Adaptives Testen in der Rehabilitation – ein Weg zur ökonomischen Erhebung von Patientenmerkmalen*

Adaptive Testing in Rehabilitation – a Way to Economic Assessment of Patients' Attributes

Autoren C. Frey[1], C. Zwingmann[2], M. Böcker[3], T. Forkmann[3], U. Kröhne[4], E. Müller[1], M. Wirtz[1]

Institute
[1] Institut für Psychologie, Pädagogische Hochschule Freiburg
[2] Evangelische Fachhochschule Rheinland-Westfalen-Lippe, Bochum
[3] Institut für Medizinische Psychologie und Medizinische Soziologie, Rheinisch-Westfälische Technische Hochschule Aachen
[4] Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main

Schlüsselwörter
- outcome-messung
- Item-Response-Theorie
- Computer-adaptives Testen (CAT)

Key words
- adaptive testing
- item response theory
- computer-adaptive testing (CAT)

Zusammenfassung

In der Diagnostik und der Outcome-Messung werden in der Rehabilitation eine Vielzahl von Fragebogen eingesetzt. In der Regel werden dabei – unabhängig vom aktuellen Gesundheitszustand und vom Schweregrad – für jeden Rehabilitanden dieselben Informationen mittels identischer Gruppen von Items erfasst, um die Vergleichbarkeit der diagnostischen Befunde gewährleisten zu können. Hierbei erweist es sich als problematisch, dass a) in der Regel nur im mittleren Belastungsspektrum eine hinreichend genaue und zuverlässige Diagnostik möglich ist und b) Messinstrumente durch die Vorgabe von vielen und im Einzelfall uninformativen Items eine unökonomische Datenerhebung erfordern. In diesem Beitrag wird gezeigt, wie durch adaptives oder maßgeschneidertes Testen die Datenerhebung an der individuellen Belastung des Patienten selbst orientiert werden kann. Hierdurch können im Idealfall eine hohe Ökonomie und eine optimierte Messpräzision der Testung und patientenseitig eine hohe Akzeptanz gewährleistet werden, da eine gute Passung zwischen Fähigkeit der Patienten und erfragten Iteminhalten gegeben ist. Adaptive Testverfahren erfordern jedoch eine aufwändige und sehr sorgfältige psychometrische Entwicklung der Erhebungsprozeduren, damit trotz Vorgabe unterschiedlicher Items an die Patienten die diagnostischen Befunde zwischen Patienten und Messzeitpunkten (z.B. Prä-Post-Vergleich) vergleichbar sind. In diesem Beitrag wird gezeigt, welche Entwick-

Abstract

For diagnostics and outcome measurement in clinical rehabilitation a multitude of questionnaires is used. In order to gain comparability of the diagnostic findings, generally, the same information is gathered of all patients, regardless of their state of health or how severely ill they are, by using identical groups of items. In this kind of assessment it is, however, problematic that (a) the assessment instrument usually only allows for adequate and reliable diagnostics of patients who suffer from injuries or illnesses of medium severity, and (b) that an uneconomic data collection is required because of an extensive set of items which may in individual cases be uninformative and unnecessary. This article shows how data assessment can be adjusted to the individual handicaps of the patient by using adaptive or tailored testing. Thus it can be ensured that the testing becomes more economical and results more precise. Furthermore, a high acceptance by the patients participating in the data collection can be achieved, as the test items submitted fit the ability of the patients adequately. Yet, adaptive test systems require a complex and very careful psychometrical development of the assessment procedure in such a way that in spite of presenting different items to patients, diagnostic outcomes are comparable between patients and for different points in time (e.g., pre-post comparison). This article shows the steps of development that have to be accomplished in order to set up a psychometrically sound adaptive test procedure. Moreover, it shows how its usage in the area of medical rehabilitation can be profitable for handlers and patients.

