Aus dem
CharitéCentrum für Innere Medizin und Dermatologie
Medizinische Klinik mit Schwerpunkt Psychosomatik
Direktor: Prof. Dr. Matthias Rose

Habilitationsschrift

Vom Instrument zum Konstrukt – standardisierte Messung gesundheitsbezogener Lebensqualität

zur Erlangung der Lehrbefähigung
für das Fach *Experimentelle Psychosomatische Medizin*

vorgelegt dem Fakultätsrat der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von
Dr. rer. nat. Herbert Felix Fischer

Eingereicht: April 2019
Dekan: Prof. Dr. med. Axel R. Pries
1. Gutachter: Prof. Dr. Harald Gündel, Ulm
2. Gutachter: Prof. Dr. Claas Lahmann, Freiburg

Inhaltsverzeichnis

Abkürzungsverzeichnis

ACSA Anamnestic Comparative Self-Assessment
BDI Beck Depression Inventory
CAT Computer-adaptiver Test
CES-D Center of Epidemiological Studies Depression Scale
DIF Differential Item Functioning
EAP Expected a Posteriori
EORTC European Organization for Research and Treatment of Cancer
EQ-5D-5L EuroQol EQ-5D
HADS Hospital Anxiety and Depression Scale
HrQoL Health-related Quality of Life, gesundheitsbezogene Lebensqualität
HUI Health Utility Index
GPCM Generalized Partial Credit Model
GRM Graded Response Model
ICC Item Characteristic Curve
IRT Item-Response Theory
KTT Klassische Testtheorie
KOOS Knee Injury and Osteoarthritis Outcome Score
ML Maximum-Likelihood
NCDIF Non-compensatory Differential Item Functioning
PHQ Patient Health Questionnaire
PRO Patient-Reported Outcome
PROM Patient-Reported Outcome Measure
PROMIS Patient-Reported Outcome Measurement Information System
QoL Quality of Life, Lebensqualität
SF-36 Short-Form 36
WHO World Health Organization

1 Hintergrund

Ziel ärztlichen Handelns ist die Linderung von durch Krankheit verursachtes Leid. Um den Erfolg dieses Strebens im klinischen Alltag und in der wissenschaftlichen Arbeit prüfen zu können, müssen wir den Einfluss von Krankheit und Therapie auf den Menschen messen. Neben Mortalität und Morbidität rückten in den letzten Jahren Symptome von Erkrankungen und daraus folgende Funktionseinschränkungen in den Fokus, die unter dem Oberbegriff *gesundheitsbezogene Lebensqualität* von den Patienten selbst berichtet werden können.

In der Folge dieser Entwicklung wurden unzählige, erkrankungsspezifische und generische Fragebögen entwickelt, um verschiedene Aspekte der gesundheitsbezogenen Lebensqualität zu erfassen. Diese Fragebögen unterscheiden sich in der theoretischen Konzeptualisierung des Zielkonstrukts, der Zielpopulation, ihrer Popularität und nicht zuletzt ihrer Qualität. Daraus folgt nicht nur, dass die Auswahl eines geeigneten Fragebogens für den spezifischen Anwendungsfall schwierig ist, sondern auch dass wir die erhobenen Daten bisher nur schlecht über Studien, Länder oder Erkrankungen vergleichen können.

In dieser Habilitationsschrift werden Arbeiten dargestellt, denen das Ziel einer stärkeren Standardisierung der Messung von gesundheitsbezogener Lebensqualität aus der Patientenperspektive gemeinsam ist. Zum einen wird die Entwicklung und Validierung einer instrumentenunabhängigen Skala dargestellt, auf der Depressivität als Kernaspekt psychischer Gesundheit mittels verschiedener Fragebögen gemessen werden kann. Zum anderen werden Arbeiten vorgestellt, in denen die Validität von Meßinstrumenten zur Erhebung von Aspekten der gesundheitsbezogenen Lebensqualität über verschiedene Sprachen hinweg untersucht wurde.

Im folgenden soll nun zunächst der Begriff der gesundheitsbezogenen Lebensqualität als Zielparameter medizinischer Forschung beschrieben werden. Danach folgt eine kurze Darstellung der Item-Response Theory (IRT), die das statistische Fundament der danach folgenden Originalarbeiten darstellt. Am Ende werden die Forschungsergebnisse eingeordnet und die Perspektiven für zukünftige Arbeiten diskutiert.

1.1 Lebensqualität als Zielparameter in der Medizin

Definition

Die Weltgesundheitsorganisation definiert *Lebensqualität*:

> …as individuals perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns.
>
> It is a broad ranging concept affected in a complex way by the person's physical health, psychological state, level of independence, social relationships, personal beliefs and their relationship to salient features of their environment.
>
> [1]

Gesundheitsbezogene Lebensqualität bezieht sich dann auf diejenigen Aspekte der Lebensqualität, die mit der Gesundheit zu tun haben – das ist im Allgemeinen der Einfluss von Krankheit und Therapie auf die individuelle Funktionsfähigkeit und die Fähigkeit, ein erfüllendes Leben zu leben [2]. Ein umfassendes Modell der gesundheitsbezogenen Lebensqualität berücksichtig nicht nur die mit einer Erkrankung einhergehenden Beschwerden und Funktionseinschränkungen, sondern auch Personen- und Umgebungsvariablen [3]. Im Gegensatz zu Mortalität und Morbidität ist damit die gesundheitsbezogene Lebensqualität ein subjektives Maß für die Belastung eines Menschen durch Krankheit. Einzelne Teilkonstrukte, die in ihrer Gesamtheit die gesundheitsbezogene Lebensqualität konstituieren, sind zum Beispiel die körperliche Funktionsfähigkeit, Symptome psychischer Belastung oder Einschränkungen in der sozialen Teilhabe.

Wichtig ist es, zwischen dem Konstrukt an sich – dem *Patient-Reported Outcome (PRO)* – und dem Meßinstrument – *Patient-Reported Outcome Measure (PROM)* – zu unterscheiden. In vielen Fällen wurden für ein spezifisches PRO mehrere PROMs entwickelt, und auch ein einzelnes PROM kann mehrere PROs abdecken. Ein Beispiel ist der Knee Injury and Osteoarthritis Outcome Score (KOOS) [4] als PROM, mit dem verschiedene Aspekte der gesundheitsbezogenen Lebensqualität als PROs (unter anderem Schmerzen, Symptome, Funktionsfähigkeit im Alltag) erhoben werden. PROMs können dabei krankheitsspezifisch oder generisch angelegt sein. Weitere Beispiele für PROs und PROMs finden sich in Tabelle 1.1.

Zielkonstrukt (PRO)	Meßinstrumente (PROM)
allgemeine Lebensqualität	Anamnestic Comparative Self-Assessment (ACSA)
gesundheitsbezogene Lebensqualität	
global	EQ-5D, Health Utility Index (HUI)
körperliche Gesundheit	SF-36 Physical Component Score
mentale Gesundheit	SF-36 Mental Component Score
Symptome & Funktionseinschränkungen	
Depressivität	PHQ-9, BDI-I/II, CES-D
Fatigue	Brief Fatigue Inventory, Fatigue Scale, Fatigue Severity Scale, EORTC QLQ-FA12
körperliche Funktionsfähigkeit	SF-36 PF Skala, Health Assessment Questionnaire, Barthel Index

Tabelle 1.1: Einige Beispiele für Patient-reported outcome measures (PROMs) für ausgewählte Patient-reported outcomes (PROs) auf verschiedenen Ebenen

Im weiteren wird also der Begriff PRO für die verschiedenen inhaltlichen und definitorisch abgrenzbaren Aspekte der gesundheitsbezogene Lebensqualität verwendet. PROM steht im Gegensatz dazu für den jeweiligen Fragebögen beziehungsweise das Meßinstrument.

Anwendung und Relevanz

PROs werden also mit PROMs gemessen. Gemeinhin werden dazu standardisierte Fragebögen verwendet. Bei der Erhebung von PROs in der Medizin lassen sich drei wesentliche Anwendungsgebiete unterscheiden:

Screening PROMs können eingesetzt werden, um Personen mit hoher Merkmalsausprägung zu identifizieren. Als Beispiel kann hier der Gesundheitsfragebogen für Patienten (Patient Health Questionnaire (PHQ)) genannt werden, der zur Detektion psychischer Symptome bei Patienten der Allgemeinmedizin entwickelt wurde [5, 6]. Der Nutzen einer breiten Implementierung von Screeningverfahren in anderen Bereichen wird dagegen kontrovers diskutiert, zum Beispiel hinsichtlich der Identifikation von Depression Frauen postpartum [7] oder bei Patienten mit Herzinsuffizienz [8, 9].

Prädiktoren PROs können als Prädiktoren anderer Zielparameter dienen und somit Zusammenhänge zwischen verschiedenen Erkrankungsaspekten aufdecken. So wurde gezeigt, das Erhebungen von PROs ungünstige Therapieverläufe bei Patienten mit chro-

nisch-obstruktiver Lungenkrankeit [10] und 1-Jahres-Mortalität bei Patienten mit Herzinsuffizient [11] vorhersagen können.

Outcomemessung Die Messung des patientenberichteten Gesundheitsstatus zur Beurteilung des Therapieerfolgs spielt eine herausgehobene Rolle. Die relevanten PROs sind hierbei Aspekte der körperlichen (zum Beispiel Schmerz, körperliche Funktionsfähigkeit) und mentalen Gesundheit (zum Beispiel Depressivität, Angst). Neben der Erfassung des Erkrankungsverlaufes konnte gezeigt werden, dass ein kontinuierliches Monitoring der Symptomatik und automatische Email-Benachrichtigung bei Verschlechterung sowohl Rettungsstellenbesuche als auch die Mortalität reduzieren kann [12]. Die Nutzung von PROs zur Messung von Outcomes hat aber im Zuge einer Hinwendung zur *value-based healthcare* weitere Dimensionen erhalten. So sind Daten aus PROs nicht nur zur Beurteilung des Therapieerfolgs eines einzelnen Patienten wertvoll, sondern können auch zur Evaluation neuer Behandlungsverfahren, zur partizipativen Entscheidungsfindung, zur Evaluation von Anbietern von Gesundheitsleistungen sowie zur Abbildung systemweiter Entwicklungen dienen [13]. Daten zur gesundheitsbezogenen Lebensqualität werden auch in Deutschland bei der Nutzenbewertung neuer Medikamente und Medizinprodukte herangezogen [14, 15].

Die Berücksichtigung der subjektiven Patientenperspektive auf die Erkrankung durch die Erhebung von PROs ist damit in Klinik, Forschung und Gesundheitspolitik gleichermaßen relevant. Dies gilt umsomehr, da in den letzten Dekaden der Anteil chronischer Erkrankungen gegenüber Infektionskrankheiten zugenommen hat [16–18]. Dadurch gewinnen Aspekte der gesundheitsbezogenen Lebensqualität an Gewicht, da neben der Behandlung der Ursache der Erkrankung das Aufrechterhalten des Funktionsniveaus als Behandlungsziel in den Vordergrund rückt.

Probleme der gegenwärtigen Praxis

Aufgrund der breiten Relevanz in Forschung und Klinik, den verschiedenen Einsatzzwecken sowie theoretischen Hintergründen wurden eine Vielzahl verschiedener Instrumente entwickelt, die oft die gleichen oder eng verwandte Konstrukte erfassen. So gibt es mehr als 100 verschiedene Skalen zur Erfassung von Depressivität [19] und in einem Review von 42 Studien der Psychotherapieforschung wurden mehr als 100 Fragebögen zur Erfassung der relevanten Zielparameter identifiziert [20].

