wie z. B. eine Facettierung „Temperatur“ in °C) differenziert werden. Qualitative und quantitative Strukturierung der Facetten lassen unterschiedliche Abbildungen der empirischen Daten im Raum erwarten. Darüber hinaus lässt sich anhand der grafischen Abbildung der empirischen Daten feststellen, ob und in welcher Weise sich die Facetten zueinander im Raum ordnen lassen, wodurch Rückschlüsse auf deren konzeptionelle Relationen möglich werden.

Das ausschließlich facettentheoretische und damit eher Theorie strukturierende, wissenschaftsorientierte Vorgehen ist nicht darauf ausgerichtet, einen Index für Pflegebedürftigkeit aufgrund der praktischen Anforderungen nach Messbarkeit von Pflegebedürftigkeit zu entwickeln (Borg 1992, S. 138). Allerdings ist es für Pflegebedürftigkeit von weitreichendem sozialpolitischem und rechtlichem Interesse, bestimmen zu können, welchen Grad der Pflegebedürftigkeit ein Mensch hat. Umgekehrt zeigt die Erfahrung, dass es kaum möglich ist, einen validen und praktikabel zu ermittelnden Index (am einfachsten als Summenscore relevanter Items) für Pflegebedürftigkeit statistisch zu entwickeln, wenn über die Struktur des latenten Konstrukts nur Vermutungen bestehen, über die sich mit inferenzstatistischen Methoden kein weiterer Erkenntnisgewinn generieren lässt.

Probabilistische Verfahren sind aufgrund ihrer weniger strengen Anforderungen an das Datenniveau durchaus geeignet typische Daten pflegerischer Sachverhalte zu verarbeiten (vgl. Brühl 2012, in diesem Band, S. 44). So sind die Ergebnisse einer Rasch-Analyse vor dem Hintergrund einer validen Indexbildung von großer Praxisrelevanz. Ohne detaillierte Kenntnisse über die Struktur zu haben, die das Konstrukt Pflegebedürftigkeit charakterisiert, wird es allerdings kaum möglich sein, ein solch komplexes Konstrukt wie Pflegebedürftigkeit in ein valides Rasch-Modell überführen zu können.

Eine Itemreduktion ausschließlich aufgrund statistischer Parameter führt zu inhaltlich unsinnigen oder trivialen Ergebnissen und ist damit nicht geeignet, den Grad der Pflegebedürftigkeit detailliert und valide zu differenzieren. Eine exakt facettierte Theorie eröffnet die Möglichkeit Korrespondenzhypothesen zu testen und damit Rückschlüsse auf die Strukturierung der Theorie zu gewinnen.

Darüber hinaus lassen sich aus rein statistischen Ergebnissen ohne Theorie kaum fachlich nützliche Erkenntnisse für die Messung von Pflegequalität und die Berechnung von Personalbedarfen, sowie die Prävention von Pflegebedürftigkeit ziehen. Da sich die Bewertung der Modellgeltung und Modellgüte probabilistischer Verfahren wiederum nicht alleine aus den empirischen Daten heraus sinnvoll begründen lässt, sondern parallel eine inhaltliche Interpretation erforderlich ist, könnte die facettentheoretische Datenanalyse wichtige Hinweise zur Konstruktion des Messinstruments und damit wichtige Hinweise für die Bildung eines validen Index bieten. Im Rahmen eines Forschungsprojekts ist zu prüfen, ob sich die Annahme einer synergetischen und damit gewinnbringenden Kombination von Facettentheorie und probabilistischer Testtheorie für die Entwicklung von Pflegetheorien und pflegerischen Test-/Messverfahren bestätigt (Balia et al. 1985).

In Tabelle 6.1 wird ein Überblick über den letzten sechs Jahren im Zusammenhang der Instrumentenentwicklung bereits getesteten Verfahren und den noch zu testenden Verfahren gegeben. Den bereits getesteten Verfahren haben wir unsere Erfahrungen beigefügt:

Programm Methodenerweiterung im Messen: 2006 - 2012				
Schritt	Frage	Theoretisches Problem	Lösungsmethoden	Ergebnis
1.	Wie werden Antwortskalenrohwerte im gewichtet?	Ordinale Daten zu Intervalldaten	Unstandardisierten Gewichte einer linearen, multivariaten Regression von Einzelfragen auf die Gesamtrohwertsumme	Stichprobenabhängig; Gewichte-Schätzer mit großen Standardfehlern
2.	Kann man mit ordinalen Daten Strukturen entdecken?	Ordinale auch ordinal behandeln	Lineare Strukturgleichungs-modelle , Konfirmatorische Faktorenanalysen mit polychorischen Korrelationen	Polychorische Korrelationen führen zu Schätzproblemen, da die Standardschätzverfahren nicht funktionieren
3.	Komme ich von dichotomen zu intervallskalierten Daten? Kann ich aus dichotomen Daten aussagekräftige Merkmalsmuster bilden?	Aus qualitativen Daten quantitative machen oder qualitative Standardisierungsmodelle entwickeln	Rasch-Modelle Latente Klassen	Rasch-Modelle nicht nachträglich anpassbar und latente Klassen können nur mit wenigen Variablen arbeiten
4.	Wie erfülle ich die Aufgabe der Aufwandsklassifikation?	Nichtlineare Beziehungen und Interaktionen zwischen Variablen entdecken	Rekursives Partitionieren Regression Trees, Regression Forrests, Bagging, Boosting, Multivariate Adaptive Regression Splines	Mit MARS ist ein einrichtungsübergreifend es Modell möglich, das über 70% der Varianz erklärt (Pflegestufen 30%)
5.	Wie operationalisiert man theoretische Konstrukte?	Überführung einer Theorie in ein Erhebungsinstrument	Facettentheorie	Erste Facettierung des Konstrukts Mobilität
6.	Wie modelliere ich Veränderungmessungen z.B. zur Bewohner-Struktur?	Modellierung von Veränderungsmessung in Panel-Daten	Mehr-Ebenen-Modelle	Noch keine
7.	Wie finde ich Dimensionen in diochotomen Ausgangsdaten?	Ich will die Struktur meines Instrumentes kennen	Multidimensionale Skalierung MDS mit unfolding-Modellen	Noch keine