Bibliografie
DOI http://dx.doi.org/10.1055/s-0031-1271816
Rehabilitation 2011; 50: 195–203
© Georg Thieme Verlag KG Stuttgart · New York
ISSN 0034-3536

Korrespondenzadresse
Prof. Dr. Markus Wirtz
Institut für Psychologie
Pädagogische Hochschule Freiburg
Kunzenweg 21
79117 Freiburg
markus.wirtz@ph-freiburg.de

* **Koordinatoren der Reihe „Methoden in der Rehabilitationsforschung“:**
Prof. Dr. Dr. Hermann Faller, Würzburg; Prof. Dr. Thomas Kohlmann, Greifswald; Prof. Dr. Dr. Christian Zwingmann, Bochum
Interessenten, die einen Beitrag zur Reihe beisteuern möchten, werden gebeten, vorab Kontakt aufzunehmen, E-Mail: christian.zwingmann@web.de.

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

lungsschritte eine fundierte adaptive Testprozedur durchlaufen muss und wie der Einsatz auch in der medizinischen Rehabilitation zum Gewinn für Anwender und Patient werden kann.

Problemstellung

Um in der Rehabilitation die spezifische Eingangsbelastung der Patienten ermitteln und Veränderungen bzw. Verbesserungen feststellen zu können, ist der Einsatz psychometrischer Testverfahren oftmals unumgänglich. Bei der Auswahl der Testverfahren bewegen sich aber die meisten Anwender in einem Spannungsfeld: Wie sehr soll auf Belange der Patienten eines möglichst wenig aufwändigen Instrumentes eingegangen werden? Welche Mindestanforderungen müssen die Gütekriterien des Tests (Gültigkeit und Zuverlässigkeit der Ergebnisse) erfüllen? Wie sehr muss auf die Vergleichbarkeit von Patienteninformationen geachtet werden? Meist werden Instrumente eingesetzt, die sich bewährt haben. Oftmals empfiehlt sich aber ein gründliches Nachdenken über die Eigenschaften des eingesetzten Tests, wie die folgende Situation zeigt.

Ein multimorbider Patient soll psychometrisch getestet werden. Dabei sind die Anforderungen, welche die Beteiligten an die Testung selbst und die Testdurchführung haben, recht unterschiedlich. 2 Beispiele für eine solche Anforderungssituation sind:

- Patientin A ist 65 Jahre alt, seit ungefähr 15 Jahren leidet sie unter Osteoporose mit starken Beschwerden in der Wirbelsäule und in der Schulter. Sie steht am Anfang der orthopädischen Rehabilitation. Sie ist in ihrer Funktionsfähigkeit im Alltag stark eingeschränkt.
- Dem Patienten B (42 Jahre) wurde vor 5 Wochen ein künstliches Hüftgelenk implantiert. Er hat bisher von der Rehabilitation profitieren können und damit seine Einschränkungen im alltäglichen Leben verringert.

Die Funktionsfähigkeit im Alltag beschreibt ein Konstrukt, das zum Inhalt hat, wie gut eine Person es mit ihren Einschränkungen (z.B. körperlich) schafft, alltäglichen Anforderungen (wie aufstehen, waschen, sich bewegen, sich selbst versorgen) gerecht zu werden.

Die Einschränkungen der beiden Patienten sind wohl kaum mit denselben Instrumenten darstellbar. Bei Patientin A ist ein Test nötig, der auch starke Beeinträchtigungen im Alltag noch erfassen kann. Patient B hingegen benötigt ein Messinstrument, das Einschränkungen auf einem weitaus geringeren Beeinträchtigungsniveau erfassen kann. Für beide Patienten sollten Instrumente gewählt werden, die optimal zur aktuellen Situation passen, aber auch im Verlauf der Behandlung noch eingesetzt werden können, d.h. auch Verbesserungen abbilden können.

Verfügbare Instrumente gehen entweder den Anforderungen der Kliniker bzw. Fragebogenentwickler nach, es handelt sich dann meist um umfangreiche Instrumente, die nicht primär den Interessen der Patienten entsprechen. Oder aber es werden sehr kurze Instrumente entwickelt, die nicht allen gängigen Kriterien für ein gutes Messinstrument genügen, dafür aber den Patienten nur wenig beanspruchen. In diesem Spannungsfeld unterschiedlicher Anforderungen bewegen sich Anwender von Messinstrumenten [1].