Die Entwicklung von Fragebögen zur Erfassung der gesundheitsbezogenen Lebensqualität folgte dabei in der Vergangenheit meist den Prinzipien der Klassische Testtheorie (KTT) [21]. Die KTT operationalisiert den beobachteten Testwert als eine fehlerbehaftete Messung eines angenommenen, aber letztlich unbeobachtbaren „wahren“ Testwertes. Testwerte sind damit aber an das jeweilige Instrument gebunden, da es keine theoretische Konzeptualisierung eines zugrundeliegenden Konstruktes gibt, und können nicht direkt über Instrumente verglichen werden. Dies erschwert die Interpretation von Studienergebnissen und die Definition von instrumentenübergreifenden Standards wie Schwellenwerte oder minimale klinischen Differenzen, da diese ebenfalls instrumentenspezifisch entwickelt werden müssen.

Da Fragebögen gemeinhin das zu messende Konstrukt präzise in seiner Gesamtheit abbilden sollen, enthalten nach der KTT entwickelte Tests viele Items, die das ganze Spektrum des Konstrukts abbilden. Um einen instrumentenspezifischen Testwert erheben zu können, müssen nun aber alle Items eines bestimmten Fragebogens erhoben werden. Daraus folgt, dass nach der KTT entwickelte Tests hinsichtlich Länge, Meßbereich und Meßpräzision Kompromisse eingehen müssen.

1.2 Item-Response Theory

Mit der Verwendung der IRT als probabilistischem Testmodell können die zuvor beschriebenen Probleme adressiert werden. IRT ist das theoretische Gerüst zur Auswertung von PROs, auf dem die Arbeiten dieser Habilitationsschrift beruhen. Daher soll zunächst der theoretische Hintergrund erläutert werden, um anschließend auf Vorteile und Anwendungen einzugehen. Zuletzt soll kurz das Patient-Reported Outcome Measurement Information System (PROMIS) als maßgebliches, auf IRT beruhende Konzept der Messung von PROs vorgestellt werden.

Theoretischer Hintergrund

Die IRT wurde in den 50er Jahren des 20. Jahrhunderts entwickelt. Das Antwortverhalten in einem Test wird hier als Ausdruck einer (oder mehreren) latenten, das heißt einer angenommenen, aber letztlich unbeobachtbaren Variable verstanden und mit Hilfe von mathematischen Gleichungen modelliert. Unter dem Begriff IRT werden eine Reihe verschiedener Modelle subsummiert, zum Beispiel das *Rasch Modell* für dichotome oder das *Graded Response*

Model (GRM) für geordnete, polytome Antwortformate [22]. Auch multidimensionale IRT-Modelle zur simultanen Erfassung mehrerer zugrundeliegender Dimensionen gibt es [23].

Gemeinsam ist all diesen Modellen, das abhängig von der Ausprägung der latenten Variable θ (theta) beim Probanden und spezifischen Itemeigenschaften (a = slope, d = thresholds) eine Wahrscheinlichkeit für die Wahl einer Antwortkategorie bei einem gegebenen Item angenommen wird.

Die Wahrscheinlichkeit, eine der möglichen Antwortkategorien zu wählen, ist im GRM:

$$P(X \geq 0|\theta, a, d) = 1$$

Für die verbleibenden Antwortoptionen gilt dann:

$$P(X \geq x|\theta, a, d) = \frac{1}{1 + e^{-a\theta + d_x}}$$

Daraus folgen die Wahrscheinlichkeiten für die Wahl einer spezifische Antwortoption:

$$P(X = x|\theta, a, d) = P(X \geq x|\theta, a, d) - P(X \geq x + 1|\theta, a, d)$$

Dieser Zusammenhang zwischen latenten Merkmal und Antwortverhalten wird als *Item Characteristic Curve (ICC)* dargestellt (siehe Abbildung 1.2).

Item- und Personenparamater der Modelle können simultan mittels Maximum-Likelihood (ML)-Methoden geschätzt werden, zum Beispiel mit dem Paket mirt [24] in R [25]. Voraussetzung für die korrekte Schätzung unidimensionaler IRT-Modelle ist die tatsächliche Unidimensionalität des Zielkonstrukts und die stochastische lokale Unabhängigkeit der Items - das heißt nach Berücksichtigung der latenten Variable sollen die Fehlerterme der einzelnen Items unkorreliert sein [22, 26]. Die Schätzung des Modells geschieht in der sogenannten Kalibrierungsstichprobe. Diese muss hinreichend groß sein, um das zu messende Konstrukt in seiner ganzen Breite beobachten zu können.

Wurden Itemparameter in einer Kalibrierungsstichprobe geschätzt, kann anschließend aus einer spezifischen Itemantwort einer Person auf die Ausprägung der latenten Variable bei dieser Person geschlossen werden. Abbildung 1.2 illustriert das Prinzip. Es ergibt sich aus

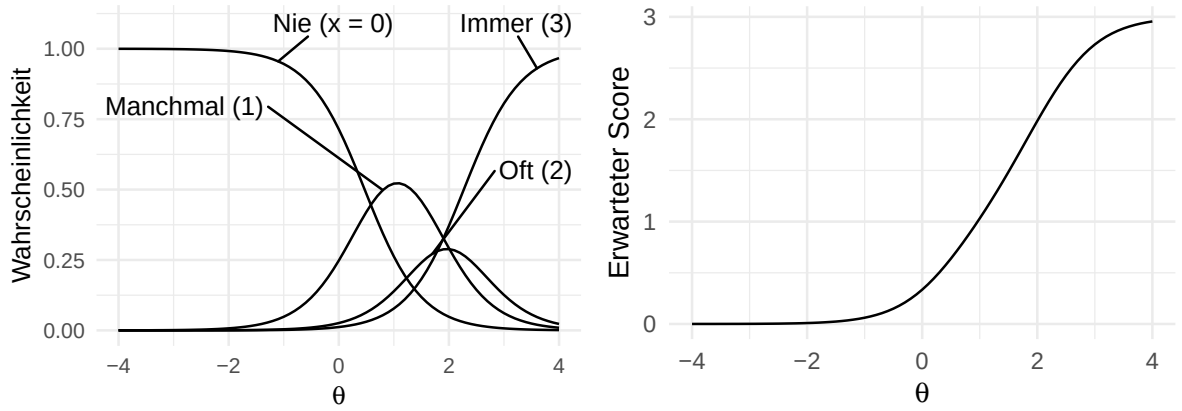

Abbildung 1.1: Item Characteristic Curve und erwarteter Summenscore eines Items in Abhängigkeit der latenten Variable θ. Im Bereich $\theta < -1$, ist die Wahrscheinlichkeit, dass ein Proband Antwortkategorie 0 ('Nie') wählt nahezu 100%. Bei höherer Merkmalsausprägung ist die Wahl einer anderen Antwortkategorie wahrscheinlicher, z.B. Antwortkategorie 1 ('Manchmal') im Bereich um $\theta = 1$. Aus diesen Wahrscheinlichkeiten lassen sich erwartete Item- und schlußendlich auch Testscores berechnen.

der Wahl der zweiten Antwortkategorie bei Item 1, dass θ zwischen -0.5 und 2.5 liegt, während die Wahl der Antwortoption 3 bei Item 2 auf ein θ zwischen 0.5 und 3.0 schließen lässt. Werden diese beiden Wahrscheinlichkeitsverteilungen multipliziert, ergibt sich ein Bereich zwischen 0.25 und 2.25 als wahrscheinliche Ausprägung für θ bei Vorliegen des beobachteten Antwortverhaltens. Mit jedem zusätzlich beantworteten Item wird der erwartete Wertebereich kleiner und damit die Schätzung der latenten Variable präziser [27].

Da im Falle extremer Antwortmuster (nur die kleinste/größte Antwortoption wird von einer Person gewählt) mit dieser ML-Schätzung kein finiter Schätzer errechnet werden kann, werden häufig Schätzer verwendet, bei denen die Wahrscheinlichkeitsverteilung zusätzlich mit einer Prior-Verteilung $N(0, 1)$ multipliziert wird. Ein Beispiel ist der Expected a Posteriori (EAP)-Schätzer, der ohne iterativen Prozess berechnet werden kann [27] und damit weniger rechenintensiv als alternative Schätzmethoden ist.

Vorteile und Anwendungen

Im Gegensatz zur KTT als reiner Messfehlertheorie [21] werden mit der IRT also die Testantworten in direktem Zusammenhang mit dem zu messenden Merkmal modelliert [28]. Eine

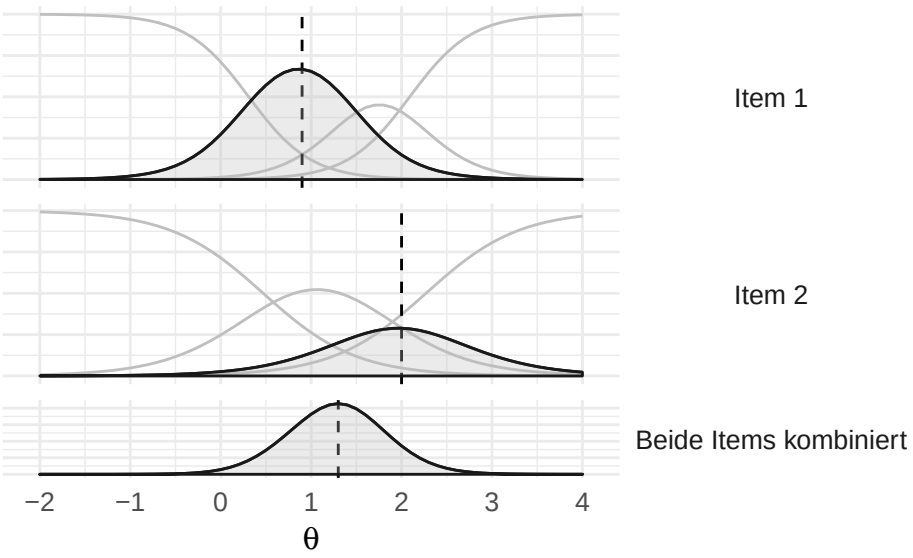

Abbildung 1.2: Anhand der gewählten Antwortoptionen lässt sich über die ICC auf die Ausprägung der latenten Variablen θ bei einer Person schließen

Sammlung von Items mit einem entsprechenden statistischen Modell zur Verknüpfung von Itemantwort und latenter Variable nennt man Itembank [22].

Ein Vorteil einer Itembank ist es, dass die latente Variable mit jedem beliebigen Subset von Items geschätzt werden kann. Neben der Möglichkeit, parallele Testformen zu erstellen [29] wird dies vor allem in der Entwicklung von Computer-adaptiver Tests (CATs) genutzt [30, 31]. In einem CAT werden jeweils nur die Items präsentiert, die für die angenommene Ausprägung der latenten Variable des Probanden möglichst informativ ist. Die Schätzung der latenten Variable wird dabei nach jeder neuen Antwort aktualisiert. Der Prozess ist exemplarisch in Abbildung 1.3 dargestellt.

Die Verwendung von CATs bietet gegenüber klassischen Fragebögen mehrere Vorteile, zum Beispiel werden unangemessene, weil zu leichte beziehungsweise schwere Items, nicht präsentiert. So kann die Gesamtzahl der Items und damit die Belastung des Probanden reduziert werden. Ein CAT kann mit 2-4 Items eine ähnliche diagnostische Treffsicherheit erreichen wie der PHQ-9 (9 Items) [32]. In den letzten Jahren wurde eine Vielzahl von CATs für PROs entwickelt, unter anderem zur Messung von Depressivität [33, 34], Angst [35, 36], Stress [37] oder Lebensqualität bei Kindern [38, 39].