Tab. 6.1 Methodenüberblick zur Instrumentenentwicklung

ENTWICKLUNGSBEDARF

Die Entwicklung eines validen Instruments von dieser enormen gesellschafts- und sozialpolitischen Bedeutung erfordert intensivere und differenziertere Forschung als die bislang durchgeführte.

Erforderlich ist die Weiterentwicklung des Inhaltsmodells zum Phänomen der Pflegebedürftigkeit, das daraufhin mit einem validen Strukturmodell unter zeitgleicher Berücksichtigung der testtheoretischen Anforderungen des ausgewählten Messmodells gemessen werden kann.

Durch die Prüfung des Strukturmodells mittels statistischer Verfahren (Testtheorie), die bei der Operationalisierung theoretischer Inhalte (Pflegetheorie) in das Instrument in Form der Items „mitzudenken“ sind, wird es möglich sein, Hinweise auf die Struktur des Phänomens zu bekommen, die wiederum der Theorieentwicklung dienlich sein werden. Nur auf diese komplexe, zirkuläre Weise wird es möglich sein, eine stimmige Theorie zum Phänomen der Pflegebedürftigkeit und ein entsprechend valides Instrument zur Messung von Pflegebedürftigkeit entwickeln zu können.

Grundsätzlich muss die Quantifizierbarkeit von komplexen Konstrukten immer geprüft werden bevor Methoden eingesetzt werden, die quantitative Daten voraussetzen. Es kann nicht einfach von linearen Beziehungen ausgegangen werden, ohne dies zu prüfen. Kurz gesagt müssen in der Regel Standardverfahren der Statistik durch nicht parametrische, nicht lineare, wahrscheinlichkeitsbasierte Verfahren ersetzt werden.

Aufgrund der engen inhaltlichen Zusammenhänge werden Ergebnisse aus Forschungsprojekten zur Pflegepersonalbedarfsbemessung, zur Messung von Pflegebedürftigkeit sowie Messungen von Pflegequalität einen Erkenntniszuwachs für die jeweils anderen Themen ermöglichen, der zu nutzen ist. Im Kern geht es darum, Faktoren als valide verallgemeinerbar zu identifizieren, mit denen sich Pflegebedürftigkeit und damit Pflegepersonalaufwand *erklären* und nicht nur abbilden lässt. Denn erst valide, weil auch überschaubare und damit anwendungsfreundliche Instrumente können den drängenden Praxisfragen gerecht werden.

LITERATUR

Balia, John R.; McDonald, Roderick P. (1985): Latent Trait Item Analysis and Facet Theory – A Useful Combination. Applied Psychological Measurement 9 (2) 191-198

Bensch, Sandra (2011): Konstruktvalidität der Module „Mobilität“ und „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessments zur Feststellung von Pflegebedürftigkeit. Inaugural-Dissertation an der Pflegewissenschaftlichen Fakultät. Philosophisch-Theologische Hochschule Vallendar

Borg, Ingwer (1992): Grundlagen und Ergebnisse der Facettentheorie. Methoden der Psychologie Bd. 13. Bern: Huber

Borg, Ingwer; Mohler, Peter Ph. (1993): Zur Indexbildung in der Facettentheorie. ZUMA Nachrichten 17; 33; 10-24

Borg, Ingwer; Groenen, Patrick (2010): Multidimensionale Skalierung. München: Hampp

Breiman, Leo (1996): Bagging Predictors. IN Machine Learning, 24: 123–140

Breiman, Leo; Friedman, Jerome H.; Olshen, Richard A.; Stone, Charles J. (1998): Classification and Regression Trees. Reprint. New York: Chapman & Hall

Breiman, Leo (2001): Random Forests. Machine Learning, 45: 5–32

Bühlmann, Peter; Yu, Bin (2003): Boosting with the L2- Loss – regression and Classification. Journal of the American Statistical Association, 98:324–339

Bühner, Markus (2006): Einführung in die Test- und Fragebogenkonstruktion. 2. aktualisierte Auflage. München: Pearson