Ein besonderes Problem aber, das die meisten gängigen Messinstrumente nicht ausreichend lösen können, ist, dass sie sich auf einen bestimmten Bereich an Merkmalsausprägungen bzw. Leistungseinschränkungen konzentrieren und dabei die übrigen Bereiche vernachlässigen. Beispielhaft sei dies für den Merkmalsbereich „Funktionsfähigkeit im Alltag“ erläutert. Wie bei den Beispielpatienten A und B beschrieben, gibt es hier verschiedene Ausprägungen: Patientin A ist sehr stark eingeschränkt, Patient B ist nur noch wenig eingeschränkt. Beide befinden sich damit, wie in **Abb. 1** dargestellt, an unterschiedlichen Stellen eines Merkmalskontinuums zwischen den Polen „0“ (maximale Einschränkung) und „100“ (volle Funktionsfähigkeit).

Abb. 1 veranschaulicht auch, dass es Instrumente gibt, die sich auf den oberen Merkmalsbereich konzentrieren, und andere Instrumente, die den unteren Merkmalsbereich erfassen. Mit diesen Instrumenten kann man entweder Patientin A oder Patient B gut diagnostizieren, beide zusammen mit einem Instrument hingegen nicht. Es treten dann bei Personen, die für den Test zu fähig bzw. zu wenig fähig sind, Decken- bzw. Bodeneffekte auf.

Ein Test konzentriert sich meist entweder auf einen mittleren Merkmalsbereich oder aber auf besonders hohe oder besonders niedrige Bereiche. Messinstrumente haben meist zum Ziel, entweder durch möglichst genaue Messungen in einem spezifischen Bereich eine gezielte Diagnostik zu ermöglichen (z.B. bei Patienten mit begründetem Verdacht auf eine Störung) oder aber durch eine weniger genaue Messung in einem breiten Bereich ein Screening durchzuführen (z.B. in der Bevölkerung nach Patienten mit einer bestimmten Erkrankung zu suchen). Beide Eigenschaften – Messung eines breiten Merkmalsbereiches *und* präzise, aussagekräftige Messergebnisse – können die meisten Messinstrumente nicht vorweisen.

Aktuell werden in der Rehabilitation mehrere Tests zur Feststellung von Fähigkeitseinschränkungen eingesetzt. Beispielhaft seien hier aus der Neurologie der FIM (Functional Independence Measure, deutsch: Funktionale Selbstständigkeitsmessung) und der Barthel-Index aufgeführt, die beide der Erfassung von Selbstständigkeit dienen. Beides sind relativ kurze Instrumente (FIM: 18 Fragen [2]; Barthel-Index: 10 Fragen [3,4]), sie sind angemessen theoretisch basiert und verfügen über beachtliche Messgenauigkeit (außer Fremdbeurteilung beim Barthel-Index). Ein Nachteil beider Verfahren ist aber, dass sie sich besonders für Patienten mit einer mittleren bis geringen Selbstständigkeitsausprägung eignen und weniger für Patienten mit einer hohen Ausprägung (für FIM siehe [5]). Problematisch ist der Einsatz

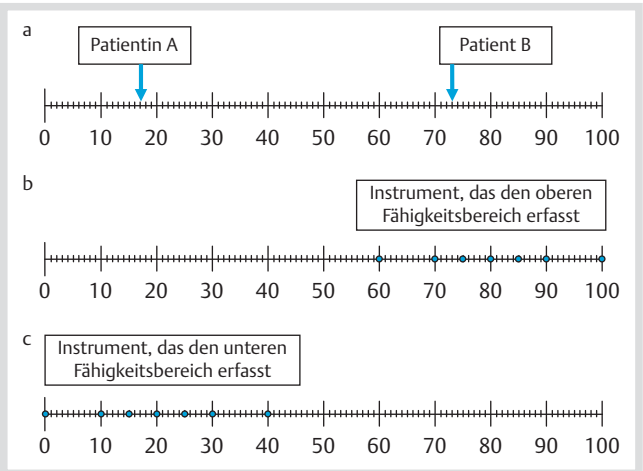

Abb. 1 Beispiel zur Passung von Eigenschaften von Erhebungsskalen und Patientenfähigkeiten im Bereich „Funktionsfähigkeit im Alltag“: **a**) Merkmalskontinuum und Lokalisation der Beispielpatienten, **b**) Messinstrument, das aus Items mit Schwierigkeiten (Punkte) im oberen Merkmalsbereich besteht, **c**) Messinstrument, das aus Items mit Schwierigkeiten (Punkte) im unteren Merkmalsbereich besteht.