CATs sind insbesondere nützlich, wenn das Zielkonstrukt eine große Bandbreite möglicher Ausprägungen hat (z.B. Intelligenz, körperliche Funktionsfähigkeit), wenn Tests besonders

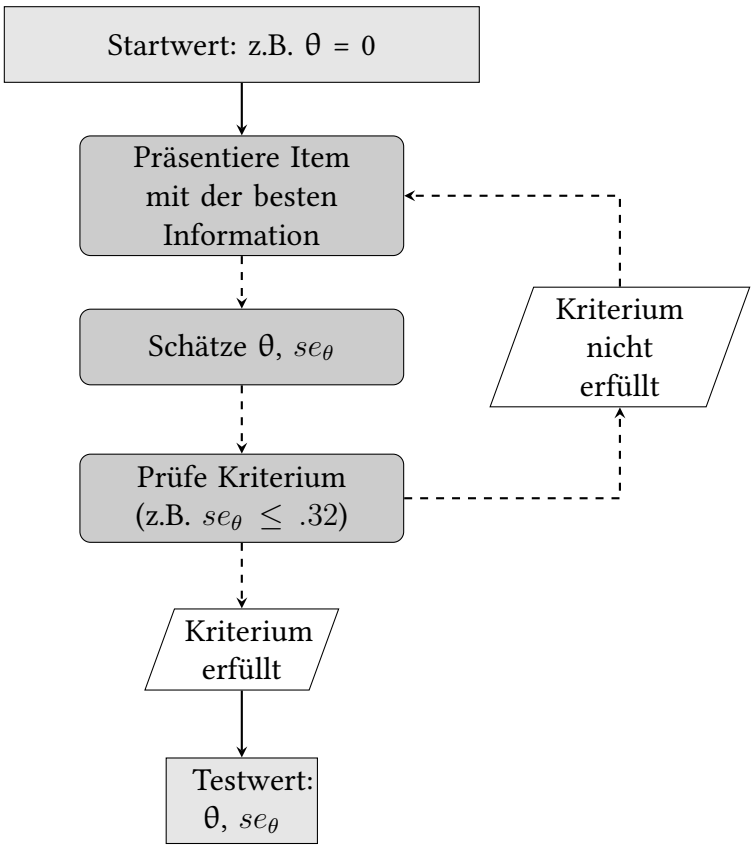

Abbildung 1.3: Algorithmus eines CAT. Es werden solange Items dargeboten, bis ein a priori spezifiziertes Kriterium erreicht wird, z.B. hinsichtlich der Genauigkeit des Testwertes (hier: $se_\theta \leq .32$, das entspricht einer Reliabilität von .9)

kurz sein müssen (z.B. weil viele Konstrukte in kurzer Zeit erhoben werden sollen) oder wenn die Meßgenauigkeit des Instruments über die Breite des Konstruktes variieren soll. Der größte Nachteil von CATs, der einer breiten Implementierung in der klinischen Praxis bisher entgegensteht, ist die Notwendigkeit einer flexiblen, computergestützten Datenerhebung.

Die zweite Anwendung, bei der man sich die Möglichkeit, die Ausprägung der latenten Variable mittels eines Subsets von Items aus einer Itembank zu schätzen, zu Nutze macht, ist das sogenannte *Linking*, um eine Vergleichbarkeit von unterschiedlichen Fragebögen beziehungsweise Testformen zu erreichen. Eine der möglichen Ansätze eine solche Vergleichbarkeit zu erreichen, ist der Einsatz von IRT [40].

Beim IRT-basierten Linking wird im Prinzip ein unidimensionales IRT-Modell konstruiert, dass die Items der verschiedenen Fragebögen beziehungsweise Testformen enthält. Dazu können zum Beispiel alle Items in einer einzigen Stichprobe (single-group design) erhoben

werden oder zwei Stichproben verbunden werden, indem einige Items in beiden Stichproben (anchor instrument designs) erhoben werden [41]. Ein solches Modell kann anschließend genutzt werden, um die Ausprägung der latenten Variable auf der gleichen Skala zu messen, obwohl verschiedene Fragebögen oder Testformen zur Erhebung genutzt wurden.

Das Patient Reported Outcome Measurement Information System (PROMIS)

Diese oben beschriebenen Beispiele zeigen, wie die Verwendung von IRT eine flexible Erfassung eines Zielkonstrukts ermöglichen kann. Ein besonders bemerkenswertes Projekt, in dessen Rahmen auf Basis der IRT PROs und die dazugehörigen PROMs entwickelt wurden, ist das Patient-Reported Outcome Measurement Information System (PROMIS). Da einige Arbeiten dieser Habilitationsschrift im Rahmen von PROMIS erarbeitet wurden, soll es hier kurz vorgestellt werden.

Seit 2011 wurde in den USA ine Reihe von Itembanken zur Erfassung der gesundheitsbezogenen Lebensqualität aus der Patientenperspektive entwickelt [42, 43]. Dazu zählen körperliche, psychische und soziale Aspekte der Gesundheit (siehe Abbildung 1). Die Entwicklung jeder Itembank erfolgte nach State-of-the-Art-Methodik und schloss Expertenreview, psychometrische Testung und Bewertung durch relevante Patientengruppen ein ([44], siehe unter anderem [45, 46]). Studien belegen die klinische Validität der entwickelten Instrumente (siehe unter anderem [47–49]).

Die Itembanken, die in PROMIS entwickelt wurden, können zur Entwicklung von Kurzformen (short forms), das heißt statischen Fragebögen mit einer festgelegten Anzahl von Items, genutzt werden. Die Auswahl der Items beruht dabei auf theoretischen und empirischen Überlegungen [45]. Um eine präzise, kurze Erhebung über den gesamten Ausprägungsbereich einer Domäne zu erreichen, können die Itembanken auch als Grundlage für die oben beschriebenen CATs dienen.

PROMIS stellt auch einen wichtigen Schritt in Richtung einer stärkeren internationalen Standardisierung der Messung von gesundheitsbezogener Lebensqualität dar. Die PROMIS Itembanken werden weltweit in einem standardisierten Prozess übersetzt [44, 50, 51]. Die deutschen Versionen zeigten in bisherigen Studien gute psychometrische Eigenschaften und hohe Vergleichbarkeit mit anderssprachigen Versionen [52–56].

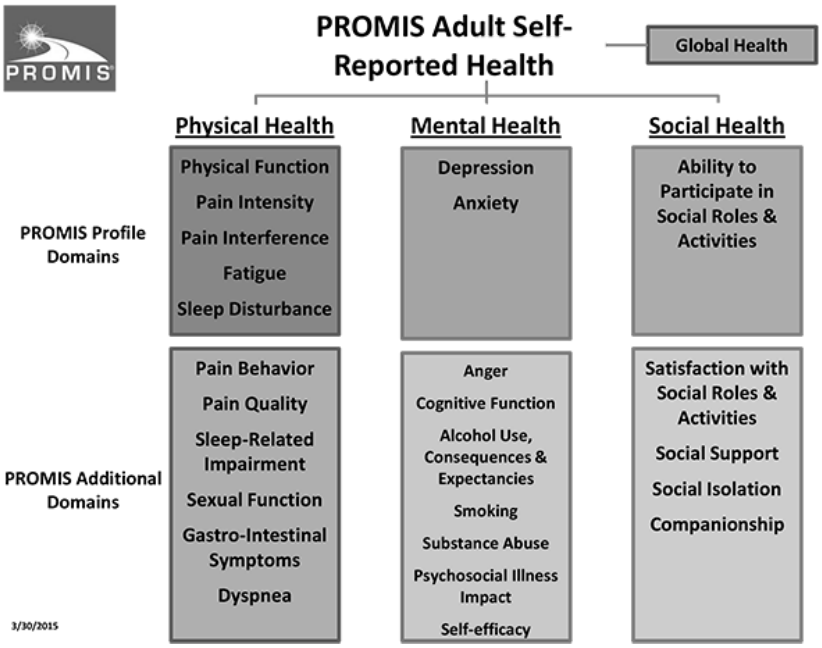

Abbildung 1.4: Das PROMIS Modell der Gesundheit – Kerndomänen gesundheitsbezogener Lebensqualität (Quelle: http://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis)

Im Rahmen des Prosetta Stone Projektes [57] wurden eine Reihe von etablierten Fragebögen auf die jeweilige PROMIS Skala (Depressivität, Angst, Schmerz, körperliche Funktionsfähigkeit, globale Gesundheit) kalibriert [58–61]. PROMIS geht damit über die Entwicklung von spezifischen Meßinstrumenten hinaus und kann auch als instrumentenübergreifende Skalendefinition auf Ebene der latenten Konstruktes verstanden werden. Langfristig dient PROMIS also einer umfassenden Standardisierung der Erhebung der gesundheitsbezogenen Lebensqualität und führt zu einer verbesserten Vergleichbarkeit und Interpretierbarkeit von Daten, die im Rahmen klinischer oder epidemiologischer Forschung über verschiedene Populationen hinweg erhoben werden.

1.3 Fragestellungen

Zusammenfassend lässt sich also sagen, dass die Erhebung von PROs in der medizinischer Forschung und Praxis eine zunehmende Bedeutung hat, um Krankheits- und Therapieverläufe adäquat aus Sicht der Patienten abzubilden. Die IRT erlaubt die Modellierung von PROs als latente Dimensionen ohne die bekannten Schwächen der nach der KTT entwickelten Instru-

mente und ermöglicht so eine flexible, präzise und effektive Erfassung dieser relevanten Zielparameter. Die amerikanische PROMIS-Initiative stellt aufgrund der hohen methodischen Qualität in der Entwicklung einen Meilenstein in der Erfassung der gesundheitsbezogenen Lebensqualität dar.

Die in dieser Habilitationsschrift eingegangenen Arbeiten addressieren zwei Fragestellungen hinsichtliche der standardisierten Erfassung von PROs auf Grundlage der IRT. Zum einen wird der Frage nachgegangen werden, ob mit Hilfe von IRT-Modellen instrumentenübergreifende Skalen auf Ebene des PRO definiert werden können. Dazu wurde eine Metrik zur Erfassung von Depressivität entwickelt, in unabhängigen Stichproben validiert und Methoden zur Schätzung der Ausprägung des latenten Konstrukte verglichen. Zum zweiten stellt sich die Frage, ob PROs über verschiedene Sprachen hinweg definiert und gemessen werden können. Dazu widmen sich zwei Arbeiten dem meßtheoretischen Vergleich der PROMIS Itembank zur Erfassung von Depressivität und dem PROMIS Profile 29 in Stichproben aus Deutschland, Frankreich, Großbritannien und der USA.

2 Originalarbeiten

Die in dieser Habilitationsschrift eingegangenen Originalarbeiten können zwei Themenkomplexen zugeordnet werden.

Drei Arbeiten haben die Entwicklung und Validierung einer instrumentenunabhängigen, konstruktbasierten Skala zum Thema, um eine standardisierte Erfassung von PROs unabhängig vom spezifischen PROM zu ermöglichen. In der ersten der drei Arbeiten wird die Entwicklung eines IRT-Modells beschrieben, dass 11 Depressionsfragebögen auf einer gemeinsamen Skala kalibriert und so eine konstruktbasierte Skala etabliert [62]. Die beiden anderen Arbeiten widmen sich anschließend der Validität dieses Modells. Dabei liegt der Fokus zunächst auf der Frage, inwieweit das Modell in unabhängigen Stichproben anwendbar ist [63]. In der dritten Arbeit wurden dann verschiedene Methoden zur Schätzung der latenten Variable verglichen, die verschieden Fehlerquellen berücksichtigen [64].

Zwei weitere Arbeiten beschäftigen sich dann mit der Standardisierung von Lebensqualitätsmaßen über verschiedene Sprachen und untersuchen die psychometrischen Eigenschaften von Übersetzungen von PROMIS Instrumenten mit der Frage der Vergleichbarkeit von Testwerten über Sprachgrenzen hinweg. Die erste Arbeit fokussiert auf die Übersetzung der PROMIS Itembank zur Depressivität [45] ins Deutsche [51], die zweite auf das PROMIS Profile 29 [65] in Großbritannien, Frankreich und Deutschland.