Carstensen, Claus H. (2000): Mehrdimensionale Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik. Kiel: IPN

Franken, Georg (2010): Konstruktvalidität der Subskala „Kognitive und kommunikative Fähigkeiten“ des Neuen Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit (NBA). Masterarbeit. Vallendar: Philosophisch Theologische Hochschule. http://opus.bsz-bw.de/kidoks/volltexte/2012/66/ zuletzt geprüft am 20.08.2012

Friedman, Jerome H. (1991) - Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1): 1-141

Friedemann, Marie-Luise; Köhlen, Christina (2003): Familien- und umweltbezogene Pflege. Bern: Huber

Guttman, Louis (1954): An Outline of some new Methodology for Social research. Public Opinion Quarterly 18; 395-404

Hasseler, Martina; Wolf-Ostermann, Karin (2010): Wissenschaftliche Evaluation zur Beurteilung der Pflege-Transparenzvereinbarungen für den ambulanten (PTVA) und stationären (PTVS) Bereich. Inklusive Empfehlungen des Beirates zur Evaluation der Pflege-Transparenzvereinbarungen. Hamburg/Berlin

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009): The Elements of Statistical Learning - Data Mining, Inference, and Prediction. 2nd Edition. New York: Springer

Hothorn, Thorsten; Hornik, Kurt; Zeileis, Achim (2006): Unbiased Recursive Partitioning: A Conditional Inference Framework. IN Journal of Computational and Graphical Statistics 15(3): 651-674

Hüngsberg, Melanie (2011): Validität von Rohsummenwerten mittels polychorischer Korrelationsberechnungen anhand der Mobilitätsskala des NBA. Masterarbeit an der Theologisch-Philosophischen Hochschule Vallendar: Pflegewissenschaftliche Fakultät

Krohwinkel, Monika (2007): Rehabilitierende Prozesspflege am Beispiel von Apoplexiekranken. Fördernde Prozesspflege als System. 2. überarbeitete und erweiterte Auflage. Bern: Huber

Mills, Ronald; Fetter, Robert B.; Riedel, Donald C.; Averill, Richard (1976): AUTOGRP: an interactive computer system for the analysis of health care data. Medical Care, 14(7): 603-615

Reuschenbach, Bernd; Mahler, Cornelia (Hg.) (2011): Pflegebezogene Assessmentinstrumente. Internationales Handbuch für Pflegeforschung und –praxis. Bern: Huber

Roper, Nancy; Logan, Winifred W.; Tierney, Alison J. (2002): Das Roper-Logan-Tierney-Modell. Basierend auf Lebensaktivitäten. Bern: Huber

Rost, Jürgen (2004): Lehrbuch Testtheorie – Testkonstruktion. Zweite, vollständig überarbeitete und erweiterte Auflage. Bern: Huber

Smith, David W.; Hogan, Andrew J.; Rohrer, James E. (1987): Activities of daily living as quantitative indicators of nursing effort. Medical Care 25(2): 120-130

Sonquist, John A.; Morgan, James N. (1964): The detection of interaction effects - a report on a computer program for the selection of optimal combinations of explanatory variables. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan

Schröder, Martina (2010): Konstruktvalidität der Subskala Mobilität des Neuen Begutachtungsassessments für Pflegebedürftigkeit (NBA). Masterarbeit. Betreut von Prof. Dr. Albert Brühl. Vallendar. Philosophisch-Theologische Hochschule Vallendar. Online verfügbar unter http://www.dip.de/datenbank-wise/detail/?no_cache=1&tx_dipwise_pi2[uid]=499, zuletzt geprüft am 09.08.2010

Strobl, Carolin; Malley, James; Tutz, Gerhard (2009): An Introduction to Recursive Partitioning. Technical Report Number 55, 2009. Department of Statistics, University of Munich. Online: http://epub.ub.uni-muenchen.de/10589/1/partitioning.pdf, 15.02.2012

Vogt, Joseph (2000): Bagging, Boosting und verwandte Methoden. Diplomarbeit. Zürich: Eidgenössische Technische Hochschule.

von Auer, Ludwig (2003): Ökonometrie – Eine Einführung. 2. Auflage. Berlin: Springer

Wingenfeld, Klaus; Büscher, Andreas; Schaeffer, Doris (2007): Recherche und Analyse von Pflegebedürftigkeitsbegriffen und Einschätzungsinstrumenten. Überarbeitete Fassung vom 23. März 2007. Universität Bielefeld: Institut für Pflegewissenschaft (IPW)

Wingenfeld, Klaus; Büscher, Andreas; Gansweid, Barbara (2008): Das neue Begutachtungsassessment zur Feststellung von Pflegebedürftigkeit. Projekt: Maßnahmen zur Schaffung eines neuen Pflegebedürftigkeitsbegriffs und eines neuen bundesweit einheitlichen und reliablen Begutachtungsinstruments zur Feststellung der Pflegebedürftigkeit nach dem SGB XI. Abschlussbericht zur Hauptphase 1: Entwicklung eines neuen Begutachtungsinstruments. Bielefeld: IPW