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

dieser Instrumente folglich für eine Verlaufsmessung im gesamten Rehabilitationsprozess. Auch die meisten anderen der in der Rehabilitation eingesetzten Messverfahren können das hier aufgezeigte Problem nicht lösen und damit nicht allen Anforderungen an ein ideales Messinstrument genügen.

Empirische Analysen zeigen, dass ökonomische Skalen stets nur in eingeschränkten Belastungsbereichen reliabel messen können [5–8]. Auch bei Vergleichen sehr aufwändig entwickelter Kurzskalen konnte gezeigt werden, dass sie der in diesem Beitrag vorgestellten adaptiven Vorgehensweise nur als gleichwertig angesehen werden können, wenn Patienten Belastungen in bestimmten Merkmalsbereichen aufweisen [7,8]: Eine Lösung dieses Problems könnte nur darin bestehen, dass verschiedene Kurzskalen in verschiedenen Belastungsbereichen eingesetzt werden [7]. Dies würde aber ebenfalls ein adaptives Vorgehen implizieren, ähnlich wie es im Folgenden beschrieben wird.

Prinzip des adaptiven Testens

Das adaptive Testen kann einen Ausweg aus dem Dilemma aufzeigen. Adaptives Testen bedeutet, dass sich der Test an die Antworten der Person, die den Test bearbeitet, anpasst. Der Test versucht, sich an die Fähigkeit der Person heranzutasten.

Dies ist z. B. bei jeder guten mündlichen Prüfung der Fall: Der Prüfer stellt Fragen; kann der Prüfling diese dann sehr gut beantworten, tastet sich der Prüfer an die Grenzen des Wissens des Prüflings heran (falls diese existieren) und stellt dann immer schwierigere Fragen. Läuft die Prüfung für den Prüfling aber nicht so gut, kann er also wenige Fragen beantworten, wird der Prüfer immer leichtere Fragen auswählen, um die Wissensbasis des Prüflings aufzufinden.

Im Rahmen der Gesundheitsversorgung entspricht dieses Vorgehen am ehesten dem klinischen Interview. Hier tastet sich der Interviewer ebenfalls an die Besonderheiten des Patienten heran: In welchen Bereichen treten Symptome auf, wo muss differenzierter nachgefragt werden, welche Bereiche sind für diesen Patienten von geringer Relevanz? Werden dann in einem Bereich Beschwerden geäußert, wird differenzierter nachgefragt, wie stark die Beschwerden sind und welche Einschränkungen diese mit sich bringen. Dieses Vorgehen ist optimal an die Eigenschaften des Patienten angepasst, es handelt sich also um ein adaptives Vorgehen.

Das adaptive Testen verfolgt ebenfalls das Ziel, die Fragen an die gemessene Eigenschaft des Patienten anzupassen, dabei aber nicht den zeitlichen und personellen Aufwand eines klinischen Interviews zu benötigen. Grundgedanke des adaptiven Testens ist es, viel mehr Fragen im Repertoire zu haben, als dem Patienten gestellt werden.

Idealerweise sollte ein Test Items in allen Schwierigkeitsbereichen enthalten, um Patienten jeder Fähigkeitsstufe einordnen zu können. **Abb. 2** stellt ein solches Instrument grafisch dar. Die Gesamtheit aller Items, die für die Erfassung eines Merkmals zur Verfügung stehen (aber nicht alle bei allen Patienten zur Anwendung kommen), wird als Itembank bezeichnet.

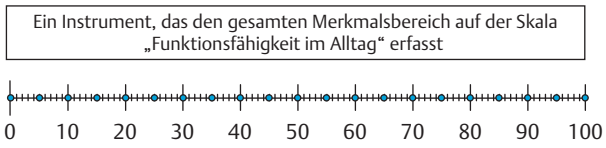

Abb. 2 Messinstrument, das Items (Punkte entsprechen deren Schwierigkeit) im gesamten Merkmalsbereich aufweist.