2.1 Standardisierung über Instrumente

2.1.1 Standardization of depression measurement: a common metric was developed for 11 self-report depression measures

I. Wahl, B. Löwe, J. B. Bjorner, H. F. Fischer, G. Langs, U. Voderholzer, S. A. Aita, N. Bergemann, E. Brähler und M. Rose. "Standardization of depression measurement: a common metric was developed for 11 self-report depression measures". *Journal of Clinical Epidemiology* 67(1) (2014), S. 73–86

Depressivität ist ein zentrales Konstrukt gesundheitsbezogener Lebensqualität - daher stehen zur Erfassung eine unüberschaubare Anzahl von Instrumenten zur Verfügung. Ziel dieser Arbeit war es, etablierte Instrumente (z.B. PHQ-9, HADS, BDI) auf einer gemeinsamen Skala zu kalibrieren, um so Depressivitätswerte unabhängig vom eingesetzten Instrument vergleichen zu können.

Als Datenbasis für die Analyse dienten insgesamt 33,844 Datensätze aus klinischen Stichproben und Normalbevölkerungserhebungen, die zu einem Zeitpunkt mindestens 2 Fragebögen zur Erfassung von Depressivität beantwortet haben. Ein unidimensionales Generalized Partial Credit Model (GPCM) wurde zur Modellierung des Zusammenhangs zwischen latenter Variable und Antwortverhalten geschätzt. Von insgesamt 143 in den verschiedenen Fragebögen enthaltenen Items definierten am Ende 89 Items die latente Dimension, die Parameter der übrigen Items wurden nachträglich geschätzt. Es zeigten sich relevante Unterschiede zwischen den eingeschlossenen Fragebögen hinsichtlich Meßbereich und -genauigkeit.

Die so entwickelte Itembank erlaubt, mit verschiedenen Instrumenten erhobene Daten zur Depressivität auf einer gemeinsamen Skala auszuwerten. Dazu können entweder die Itemparameter des Modells genutzt werden; alternativ werden in der Arbeit die Schätzer der latenten Variable für jeden erreichbaren Summenscore berichtet. Als Beleg für die Validität des Modell wurde gezeigt, dass für verschiedene Instrumente berichtete Grenzwerte zur Trennung von mild, mittel und stark ausgeprägter Depressivität auf der gemeinsamen Skala gut übereinstimmen.

I. Wahl, B. Löwe, J. B. Bjorner, H. F. Fischer, G. Langs, U. Voderholzer, S. A. Aita, N. Bergemann, E. Brähler und M. Rose. “Standardization of depression measurement: a common metric was developed for 11 self-report depression measures”. *Journal of Clinical Epidemiology* 67(1) (2014), S. 73–86

https://doi.org/10.1016/j.jclinepi.2013.04.019

2.1.2 Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation

G. Liegl, I. Wahl, A. Berghöfer, S. Nolte, C. Pieh, M. Rose und H. F. Fischer. "Using PHQ-9 item parameters of a common metric resulted in similar depression scores compared to independent IRT model reestimation". *Journal of Clinical Epidemiology* 71 (2016), S. 25–34

Ziel dieser Arbeit war es, das statistische Modell der zuvor entwickelten Depressionsskala [62] in vier unabhängigen Stichproben aus Deutschland (psychiatrische und psychosomatische Patienten) und Österreich (Patienten aus Allgemeinmedizin und orthopädischer Rehabilitation) zu validieren.

Dazu wurden in jeder der 4 Stichproben (n = 499, 507, 1.049, 1.260) aus dem jeweils erhobenen Depressionsfragebogen (PHQ-9 bzw. PHQ-8) ein IRT-Modell geschätzt und mittels Stocking-Lord Methode [66] oder Anpassung des Priors auf die zuvor entwickelte Depressionsskala skaliert. Aus beiden Modellen wurde dann die Ausprägung der Depressivität jedes Probanden geschätzt und mit der Schätzung der Depressivität anhand der Parameter des zuvor entwickelten Modells verglichen.

In der Analyse zeigte sich, dass sich zwar die Itemparameter zwischen den Stichproben und von der zuvor entwickelten Depressionsmetrik unterschieden, es aber auf Ebene der Schätzung der individuellen Depressivität kaum Unterschiede zwischen den angewendeten Methoden gab. So waren die Differenzen der Depressivität zwischen den Stichproben 16 bis 130 mal größer als die Unterschiede, die auf die Wahl der Schätzmethode zurückzuführen sind.

Zusammenfassend zeigt diese Arbeit, dass die Schätzung der Depressivität anhand eines IRT-Modells, das zuvor und in anderen Stichproben entwickelt wurde, zu keinen praktisch relevanten Unterschieden gegenüber einer Neuschätzung des zugrundeliegenden IRT-Modells führt. Damit liefert die Arbeit Evidenz, dass die Nutzung einer instrumentenunabhängigen Skala in unabhängigen Stichproben möglich ist.

G. Liegl, I. Wahl, A. Berghöfer, S. Nolte, C. Pieh, M. Rose und H. F. Fischer. “Using PHQ-9 item parameters of a common metric resulted in similar depression scores compared to independent IRT model reestimation”. *Journal of Clinical Epidemiology* 71 (2016), S. 25–34

https://doi.org/10.1016/j.jclinepi.2015.10.006

2.1.3 Scoring Depression on a Common Metric: a Comparison of EAP Estimation, Plausible Value Imputation and Full Bayesian IRT Modeling

> H. F. Fischer und M. Rose. "Scoring Depression on a Common Metric: A Comparison of EAP Estimation, Plausible Value Imputation, and Full Bayesian IRT Modeling". *Multivariate Behavioral Research* 54 (2019), S. 85–99

Bei der Verwendung einer instrumentenunabhängigen Skala sollten unterschiedliche Instrumente zu ähnlichen Schätzungen der latenten Variablen kommen - es wurden aber in einer Untersuchung von 461 Patienten mit multipler Sklerose relevante Unterschiede zwischen den verwendeten Instrumenten berichtet [67]. In der folgenden Arbeit haben wir daher untersucht, inwiefern diese Unterschiede durch Berücksichtung (1) des vollständigen Antwortmusters, (2) des Meßfehlers und (3) Unsicherheit der Itemparameter beeinflusst werden können.

In einer Reanalyse der von Kim et al. [67] Daten wurde untersucht, inwieweit die Mittelwerte und Differenzen von Depressivitätsschätzungen auf Gruppenebene von verschiedenen Schätzmethoden beeinflusst werden. Die in Umrechnungstabellen genutzten EAP-Schätzern für Summenscores dienten als Referenz - damit verglichen wurde der EAP-Schätzer für spezifische Antwortmuster, die Imputation plausibler Werte sowie eine vollständige Bayesianische Modellierung.

Als Ergebnis zeigt sich, dass eine Berücksichtigung des vollständigen Antwortmusters keinen relevanten Vorteil gegenüber der Nutzung von auf dem Summenscore beruhenden Schätzern hat. Die Berücksichtigung des Meßfehlers mittels Imputation plausibler Werte dagegen führte zu ca. 10% größeren Standardfehlern der Mittelwerte und Differenzen. Bei zusätzlicher Berücksichtigung der Unsicherheit über die genutzten Itemparameter beobachteten wir einen bessere Modellanpassung und eine relevante Verkleinerung der Differenzen zwischen verschiedenen Instrumenten.

Zusammenfassend läßt sich sagen, dass Umrechnungstabellen für Summenscores mit gutem Gewissen genutzt werden können, der Meßfehlers der Schätzung aber insbesondere im Rahmen konfirmatorischer Studien berücksichtigt werden sollte, um die α-Fehlerwahrscheinlichkeit nicht zu unterschätzen. Bayesianische Methoden bieten darüberhinaus eine langfristige Perspektive, die Itemparameter von konstruktbasierten Skalen anhand neu erhobener Daten zu aktualisieren.

H. F. Fischer und M. Rose. “Scoring Depression on a Common Metric: A Comparison of EAP Estimation, Plausible Value Imputation, and Full Bayesian IRT Modeling”. *Multivariate Behavioral Research* 54 (2019), S. 85–99

https://doi.org/10.1080/00273171.2018.1491381

2.2 Standardisierung über Sprachen

2.2.1 Language-related differential item functioning between English and German PROMIS Depression items is negligible

H. F. Fischer, I. Wahl, S. Nolte, G. Liegl, E. Brähler, B. Löwe und M. Rose. "Language-related Differential Item Functioning between English and German PROMIS Depression Items is negligible". *International Journal of Methods in Psychiatric Research* 26(4) (2016), e1530

Vergleiche von Testwerten zwischen der Originalversion der PROMIS Itembank zur Depressivität und ihrer deutschen Übersetzung können durch Differential Item Functioning (DIF), das heißt systematisch unterschiedliche Itemparameter, verzerrt sein. Ziel dieser Arbeit war es, das Ausmaß eines solchen Bias zu quantifizieren.

Dazu wurden Daten aus der PROMIS Wave 1 Stichprobe (n = 780), einer Erhebung der deutschen Normalbevölkerung (n = 2.500) und einer klinischen Stichprobe (n = 621) genutzt. Mit Hilfe ordinaler logistischer Regressionen wurden die erwarteten Itemscores auf Basis der geschätzten latenten Depressivität modelliert - Kriterien für DIF waren eine Zunahme von Nagelkerkes' R^2 um 0.02 Punkte sowie Raju's N (NCDIF) größer als 0.096. Itemparameter für Items, die diese Kriterien erfüllten, wurden in den deutschen Stichproben neu geschätzt. Zur Bewertung der Relevanz von DIF wurden Schätzungen der latenten Depressivität auf Basis dieses korrigierten Modells mit denen auf Basis US-amerikanischer Itemparametern verglichen. Zur Berücksichtigung des Meßfehlers wurden plausible Werte imputiert.

Insgesamt erfüllten nur 4 von 28 Items die Kriterien für DIF. Wurde DIF in der Schätzung der latenten Depressivität berücksichtigt, ergaben sich vernachlässigbare Unterschiede zum US-Modell im Falle der vollen Itembank sowie bei der Simulation von CATs ($_{\Delta}\theta < 0.1$). Bei Kurzformen mit 4,6 beziehungsweise 8 Items war der Effekt klein ($_{\Delta}\theta < 1$).

Es zeigen sich lediglich kleine Unterschiede in den Meßeigenschaften der PROMIS Itembank für Depressivität zwischen US-amerikanischen und deutschen Stichproben, sodass man von einer prinzipiellen Vergleichbarkeit der Testwerte ausgehen kann. Dies gilt insbesondere, da man durch den Einsatz der IRT bei der Erhebung auf Items mit DIF verzichten oder die korrigierten Itemparameter nutzen kann.

H. F. Fischer, I. Wahl, S. Nolte, G. Liegl, E. Brähler, B. Löwe und M. Rose. “Language-related Differential Item Functioning between English and German PROMIS Depression Items is negligible”. *International Journal of Methods in Psychiatric Research* 26(4) (2016), e1530

https://doi.org/10.1002/mpr.1530

2.2.2 Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France, and Germany

> H. F. Fischer, C. Gibbons, J Coste, J. M. Valderas, M. Rose und A. Leplege. "Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France and Germany". *Quality of Life Research* 27(4) (2018), S. 999–1014

Das PROMIS Profile 29 ist ein Fragebogen, der Kerndomänen der gesundheitsbezogenen Lebensqualität (Körperliche Funktionsfähigkeit, Angst, Depression, Fatigue, Schlafbeeinträchtigung, Beeinträchtigung durch Schmerzen, Teilhabe an sozialen Aktivitäten) erfasst. Ziel der Arbeit war es, die psychometrische Äquivalenz des Fragebogens zwischen Großbritannien, Frankreich und Deutschland und das Ausmaß etwaiger Gesundheitsunterschiede zu untersuchen sowie Referenzwerte aus der Normalbevölkerung bereitzustellen.