Aus dieser Vielzahl an Fragen, die potenziell zur Verfügung stehen, können die für den Patienten relevanten ausgewählt werden. Damit kann das oben dargestellte Problem umgangen werden. Patientin A und Patient B können mit demselben Instrument erfasst werden, auch Veränderungen der beiden Patienten können mit einem solchen Instrument gemessen werden.

Beim adaptiven Testen werden Fragen aussortiert, die für den Patienten irrelevant sind, genauso wie beim klinischen Interview ein Patient, der Ängste generell verneint, nicht nach Ängsten unterschiedlicher Ausprägungen gefragt wird.

Fragen können dann irrelevant sein, wenn sie

- durch eine vorangegangene, allgemeine Frage schon beantwortet wurden,
- durch bestimmte Merkmale des Patienten ausgeschlossen sind (z. B. die Frage nach Treppensteigen bei einem Rollstuhlfahrer),
- nicht zur Situation des Patienten passen, weil
  - die Fragen den Patienten überfordern (z. B. die Frage, ob jemand 1 km joggen kann, wenn schon verneint wurde, dass diese Person 1 km gehen kann) oder
  - die Fragen den Patienten unterfordern (z. B. die Frage, ob jemand 1 km gehen kann, wenn schon bejaht wurde, dass diese Person 1 km joggen kann).

Im Rahmen eines adaptiven Tests sollten alle 3 Bedingungen für irrelevante Fragen integriert sein, besonderes Augenmerk soll im Folgenden aber auf den dritten Punkt gelegt werden: Wie kann sichergestellt werden, dass Fragen, die für den Patienten zu einfach oder zu schwierig sind, vermieden werden?

Die Umsetzung eines solchen adaptiven Tests kann auf verschiedenen Wegen realisiert werden. Zum einen können durch Sprungbefehle oder vorab festgelegte Verzweigungsregeln, welche in die Fragebogen fest eingearbeitet werden, irrelevante Fragen übersprungen werden. Zum anderen gibt es zunehmend die Tendenz, sich die Möglichkeiten des Computers zunutze zu machen und den Computer über einen zuvor definierten Algorithmus die passenden Fragen aussuchen zu lassen (sog. Computeradaptives Testen, kurz CAT, [7]). Der hier vorliegende Artikel beschäftigt sich zentral mit der Vorgehensweise im Rahmen einer flexiblen adaptiven Vorgabe von Fragen an Patienten, bei der jedes Mal aufs Neue entschieden wird, welche Frage als nächstes vorgelegt werden soll. Dies ist ausschließlich im Rahmen eines CAT zu realisieren.

Bei der Testdurchführung sitzt der Patient vor einem Computer und beantwortet durch Klicken mit der Maus die ihm vorgelegten Fragen. Auf Grundlage der Antworten des Patienten wählt der Rechner sodann die nächsten Fragen aus. Wichtig ist dabei, nur Fragen zu wählen, die für den Patienten mit seinen Fähigkeiten relevant sind, und irrelevante Fragen auszusortieren.

Abb. 3 verdeutlicht das Vorgehen. Gestartet wird mit einer Frage aus dem mittleren Bereich; diese kann von der Patientin (vgl. Beispielpatientin A) aufgrund ihrer Einschränkungen nicht gelöst werden. Das zweite Item ist ebenfalls zu schwierig, das dritte kann gelöst werden. Die Schwierigkeit des Items liegt unter der Fähigkeit der Patientin. Das nächste Item (Item 4) ist schwieriger als Item 3, aber leichter als Item 2 und kann wiederum nicht gelöst werden. Bei Item 5 ist die Passung zwischen Personenfähigkeit und Itemschwierigkeit dann gut, die Fähigkeit der Patientin kann ausreichend gut geschätzt und die Testung beendet werden.