Neben dem PROMIS Profile 29 wurden soziodemographische Daten und der EQ-5D-5L als Maß der gesundheitsbezogenen Lebensqualität in repräsentativen Stichproben von jeweils circa 1.500 Personen aus den drei Ländern beantwortet. Meßinvarianz über die Sprachversionen wurde mittels konfirmatorischer Faktorenanalysen untersucht. Unterschiede im wahrgenommenen Gesundheitszustand wurden mit linearen Regressionen, die Verteilungen der jeweiligen Zielparameter in der Normalbevölkerung mit Quantilregressionen modelliert.

Es zeigt sich, dass ein Modell mit der theoretisch angenommene Struktur von 7 korrelierten Faktoren gut an die Daten angepasst werden kann und auch strenge Annahmen der Meßinvarianz keine nennenswerten Effekte auf die Schätzungen der latenten Variablen haben. Mittelwertsunterschiede in den Domänen des PROMIS Profile 29 zwischen den Ländern waren eher klein und konnten zum Teil durch Unterschiede in den Gesundheitsratings des EQ-5D-5L erklärt werden. Durch den Einsatz von plausiblen Werten können die berichteten Referenzwerte für die Normalbevölkerung auch für andere PROMIS Short Forms oder CATs genutzt werden. Die Modellierung mit Quantilregression erlaubt auch Referenzwerte für bestimmte Subgruppen zu konstruieren, zum Beispiel stratifiziert nach Geschlecht oder Bildungsgrad.

Das PROMIS Profile 29 ist also ein für länderübergreifende Studien geeignetes Instrument zur Erfassung der gesundheitsbezogenen Lebensqualität, da die erhobenen Daten über Sprachversionen vergleichbar sind und flexibel nutzbare Referenzwerte zur Verfügung stehen.

H. F. Fischer, C. Gibbons, J Coste, J. M. Valderas, M. Rose und A. Leplege. “Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France and Germany”. *Quality of Life Research* 27(4) (2018), S. 999–1014

https://doi.org/10.1007/s11136-018-1785-8

3 Diskussion

Im folgenden werden die Ergebnisse der in dieser Habilitationschrift eingegangen Arbeiten in den größeren Forschungszusammenhang eingeordnet. Abschließend sollen die Perspektiven einer weiteren Standardisierung der Erhebung von PROs diskutiert werden.

3.1 Einordnung der Ergebnisse in den Forschungskontext

Die in dieser Arbeit vorgestellten Arbeiten sind im Kontext eines Paradigmenwechsels bei der Messung gesundheitsbezogener Lebensqualität zu sehen. Bisher gab es keine klare Trennung zwischen Patient-Reported Outcome (PRO) und Patient-Reported Outcome Measure (PROM), da die zugrundeliegende zu messende Dimension im Prinzip erst durch die Entwicklung des betreffenden Meßinstruments definiert wurde. Durch die Nutzung probabilistischer Methoden der IRT wird eine vom Meßinstrument unabhängige Definition eines PRO möglich und verschiedene PROMs können genutzt werden, um das gleiche PRO zu messen.

Die Entwicklung der Depressionsmetrik [62] baut dabei auf früheren Arbeiten auf, in denen Umrechnungstabellen einzelner PROMs entwickelt wurden [68, 69]. Ein wesentlicher Unterschied zu den früheren Arbeiten besteht aber nicht nur in der Anzahl der eingeschlossenen PROMs, sondern auch darin, dass die zugrundeliegende Skala (das PRO Depressivität) empirisch definiert wurde. So wurden Items, die Nebenaspekte des PROs abbilden (wie Schlafstörungen, Appetitverlust), nicht zur Definition des statistischen Modells genutzt, sondern im Nachgang in das Modell integriert. Einen etwas anderen Ansatz verfolgt das Projekt Prosetta Stone [57]. Hier werden existierende PROMs auf die in PROMIS definierte Skala kalibriert (siehe zum Beispiel [58, 61]). Im Falle von Depressivität sind beide Definitionen allerdings

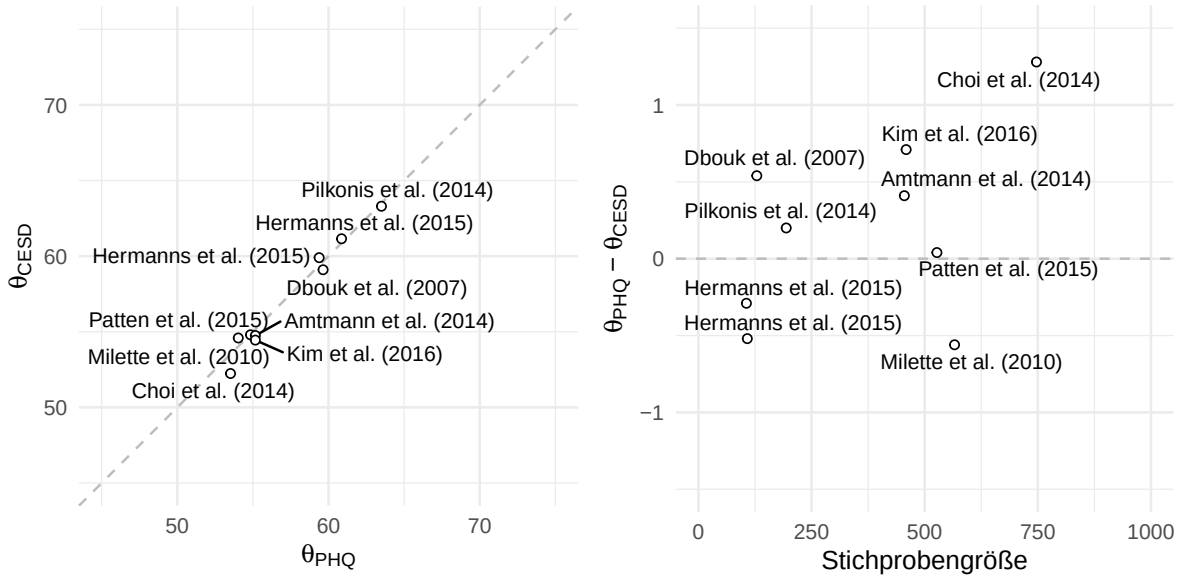

Abbildung 3.1: Differenzen der Depressivität (θ, jeweils geschätzt aus PHQ-9 und CESD und auf die PROMIS Depression Metrik skaliert) in 8 verschiedenen Studien. Zur Auswahl der Studien und Methodik siehe Chung [70]

inhaltlich ähnlich und fokussieren auf die kognitiv-affektive Aspekte depressiver Erkrankungen. Ein Vergleich beider Metriken in unabhängigen Stichproben, der Aufschluss über mögliche Unterschiede zwischen den beiden Ansätze geben könnte, steht leider noch aus.

Zur Genauigkeit bei der Kalibrierung anderer Depressivitätsmaße auf die PROMIS Metrik gibt es allerdings erste Befunde. Wie berichtet zeigten sich in der Arbeit von Kim et al. [67] relevante Unterschiede in der Schätzung der latenten Depressivität, wenn verschiedene Depressionsfragebögen zum Einsatz kamen. Eine australische Arbeit spricht dagegen für die Validität der instrumentenübergreifenden PROMIS Depressivitätsskala [71]. Weitere Evidenz für die Validität liefert eine unveröffentlichten Hausarbeit, die im Rahmen des Modellstudiengangs Medizin an der Charité entstanden ist [70]. Hier konnte gezeigt werden, dass in acht veröffentlichten Studien die Differenz zwischen den Schätzungen der Depressiviät gemessen mit dem PHQ-9 beziehungsweise dem CESD in den meisten Fällen weniger als 1 Punkt auf der von PROMIS verwendeten T-Metrik (M: 50, SD: 10) beträgt und damit vernachlässigbar klein ist. Darüberhinaus ist kein Zusammenhang zwischen Stichprobengröße und beobachteter Differenz zu erkennen (siehe Abbildung 3.1). Einschränkend muss aber gesagt werden, dass auch bei Kim et al. [67] die Differenz zwischen PHQ-9 und CESD kleiner war als die Differenz zu der verwendetet PROMIS Kurzform. Gegenwärtig ist noch offen, ob die beobach-

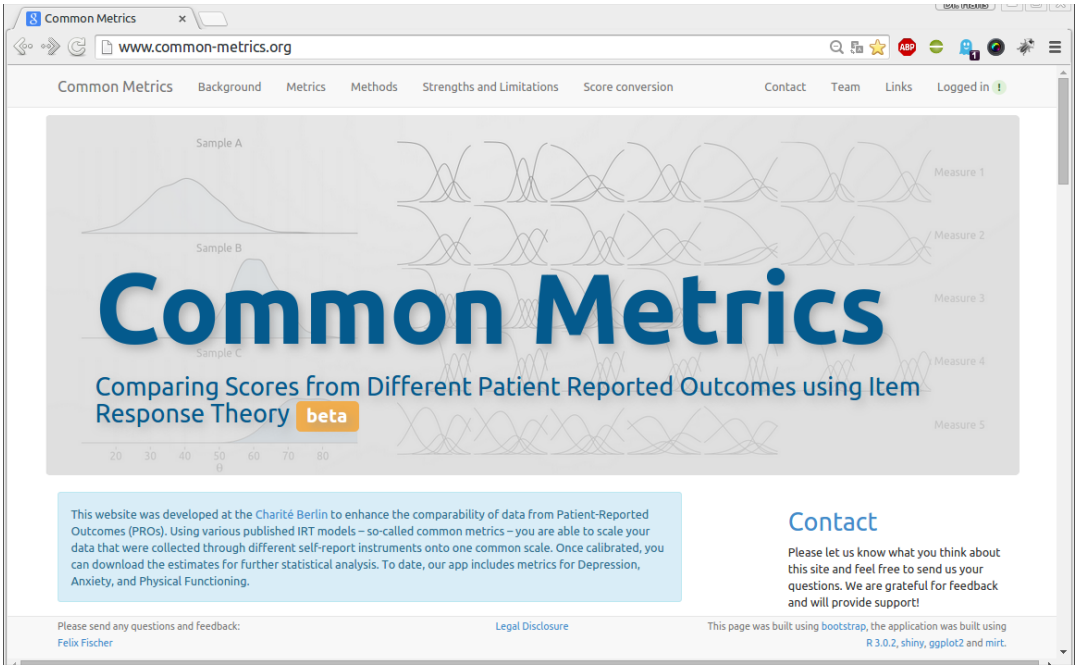

Abbildung 3.2: www.common-metrics.org: eine Website zur Schätzung latenter Konstrukte auf instrumentenunabhängigen Skalen

teten Unterschiede Ausdruck systematischer Fehler, also auf einen Mangel an Validität der Modelle, zurückzuführen sind oder ob sie Ausdruck eines unsystematischen Fehlers sind.

Eine Herausforderung bei der Implementierung instrumentenunabhängiger Skalen in Forschung und klinischer Praxis ist die Schätzung der latenten Variable auf der jeweiligen Skala für die einzelne Person. Dies setzt psychometrisches Spezialwissen und -software voraus. Ein häufig gewählter Weg ist es daher, die latente Variable für alle möglichen Summenwerte zu berechnen und diese in Umrechnungstabellen zu berichten [72]. Dies hat den Vorteil, dass eine einfache Ersetzung von Summenscore mit dem entsprechenden Schätzwert der latenten Variable erfolgen muss, allerdings geht dabei die im spezifischen Antwortmuster enthaltene Information verloren. Um die Schätzung der latenten Variable auf Basis aller verfügbaren Informationen und auch im Falle fehlender Antworten auch einem breiteren Publikum möglich zu machen, wurde eine Website zur Verfügung gestellt (siehe Abbildung 3.2) [73]. Allerdings konnte zuletzt auch gezeigt werden, dass die Berücksichtigung des kompletten Antwortmusters gegenüber dem Summenscore keinen großen Einfluss auf die Schätzung von Populationsmittelwert und -streuung hat [64].

Zusammengenommen erscheinen auf IRT basierende, instrumentenunabhängige Skalen als ein vielversprechender Weg, die Erfassung von PROs zu standardisieren. Neben der in die-

Domäne	Land	MW	SD	Referenz
Depression	Niederlande	49.6	10.0	[79]
Depression	Spanien	48.8	9.3	[80]
Angst	Niederlande	49.9	10.1	[79]
Zufriedenheit mit sozialer Rollenfunktion	Niederlande	47.5	8.3	[78]
Teilhabe an sozialen Rollen und Aktivitäten	Niederlande	50.6	9.5	[78]

Tabelle 3.1: Mittelwerte verschiedener PROMIS Domänen in europäischen Normalbevölkerungsstichproben

ser Arbeit vorgestellten Metrik [62] sind insbesondere die in Prosetta Stone entwickelten Modelle und Umrechnungstabellen zu erwähnen [57] – erste Daten zur Anwendung solcher Modelle belegen ihre Validität und Nützlichkeit.

Hinsichtlich der Standardisierung von PROs über verschiedene Sprachen zeigt sich, dass die hohen Qualitätskriterien, die bei der Übersetzung der PROMIS Instrumente angelegt werden [44], im Ergebnis zu einer guten Vergleichbarkeit der verschiedenen Sprachversionen führen. Neben den beiden hier berichteten Arbeiten [55, 56], zeigen bisherige Studien, dass zwischen verschiedensprachige Versionen von PROMIS Instrumenten kein klinisch relevanter Bias im Sinne von DIF beobachtet werden konnte. Dies gilt für die niederländischen Versionen der Domänen körperliche Funktionsfähigkeit und Beeinträchtigung durch Schmerzen [74–76] sowie die spanische [77] und die niederländische [78] Version der Domäne soziale Rollenfunktion. Auch für die Domänen Angst und Depressivität wurde vernachlässigbares DIF für die niederländische [79] und spanische [80] Version berichtet. Eine Ausnahme ist die Itembank zur körperlichen Funktionsfähigkeit, wo ungefähr ein Drittel der spanischsprachigen Items DIF zu den englischen Items zeigten [81]. Insgesamt können diese Befunde als Beleg dafür interpretiert werden, dass PROMIS Itembanken und die daraus abgeleiteten Kurzformen und CATs für valide Vergleiche über Sprachen nutzbar sind.

Betrachtet man allerdings die Mittelwerte verschiedener PROMIS Domänen in Normalbevölkerungsstichproben in Europa (siehe Tabelle 3.1 und Fischer et al. [55]), wird deutlich, dass zur statistischen Absicherung von Unterschieden zwischen diesen Stichproben DIF in der Größenordnung eines Punktes, welcher gemeinhin als vernachlässigbar interpretiert wird, durchaus relevant sein kann. Vergleiche zwischen Ländern müssen daher die Möglichkeit von Verzerrungen durch DIF in die Analyse explizit einbeziehen. Ein Vorteil IRT-basierter Analysen ist, dass eine solche Berücksichtigung möglich ist, indem zum Beispiel für die Items,

die DIF zeigen, sprachspezifische Itemparameter eingesetzt werden. Die – bis auf geringe Abweichungen – konsistente Replizierbarkeit des PROMIS Meßmodells in vielen verschiedenen Sprachen spricht aber für die prinzipielle Validität der in PROMIS entwickelten Meßinstrumente und ihrer Übersetzungen.

3.2 Perspektiven zukünftiger Forschung

Die in dieser Habilitationsschrift eingegangenen Arbeiten haben also gezeigt, dass mittels IRT eine Standardisierung der Erhebung von PROs sowohl über Sprachen als auch über Instrumente möglich ist. Im letzten Abschnitt dieser Arbeit soll nun auf die Perspektiven weiterer Forschung eingegangen werden.

Wie oben beschrieben sind bisher nur erste Befunde über die Validität konstruktbasierter Skalen publiziert worden. Eine offene Frage ist zum Beispiel, inwieweit die verfügbaren instrumentenunabhängigen Skalen über verschiedene Sprachen hinweg valide sind. Allerdings zeigt die oben beschriebene Hausarbeit [70] exemplarisch auf, wie bereits erhobene und publizierte Daten zur Validierung von konstruktbasierten Skalen genutzt werden können. Dies ist insbesondere interessant, da mittels der verfügbaren Umrechnungstabellen auch aggregierte Daten wie Mittelwerte zur Validierung genutzt werden können. So erscheint eine metaanalytische Validierung von konstruktbasierten Skalen möglich und aufgrund der hohen externen Validität auch sinnvoll.

Neben einer umfassenden Validierung ist eine Erweiterung bestehender instrumentenunabhängiger Skalen um weitere, häufig verwendete Fragebögen wünschenswert. Die größte Herausforderung besteht dabei in der Sammlung von Daten für die notwendigen statistischen Analysen. Eine vielversprechende Perspektive ist auch hier die Nutzung bereits erhobener Daten. Es ist nicht unüblich, innerhalb einer Studie mehrere Instrumente zur Erfassung des gleichen Outcomes zu nutzen – eine Sammlung der Rohdaten solcher Studien wäre eine gute Datenbasis für weitere Forschung, da hier realistische Nutzungsszenarien von PROs abgebildet sind. Im Zuge einer erwartbaren häufigeren computergestützten Erhebung von PROs in der klinischen Routine wäre auch denkbar, Datenerhebungen zum Linking von Fragebögen in diesem Kontext durchzuführen.

Auch die statistischen Modelle, die zur Definition instrumentenunabhängiger Skalen eingesetzt werden, können in der Zukunft erweitert werden. Die bisher verwendeten IRT-Modelle

gehen von festen (fixed) Itemparametern aus, obwohl – zum Beispiel im Rahmen von Untersuchungen zu DIF – die Problematik irrelevanter Abweichungen von diesen Modellparametern in anderen Stichproben deutlich wurde. Eine Perspektive, solche Variationen über Stichproben in die verwendeten Modelle einzubeziehen, wäre die explizite Modellierung von zufälligen (random) Itemparametern. Die Schätzung eines kompletten IRT-Modells mit zufälligen Itemparametern wurden bisher lediglich im Rahmen der PISA-Studie durchgeführt und beschränkte sich auf dichotome Itemantworten im Rasch-Modell [82, 83].

Bayesianische Ansätze bieten aber auch bei der Anwendung in einer einzelnen Stichprobe die Möglichkeit, Wissen über mögliche irrelevante Variation der Modellparameter in die Schätzung der latenten Variable einfließen zu lassen. Dazu wären insbesondere Studien notwendig, die diese Varianz untersuchen und so die Konstruktion adäquater Priors auf die jeweiligen Modellparameter informieren können. Es ergibt sich also, dass sowohl für die Entwicklung von konstruktbasierten Skalen, die eine Erhebung von PROs mit verschiedenen PROMs ermöglichen, als auch für die sprachenübergreifende Vergleichbarkeit von PROs und PROMs, Bayesianische IRT-Modelle eine Perspektive bieten, klinisch irrelevante, aber durchaus erwartbare, stichprobenabhängige Unterschiede der Itemparameter in der statistischen Modellierung zu berücksichtigen. Die Digitalisierung der Erhebung von PROs in klinischen Studien und Routineversorgung bietet dabei eine Chance, die Datenbasis für eine Weiterentwicklung solcher Modelle zu legen.

Darüberhinaus wurden die Grundannahmen der in dieser Arbeit genutzten reflexiven Modelle, nämlich dass eine nicht direkt beobachtbare Variable für das beobachtete Antwortverhalten verantwortlich ist, in Frage gestellt, zum Beispiel für das Konstrukt Depressivität [84, 85]. Die Netzwerkanalyse bietet hier eine andere Perspektive an, indem Symptome, die als Items erhoben werden, über kausale Pfade miteinander verbunden sind [86]. Bestimmte Netzwerkeigenschaften konnten zum Beispiel die Vorhersage von Therapieabbrüchen verbessern [87], allerdings ist bisher die Stabilität solcher Netzwerkmodelle noch wenig erforscht [88]. Ein Schwerpunkt zukünftiger Forschung sollte daher sein, Vor- und Nachteile dieser beiden konkurrierenden Ansätze zu explorieren, um eine valide und präzise Messung der Gesundheit aus Patientenperspektive zu ermöglichen.

4 Zusammenfassung

Die Erhebung patientenberichteter Endpunkte ist in klinischer Forschung und Versorgung gleichermaßen relevant. Die gegenwärtig genutzten Instrumente zur Erhebung dieser Endpunkte sind allerdings wenig standardisiert, was die Vergleichbarkeit der erhobenen Daten einschränkt. Methoden der Item-Response Theory bieten die Möglichkeit, eine Standardisierung der Erhebung sowohl über Instrumente als auch über Sprachen zu erreichen.

Im Rahmen dieser Habilitationsschrift wird zum einen die Entwicklung und Validierung einer konstruktbasierten, instrumentenunabhängigen Skala auf Basis eines probabilistischen Testmodells zur Erhebung von Depressivität beschrieben. Dabei zeigte sich, dass eine Messung der latenten Variable Depressivität mit verschiedenen Instrumenten möglich ist, Modellparameter in unabhängigen Stichproben angewendet werden können und Bayesianische Methoden genutzt werden können, um Modellparameter anhand neu erhobener Daten zu aktualisieren.

Zum zweiten wird die Vergleichbarkeit von PROMIS Instrumenten über verschiedene Sprachen untersucht. Dabei zeigt sich, dass sowohl für die PROMIS Itembank zur Erhebung von Depressivität als auch im PROMIS Profile 29, das die zentralen Gesundheitsdomänen vereint, die Unterschiede in Itemparametern über Sprachen vernachlässigbar gering sind. Somit ist ein Vergleich der Testwerte über verschiedene Sprachversionen der getesteten Instrumente hinweg valide.

Weitere Befunde bestätigen die Validität konstruktbasierter Skalen und auch die Ergebnisse zur Vergleichbarkeit über Sprachen wurden für andere Domänen beziehungsweise Sprachen berichtet. Eine Perspektive zur Weiterentwicklung der verwendeten Modelle sind Bayesianische IRT-Modelle, die klinisch irrelevante Unterschiede in den Itemparametern in verschiedenen Stichproben explizit modellieren können.

Literatur

[1] World Health Organization. *WHOQOL: measuring quality of life*. 1997.

[2] International Society for Quality of Life Research. *Dictionary of Quality of Life and Health Outcomes Measurement*. Hrsg. von N. Mayo. Milwaukee, MI: ISOQOL, 2015.

[3] I. B. Wilson und P. D. Cleary. "Linking Clinical Variables with Health-Related Quality of Life". *JAMA* 273(1) (1995), S. 59–65.

[4] E. M. Roos, H. P. Roos, L. S. Lohmander, C. Ekdahl und B. D. Beynnon. "Knee Injury and Osteoarthritis Outcome Score (KOOS) –Development of a Self-Administered Outcome Measure". *Journal of orthopaedic & Sports Physical Therapy* 78(2) (1998), S. 88–96.

[5] R. L. Spitzer, K. Kroenke und J. B. W. Williams. "Validation and Utility of a Self-report Version of PRIME-MD: The PHQ Primary Care Study". *JAMA* 282(18) (1999), S. 1737–1744.

[6] K. Kroenke, R. L. Spitzer, J. B. W. Williams und B. Löwe. "The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review." *General Hospital Psychiatry* 32(4) (2010), S. 345–59.

[7] B. D. Thombs, E. Arthurs, S. Coronado-montoya, M. Roseman, V. C. Delisle, A. Leavens, B. Levis, L. Azoulay, C. Smith, L. Ciofani, J. C. Coyne, N. Feeley, S. Gilbody, J. Schinazi, D. E. Stewart und P. Zelkowitz. "Depression screening and patient outcomes in pregnancy or postpartum : A systematic review". *Journal of Psychosomatic Research* 76(6) (2014), S. 433–446.