Das geschilderte Vorgehen ist vereinfacht dargestellt; es kann auch mit einer gewissen Wahrscheinlichkeit vorkommen, dass

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

ein Patient ein zu schwieriges Item lösen bzw. ein zu leichtes Item nicht lösen kann. Somit kann es auch zu weniger idealen Verläufen bei einem adaptiven Test kommen, wobei aber immer das Ziel verfolgt wird, die Schwierigkeit des Items der (geschätzten) Fähigkeit der Person anzupassen.

Abb. 3 stellt dar, wie sich beim adaptiven Testen die vorgelegten Fragen an den vorangegangenen Antworten der Person orientieren. Es wird angestrebt, immer eine Passung zwischen den Schwierigkeiten der Items und dem wahren Fähigkeitswert der Person zu erreichen. Wie dies dann konkret für einen Patienten aussieht, verdeutlicht **Abb. 4**. Dargestellt sind hier die Fragen, die ein adaptiver Test der Beispielpatientin A nacheinander vorlegt, um die Fähigkeit dieser Person im Bereich „Funktionsfähigkeit im Alltag" zu erfassen. Die Reihenfolge der 5 Items in **Abb. 4** korrespondiert dabei mit der Nummerierung in **Abb. 3**.

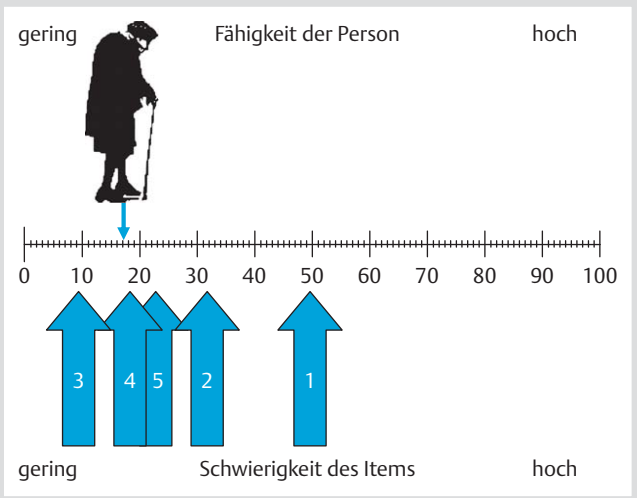

Abb. 3 Annäherung der Itemschwierigkeiten an die Fähigkeit der Person im Rahmen einer adaptiven Testung (Anmerkung: Eine solche Verortung von Schwierigkeiten und Fähigkeiten auf einer Skala ist das Ergebnis einer sorgfältigen Analyse auf Grundlage der Item Response Theory, IRT; s. weiter unten).

Bitte kreuzen Sie an, in welchem Maße Sie momentan Schwierigkeiten haben, folgendes zu tun	(1) Unmöglich	(2) Starke Schwierigk.	(3) Mäßige Schwierigk.	(4) Geringe Schwierigk.	(5) Keine Schwierigk.
15 Minuten körperlich arbeiten (wie z. B. Staub saugen, Rasen mähen)	☒	☐	☐	☐	☐
Bitte kreuzen Sie an, in welchem Maße Sie momentan Schwierigkeiten haben, folgendes zu tun	**(1) Unmöglich**	**(2) Starke Schwierigk.**	**(3) Mäßige Schwierigk.**	**(4) Geringe Schwierigk.**	**(5) Keine Schwierigk.**
Einen schweren Gegenstand über 5 Meter tragen	☒	☐	☐	☐	☐
Bitte kreuzen Sie an, in welchem Maße Sie momentan Schwierigkeiten haben, folgendes zu tun	**(1) Unmöglich**	**(2) Starke Schwierigk.**	**(3) Mäßige Schwierigk.**	**(4) Geringe Schwierigk.**	**(5) Keine Schwierigk.**
Aus dem Bett aufstehen	☐	☐	☐	☒	☐
Bitte kreuzen Sie an, in welchem Maße Sie momentan Schwierigkeiten haben, folgendes zu tun	**(1) Unmöglich**	**(2) Starke Schwierigk.**	**(3) Mäßige Schwierigk.**	**(4) Geringe Schwierigk.**	**(5) Keine Schwierigk.**
In die Badewanne ein- und aussteigen	☐	☒	☐	☐	☐
Bitte kreuzen Sie an, in welchem Maße Sie momentan Schwierigkeiten haben, folgendes zu tun	**(1) Unmöglich**	**(2) Starke Schwierigk.**	**(3) Mäßige Schwierigk.**	**(4) Geringe Schwierigk.**	**(5) Keine Schwierigk.**
Duschen	☐	☐	☒	☐	☐

Abb. 4 Beispiel für den Verlauf eines adaptiven Tests im Bereich „Funktionsfähigkeit im Alltag".