[8] B. D. Thombs, M. Roseman, J. C. Coyne, P. de Jonge, V. C. Delisle, E. Arthurs, B. Levis und R. C. Ziegelstein. "Does evidence support the American Heart Association's recommendation to screen patients for depression in cardiovascular care? An updated systematic review." *PloS one* 8(1) (2013), e52654.

[9] J. H. Lichtman, J. T. Bigger, J. A. Blumenthal, N. Frasure-Smith, P. G. Kaufmann, F. Lespérance, D. B. Mark, D. S. Sheps, C. B. Taylor und E. S. Froelicher. "Depression and coronary heart disease: recommendations for screening, referral, and treatment". *Circulation* 118(17) (2008), S. 1768–75.

[10] J. W. H. Kocks, J. W. K. van den Berg, H. A. Kerstjens, S. M. Uil, J. M. Vonk, Y. P. de Jong, I. G. Tsiligianni und T. van der Molen. "Day-to-day measurement of patient-reported outcomes in exacerbations of chronic obstructive pulmonary disease". *International Journal of Chronic Obstructive Pulmonary Disease* 8 (2013), S. 273–286.

[11] S. K. Berg, C. B. Thorup, B. Borregaard, A. V. Christensen, L. Thrysoee, T. B. Rasmussen, O. Ekholm, K. Juel und M. Vamosi. "Patient-reported outcomes are independent predictors of one-year mortality and cardiac events across cardiac diagnoses: Findings from the national DenHeart survey". *European Journal of Preventive Cardiology* (2018).

[12] E. Basch, A. M. Deal, M. G. Kris, H. I. Scher, C. A. Hudis, P. Sabbatini, L. Rogak, A. V. Bennett, A. C. Dueck, T. M. Atkinson, J. F. Chou, D. Dulko, L. Sit, A. Barz, P. Novotny, M. Fruscione, J. A. Sloan, D. Schrag, C. A. Hudis und P. Sabbatini. "Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment : A Randomized Controlled Trial". *Journal of Clinical Oncology* 34(6) (2016).

[13] L. Squitieri, K. J. Bozic und A. L. Pusic. "The Role of Patient-Reported Outcome Measures in Value-Based Payment Reform". *Value in Health* 20(6) (2017), S. 834–836.

[14] Gemeinsamer Bundesausschuss. *Verfahrensordnung*. 2017.

[15] Institut für Wirtschaftlichkeit im Gesundheitswesen. *Allgemeine Methoden*. 5. Aufl. 2017.

[16] A. Maaz, M. H.-J. Winter und A. Kuhlmey. "Der Wandel des Krankheitspanoramas und die Bedeutung chronischer Erkrankungen (Epidemiologie, Kosten)". *Fehlzeiten-Report 2006: Chronische Krankheiten*. Hrsg. von B. Badura, H. Schellschmidt und C. Vetter. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, S. 5–23.

[17] S. H. Van Oostrom, R. Gijsen, I. Stirbu, J. C. Korevaar, F. G. Schellevis, H. S. J. Picavet und N. Hoeymans. "Time trends in prevalence of chronic diseases and multimorbidity not only due to aging: Data from general practices and health surveys". *PLoS ONE* (2016).

[18] D. Plass, T. Vos, C. Hornberg, C. Scheidt-Nave, H. Zeeb und A. Krämer. "Entwicklung der Krankheitslast in Deutschland: Ergebnisse, potenziale und grenzen der Global Burden of disease-studie". *Deutsches Arzteblatt International* 111(38) (2014), S. 629–638.

[19] R.-D. Stieglitz. *Diagnostik und Klassifikation in der Psychiatrie*. Stuttgart: Kohlhammer, 2008.

[20] I. Wahl, B. Meyer, B. Löwe und M. Rose. "Die Erfassung der Lebensqualität in der Psychotherapieforschung". *Klinische Diagnostik und Evaluation* 3(1) (2010), S. 4–21.

[21] M. Bühner. *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium, 2010.

[22] S. E. Embretson und S. P. Reise. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[23] M. D. Reckase. *Multidimensional Item Response Theory*. Dordrecht: Springer, 2009.

[24] R. P. Chalmers. "mirt: A Multidimensional Item Response Theory Package for the R Environment". *Journal of Statistical Software* 48(6) (2012).

[25] R Development Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2008.

[26] J. B. Bjorner, C.-H. Chang, D. Thissen und B. B. Reeve. "Developing tailored instruments: item banking and computerized adaptive assessment." *Quality of Life Research* 16 Suppl 1 (2007), S. 95–108.

[27] D. Thissen und H Wainer. *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates, 2001.

[28] R. K. Hambleton und R. W. Jones. "Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development". *Educational Measurement: Issues and Practice* 12(3) (1993), S. 38–47. arXiv: 1305.4210.

[29] T. Forkmann, M. Boecker, M. Wirtz, N. Eberle, M. Westhofen, P. Schauerte, K. Mischke, T. Kircher, S. Gauggel und C. Norra. "Development and validation of the Rasch-based depression screening (DESC) using Rasch analysis and structural equation modelling". *Journal of Behavior Therapy and Experimental Psychiatry* 40(3) (2009), S. 468–478.

[30] W. J. Van der Linden und C. A. W. Glas. *Computerized Adaptive Testing: Theory and Practice*. Hrsg. von W. J. Linden und G. A. Glas. Dordrecht: Kluwer Academic Publishers, 2000.

[31] D. Cella, R. Gershon, J.-S. Lai, S. W. Choi, S. Yount, N. Rothrock, K. F. Cook, B. B. Reeve, D. Ader, J. F. Fries, B. Bruce und M. Rose. "The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment." *Quality of Life Research* 16 Suppl 1(5 Suppl 1) (2007), S. 133–41.

[32] H. F. Fischer, C. Klug, K. Roeper, E. Blozik, F. Edelmann, M. Eisele, S. Störk, R. Wachter, M. Scherer, M. Rose und C. Herrmann-Lingen. "Screening for mental disorders in heart failure patients using computer-adaptive tests". *Quality of Life Research* 23(5) (2014), S. 1609–1618.

[33] H. Fliege, J. Becker, O. B. Walter, J. B. Bjorner, B. F. Klapp und M. Rose. "Development of a computer-adaptive test for depression (D-CAT)." *Quality of Life Research* 14(10) (2005), S. 2277–2291.

[34] T. Forkmann, U Kroehne, M Wirtz, C Norra, H Baumeister, S Gauggel, A. H. Elhan, A Tennant und M Boecker. "Adaptive screening for depression–recalibration of an item bank for the assessment of depression in persons with mental and somatic diseases and evaluation in a simulated computer-adaptive test environment". eng. *J Psychosom Res* 75(5) (2013), S. 437–443.

[35] O. B. Walter, J. Becker, H. Fliege, J. B. Bjorner, M. Kosinski, M. Walter, B. F. Klapp und M. Rose. "Entwicklungsschritte für einen computeradaptiven Test zur Erfassung von Angst (A-CAT)". *Diagnostica* 51(2) (2005), S. 88–100.

[36] B. Abberger, A. Haschke, M. Wirtz, U. Kroehne, J. Bengel und H. Baumeister. "Development and evaluation of a computer adaptive test to assess anxiety in cardiovascular rehabilitation patients". *Archives of physical medicine and rehabilitation* 94(12) (2013), S. 2433–2439.

[37] R.-D. Kocalevent, M. Rose, J. Becker, O. B. Walter, H. Fliege, J. B. Bjorner, D. Kleiber und B. F. Klapp. "An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception." *Journal of Clinical Epidemiology* 62(3) (2009), S. 278–287.

[38] J. Devine, C. Otto, M. Rose, D. Barthel, H. F. Fischer, H. Mühlan, S. Nolte, S. Schmidt, V. Ottova-Jordan, U. Ravens-Sieberer und H Mülhan. "A new Computerized Adaptive Test advancing the measurement of Health-Related Quality of Life (HRQoL) in children: The Kids-CAT". *Quality of Life Research* 24(4) (2015), S. 871–884.

[39] D. Barthel, K. Fischer, S. Nolte, C. Otto, A.-K. Meyrose, S. Reisinger, M. Dabs, U. Thyen, M. Klein, H. Muehlan, T. Ankermann, O. Walter, M. Rose und U. Ravens-Sieberer. "Implementation of the Kids-CAT in clinical settings: a newly developed computer-adaptive test to facilitate the assessment of patient-reported outcomes of children and adolescents in clinical practice in Germany." *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 25(3) (2016), S. 585–594.

[40] M. J. Kolen und R. L. Brennan. *Test Equating, Scaling, and Linking*. New York: Springer, 2014.

[41] N. J. Dorans. "Linking scores from multiple health outcome instruments." *Quality of Life Research* 16 Suppl 1(December 2006) (2007), S. 85–94.

[42] D. Cella, S. Yount, N. Rothrock, R. Gershon, K. F. Cook, B. B. Reeve, D. Ader, J. F. Fries, B. Bruce und M. Rose. "The patient-reported outcomes measurement information system (PROMIS) – Progress of an NIH Roadmap Cooperative Group During its First Two Years". *Medical Care* 45(5) (2007), S. 3–11.

[43] B. B. Reeve, R. D. Hays, J. B. Bjorner, K. F. Cook, P. K. Crane, J. A. Teresi, D. Thissen, D. A. Revicki, D. J. Weiss, R. Hambleton, H. Liu, R. Gershon, S. P. Reise, J.-S. Lai und D. Cella. "Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS)." *Medical Care* 45(5 Suppl 1) (2007), S. 22–31.

[44] Patient-Reported Outcomes Measurement Information System. *PROMIS Instrument Development and Validation Scientific Standards Version 2.0*. 2013.

[45] P. A. Pilkonis, S. W. Choi, S. P. Reise, A. M. Stover, W. T. Riley und D. Cella. "Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): depression, anxiety, and anger." *Assessment* 18(3) (2011), S. 263–83.

[46] D. Amtmann, K. F. Cook, M. P. Jensen, W.-H. Chen, S. W. Choi, D. A. Revicki, D. Cella, N. Rothrock, F. Keefe und L. Callahan. "Development of A PROMIS Item Bank to Measure Pain Interference". *Pain* 150(1) (2010), S. 173–182.

[47] K. F. Cook, S. E. Jensen, B. D. Schalet, J. L. Beaumont, D. Amtmann, S. Czajkowski, D. A. Dewalt, J. F. Fries, P. A. Pilkonis, B. B. Reeve, A. A. Stone, K. P. Weinfurt und D. Cella. "PROMIS Measures of Pain, Fatigue, Negative Affect, Physical Function and Social Function Demonstrate Clinical Validity across a Range of Chronic Conditions". *Journal of Clinical Epidemiology* 73 (2016), S. 89–102.

[48] B. D. Schalet, P. A. Pilkonis, L. Yu, N. Dodds, K. L. Johnston, S. Yount, W. Riley und D. Cella. "Clinical validity of PROMIS Depression, Anxiety, and Anger across diverse clinical samples". *Journal of Clinical Epidemiology* 73 (2016), S. 119–127.

[49] B. D. Schalet, R. D. Hays, S. E. Jensen, J. L. Beaumont, J. F. Fries und D. Cella. "Validity of PROMIS physical function measures in diverse clinical samples". *Journal of Clinical Epidemiology* 73 (2016), S. 112–118.

[50] S. L. Eremenco, D. Cella und B. J. Arnold. "A comprehensive method for the translation and cross-cultural validation of health status questionnaires". *Evaluation and the Health Professions* 28(2) (2005), S. 212–232.

[51] I. Wahl, B. Löwe und M. Rose. "Das Patient-Reported Outcomes Measurement Information System (PROMIS): Übersetzung der Item-Banken für Depressivität und Angst ins Deutsche". *Klinische Diagnostik und Evaluation* 4 (2011), S. 236–261.