Voraussetzungen des adaptiven Testens

Die meisten vorliegenden Testinstrumente sind im Rahmen der klassischen Testtheorie (KTT) entwickelt worden [9], und die meisten dieser Instrumente konzentrieren sich dabei auf einen eingeschränkten Fähigkeitsbereich, wie in **Abb. 1** dargestellt. Um vergleichbare Aussagen über die Leistungsfähigkeit einer Person im Rahmen dieser Messinstrumente machen zu können, ist es erforderlich, dass jedes Individuum *alle* Items bearbeitet, die zur Erfassung der Leistungsfähigkeit vorgesehen sind. Will man für ein Merkmal den gesamten Fähigkeitsbereich abdecken, sollte das Instrument idealerweise viele Items mit hoher Heterogenität der Itemschwierigkeiten umfassen (siehe **Abb. 5a**), um in allen Bereichen eine zuverlässige Messung zu ermöglichen. Reduziert man die Itemanzahl (siehe **Abb. 5b**), so wird die Messung insgesamt weniger zuverlässig und weniger präzise, oder sie ist nur noch geeignet, in einem eingeschränkten Bereich des Merkmalsspektrums zuverlässig zu diagnostizieren (siehe **Abb. 5c**).

Bei klassisch konstruierten Skalen werden in der Regel mithilfe des aus vielen Einzelitems errechenbaren Summenwerts (z. B. einer Skala) die Leistungen verschiedener Personen miteinander verglichen. Der Summenwert einer Skala der KTT kann dabei jedoch durch verschiedene Antwortmuster zustande kommen (s. schwarzes und graues Antwortmuster in **Abb. 6**). Um alle Informationen, die in einem Summenwert stecken, auszuschöpfen, muss das Antwortmuster mit berücksichtigt werden.

Das Antwortmuster zweier unterschiedlicher orthopädischer Patienten gibt **Abb. 6** wieder. Die Beispielpatientin A (65-jährig, große Einschränkungen im Bereich „Funktionsfähigkeit im Alltag") weist das schwarze Antwortmuster auf. Das graue Antwortmuster ist Beispielpatient B (geringe Einschränkungen im alltäglichen Leben, zusätzlich depressive Symptomatik) zuzuordnen. Bildet man jeweils den Summenwert, so ergibt sich

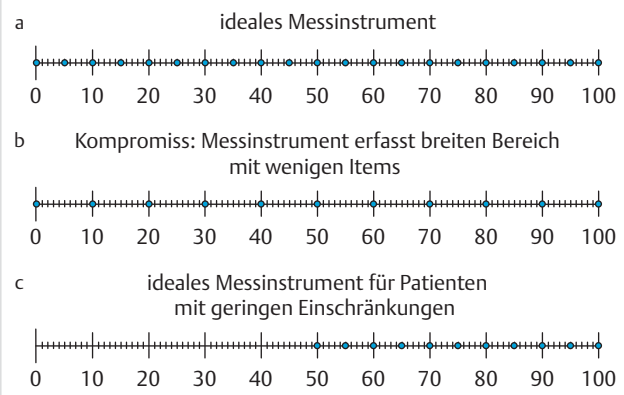

Abb. 5 Verteilung von Itemschwierigkeiten bei 3 unterschiedlichen Messinstrumenten.