[52] G. Liegl, M. Rose, H. Correia, H. F. Fischer, S. Kanlidere, A. Mierke, A. Obbarius und S. Nolte. "An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions". *Clinical Rehabilitation* (2017), S. 026921551771429.

[53] T. Jakob, M. Nagl, L. Gramm, K. Heyduck, E. Farin und M. Glattacker. "Psychometric Properties of a German Translation of the PROMIS(R) Depression Item Bank". *Evaluation & the Health Professions* 40(1) (2015), S. 106–120.

[54] E. Farin, M. Nagl, L. Gramm, K. Heyduck und M. Glattacker. "Development and evaluation of the PI-G: a three-scale measure based on the German translation of the PROMIS pain interference item bank." *Quality of Life Research* 23(4) (2013), S. 1255–1265.

[55] H. F. Fischer, C. Gibbons, J Coste, J. M. Valderas, M. Rose und A. Leplege. "Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France and Germany". *Quality of Life Research* 27(4) (2018), S. 999–1014.

[56] H. F. Fischer, I. Wahl, S. Nolte, G. Liegl, E. Brähler, B. Löwe und M. Rose. "Language-related Differential Item Functioning between English and German PROMIS Depression Items is negligible". *International Journal of Methods in Psychiatric Research* 26(4) (2016), e1530.

[57] S. W. Choi, T. Podrabsky, N. McKinney, B. D. Schalet, K. F. Cook und D. Cella. *Prosetta Stone Methodology – A Rosetta Stone for Patient Reported Outcomes.* 2015.

[58] S. W. Choi, B. D. Schalet, K. F. Cook und D. Cella. "Establishing a Common Metric for Depressive Symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression." *Psychological Assessment* 26(2) (2014), S. 513–527.

[59] B. D. Schalet, D. A. Revicki, K. F. Cook, E. Krishnan, J. F. Fries und D. Cella. "Establishing a Common Metric for Physical Function: Linking the HAQ-DI and SF-36 PF Subscale to PROMIS Physical Function". *Journal of General Internal Medicine* 30(10) (2015), S. 1517–1523.

[60] A. J. Kaat, M. E. Newcomb, D. T. Ryan und B. Mustanski. "Expanding a common metric for depression reporting: linking two scales to PROMIS depression". *Quality of Life Research* 26(5) (2017), S. 1119–1128.

[61] K. F. Cook, B. D. Schalet, M. A. Kallen, J. P. Rutsohn und D. Cella. "Establishing a common metric for self-reported pain: linking BPI Pain Interference and SF-36 Bodily Pain Subscale scores to the PROMIS Pain Interference metric". *Quality of Life Research* 24(10) (2015), S. 2305–18.

[62] I. Wahl, B. Löwe, J. B. Bjorner, H. F. Fischer, G. Langs, U. Voderholzer, S. A. Aita, N. Bergemann, E. Brähler und M. Rose. "Standardization of depression measurement: a common metric was developed for 11 self-report depression measures". *Journal of Clinical Epidemiology* 67(1) (2014), S. 73–86.

[63] G. Liegl, I. Wahl, A. Berghöfer, S. Nolte, C. Pieh, M. Rose und H. F. Fischer. "Using PHQ-9 item parameters of a common metric resulted in similar depression scores compared to independent IRT model reestimation". *Journal of Clinical Epidemiology* 71 (2016), S. 25–34.

[64] H. F. Fischer und M. Rose. "Scoring Depression on a Common Metric: A Comparison of EAP Estimation, Plausible Value Imputation, and Full Bayesian IRT Modeling". *Multivariate Behavioral Research* 54 (2019), S. 85–99.

[65] Patient-Reported Outcomes Measurement Information System. *PROMIS Short Form Scoring Manual*. 2013.

[66] M. L. Stocking und F. M. Lord. "Developing a Common Metric in Item Response Theory". *Applied Psychological Measurement* 7(2) (1983), S. 201–210.

[67] J. Kim, H. Chung, R. L. Askew, R. Park, S. M. W. Jones, K. F. Cook und D. Amtmann. "Translating CESD-20 and PHQ-9 Scores to PROMIS Depression". *Assessment* 24(3) (2017), S. 300–307.

[68] H. F. Fischer, K. Tritt, B. F. Klapp und H. Fliege. "How to compare scores from different depression scales: equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using Item Response". *International Journal of Methods in Psychiatric Research* 20(4) (2011), S. 203 –214.

[69] P. M. ten Klooster, M. A. H. Oude Voshaar, B. Gandek, M. Rose, J. B. Bjorner, E. Taal, C. Glas, P. L.C. M. van Riel und M. A.F. J. van de Laar. "Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment

Questionnaire disability index in rheumatoid". *Health and quality of Life outcomes* 11 (2013), S. 199.

[70] J. Y. Chung. "Validation of common metrics to measure depression with PHQ-9 and CES-D: a review of literatures". Unveröffentlichte Hausarbeit. 2017.

[71] M. Batterham, P. Sunderland, A. Carragher und N. Calear. "Validity of the PROMIS depression and anxiety common metrics in an online sample of Australian adults". *Quality of Life Research* (2018), epub first.

[72] D. Thissen, M. Pommerich, K. Billeaud und V. S. L. Williams. "Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses". *Applied Psychological Measurement* 19(1) (1995), S. 39–49.

[73] H. F. Fischer und M. Rose. "www.common-metrics.org: a web application to estimate scores from different patient-reported outcome measures on a common scale". *BMC Medical Research Methodology* 16(142) (2016).

[74] M. A. H. Oude Voshaar, P. M. ten Klooster, C. Glas, H. E. Vonkeman, E. Taal, E. Krishnan, H. J. B. Moens, M. Boers, C. B. Terwee, P. L.C. M. van Riel und M. A.F. J. van de Laar. "Calibration of the PROMIS Physical Function Item Bank in Dutch Patients with Rheumatoid Arthritis." *PloS one* 9(3) (2014), e92367.

[75] M. H. P. Crins, L. D. Roorda, N. Smits, H. C. W. de Vet, R. Westhovens, D. Cella, K. F. Cook, D. A. Revicki, J. van Leeuwen, M. Boers, J. Dekker und C. B. Terwee. "Calibration and Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients with Chronic Pain". *Plos One* 10(7) (2015), e0134094.

[76] W. Schuller, C. B. Terwee, T. Klausch, L. D. Roorda, D. C. Rohrich, R. W. Ostelo, B. Terluin und H. C. W. de Vet. "Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients With Musculoskeletal Complaints." *Spine* (2018), epub first.

[77] E. A. Hahn, D. A. DeWalt, R. K. Bode, S. F. Garcia, R. F. Devellis, H. Correia und D. Cella. "New English and Spanish Social Health Measures Will Facilitate Evaluating Health Determinants." *Health Psychology* 33(5) (2014), S. 490–499.

[78] C. B. Terwee, M. H. P. Crins, M. Boers, H. C. W. de Vet und L. D. Roorda. "Validation of two PROMIS item banks for measuring social participation in the Dutch general population". *Quality of Life Research* (2018), epub first.

[79] J. van Bebber, G. Flens, J. T. Wigman, E. de Beurs, S. Sytema, L. Wunderink und R. R. Meijer. "Application of the Patient-Reported Outcomes Measurement Information System (PROMIS) item parameters for Anxiety and Depression in the Netherlands". *International Journal of Methods in Psychiatric Research*(2) (2018), e1744.

[80] G. Vilagut, C. G. Forero, J. I. Castro-Rodriguez, E. Olariu, G. Barbaglia, M. Astals, C. Diez-Aja, M. Gárriz, A. Abellanas, J. M. López-Santín, C. Sanchez-Gil und J. Alonso. "Measurement equivalence of PROMIS depression in Spain and the United States." *Psychological Assessment* (2018), epub first.

[81] S. H. Paz, K. L. Spritzer, L. S. Morales und R. D. Hays. "Evaluation of the Patient-Reported Outcomes Information System (PROMIS) Spanish-language physical functioning items." *Quality of Life Research* 22(7) (2013), S. 1819–1830.

[82] J. P. Fox. *Bayesian item response modeling: Theory and applications*. New York: Springer, 2010.

[83] A. J. Verhagen und J. P. Fox. "Bayesian tests of measurement invariance". *British Journal of Mathematical and Statistical Psychology* 66(3) (2013), S. 383–401.

[84] E. I. Fried, C. D. van Borkulo, S. Epskamp, R. A. Schoevers, F. Tuerlinckx und D. Borsboom. "Measuring depression over time... or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression." *Psychological Assessment* 28(11) (2016), S. 1354–1367.

[85] E. I. Fried, R. M. Nesse, K. Zivin, C. Guille und S. Sen. "Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors." *Psychological medicine*(December) (2013), S. 1–10. arXiv: 15334406.

[86] V. D. Schmittmann, A. O. Cramer, L. J. Waldorp, S. Epskamp, R. A. Kievit und D. Borsboom. "Deconstructing the construct: A network perspective on psychological phenomena". *New Ideas in Psychology* 31(1) (2013), S. 43–53.

[87] W. Lutz, B. Schwartz, S. G. Hofmann, A. J. Fisher, K. Husen und J. A. Rubel. "Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study". *Scientific Reports* 8(1) (2018), S. 1–9.

[88] S. Epskamp, D. Borsboom und E. I. Fried. "Estimating psychological networks and their accuracy : A tutorial paper". *Behavior Research Methods* 50 (2018), S. 195–212.

Danksagung

Großen Anteil an der Entstehung dieser Arbeit haben meine akademischen Lehrer. Ich danke Prof. Upmeyer, in dessen Arbeitsgruppe an der TU Berlin ich erste Schritte in der wissenschaftlichen Welt gehen konnte. Ich danke Prof. Klapp (†) und PD Dr. Fliege für die Unterstützung bei meiner Promotion. Prof. Willich & Prof. Witt danke ich für die Jahre am Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie, in denen ich mein methodisches Wissen vertiefen konnte. Prof. Rose danke ich für die große Unterstützung und den weiten Rahmen, meine Forscherpersönlichkeit entwickeln zu können.

Neben meinen akademischen Lehrern haben mich in den letzten Jahren unzählige Kollegen unterstützt. Ich danke Oliver Fellmann, Nina Gerard, Anne Ahnis, Fatma Atalay, Janine Devine, Petra Georgiewa, Anne Grimm, Rüya Kocalevent, Ute Meissner, Hedda Neumann, Christina Papachristou, Iris Bartsch, Anne Berghöfer, Sylvia Binting, Benno Brinkhaus, Linus Grabenhenrich, Stefanie Helmer, Katja Icke, Thomas Keil, Ralf Krause, Lilian Krist, Julia Ostermann, Andreas Reich, Thomas Reinhold, Alizee Rogge, Stephanie Roll, Ulrike Stasun, Barbara Stöckigt, Kathrin Fischer, Tobias Hoffmann, Paul Klapproth, Gregor Liegl, Annett Mierke, Sandra Nolte, Alexander und Nina Obbarius, Eva Winter und vielen anderen.

Ich danke außerdem all meinen Kooperationspartnern für die Möglichkeit, an spannenden und lehrreichen Projekten mitzuarbeiten, den anonymen Reviewern für ihre konstruktiven Verbesserungsvorschläge und den unsichtbaren Händen, die einem das schöne Leben im Elfenbeinturm erst ermöglichen.

Mein größter Dank aber gilt Justus, Lukas und Karla, die mich gelehrt haben, das Glück nicht erst zu suchen, sondern einfach im Moment zu finden.

Eidesstattliche Versicherung

Hiermit erkläre ich, dass

- weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wurde,
- die vorgelegte Habilitationsschrift ohne fremde Hilfe verfasst, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden,
- mir die geltende Habilitationsordnung bekannt ist.

Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

Datum, Unterschrift