Bitte kreuzen Sie an, in welchem Maße Sie momentan Schwierigkeiten haben, folgendes zu tun	(1) Unmöglich	(2) Starke Schwierigk.	(3) Mäßige Schwierigk.	(4) Geringe Schwierigk.	(5) Keine Schwierigk.
Anstrengende Sportarten betreiben, wie z. B. Joggen	☒	☒	☐	☐	☐
Einen schweren Gegenstand über 5 Meter tragen	☒	☐	☐	☒	☐
Zu einer kulturellen Veranstaltung oder einem Fest gehen	☒	☐	☐	☐	☒

Abb. 6 Auswirkungen zweier Antwortmuster auf den Summenwert [6] (schwarz: Beispielpatientin A, grau: Beispielpatient B).

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

sowohl für Patientin A als auch für Patient B ein Score von 7. Wenn man nur diesen Wert kennt, ist man geneigt zu behaupten, dass beide Patienten ähnlich stark belastet sind. Die Summenwerte beruhen aber auf vollkommen unterschiedlichen Angaben und unterschiedlichen Personenmerkmalen.

Im Rahmen der Item-Response-Theorie (IRT) gehen in den Summenwert tatsächlich alle Informationen über das Ankreuzverhalten einer Person ein. Dies wird durch verschiedene Annahmen der IRT sichergestellt [6, 10]:

- Alle Items einer Skala messen ausschließlich dieselbe Personeneigenschaft,
- allein die Personenfähigkeit beeinflusst die Antwort auf ein vorgelegtes Item,
- die Rangordnung der Schwierigkeiten aller Items ist für alle Personen gleich. Das bedeutet, dass ein einfaches Item für alle Personen einfacher ist als ein schwieriges.

Diese Annahmen sind als Voraussetzungen zu verstehen, die im Rahmen einer Konstruktion einer solchen Skala zu überprüfen sind.

Gelten die Annahmen der IRT, kann das in ▶ **Abb. 6** dargestellte diagnostische Problem der Vermischung unterschiedlicher Informationsaspekte nicht auftreten, sodass die Interpretierbarkeit von Summenwerten nicht eingeschränkt ist. Mit den Vorgaben der IRT wäre die in ▶ **Abb. 6** aufgeführte Befragung lediglich als homogener Indikator für die „Funktionsfähigkeit im Alltag" geeignet, wenn Item 1 und 2 als Elemente einer entsprechenden Skala akzeptiert würden. Für Item 3 wäre eine theoretische Klärung oder Fokussierung der Inhalte notwendig, um eine eindeutig inhaltliche Zuordnung zu dieser oder einer anderen Skala (z. B. „Depressive Symptome") inhaltlich rechtfertigen und ggf. anschließend auch empirisch als zum Modell zugehörig akzeptieren zu können. IRT-Skalen werden selbstverständlich auch für die Erfassung mehrerer Dimensionen eingesetzt: Die Diagnostik muss dann aber mit getrennten eindimensionalen Skalen erfolgen, sodass die oben aufgeführten Annahmen für jede einzelne Skala gelten.

Durch die erläuterten Eigenschaften kann ein weiterer Vorteil der IRT ausgenutzt werden, der für jedes adaptive Testen grundlegend ist. Jedes einzelne Item kann für Vergleichszwecke separat herangezogen werden. Es muss nicht die komplette Skala vorgelegt werden, um Aussagen über die Leistungsfähigkeit einer Person machen zu können. Mit allein einer Frage könnte man theoretisch schon Unterschiede zwischen den Fähigkeiten zweier Personen feststellen. Nur ist die Antwort auf eine einzige Frage mit relativ großer statistischer Unsicherheit verbunden, sodass hieraus noch kein zuverlässiger Wert für die Personenfähigkeit geschätzt werden kann. Um die Zuverlässigkeit zu erhöhen, bedarf es Antworten auf verschiedene Items, aber eben nicht auf alle Items. Es muss aber betont werden, dass eine adaptive Testung die Gültigkeit der Annahmen der IRT für die verwendete Itemgruppe oder Itembank voraussetzt.

Vorgehen beim adaptiven Testen

▼

Ziel beim adaptiven Testen ist es, mit möglichst wenigen Items eine möglichst präzise und zuverlässige Schätzung der wahren Personenfähigkeit zu erreichen.

Zuverlässigkeit der Schätzung der Personenfähigkeit

Sowohl in der KTT als auch in der IRT gibt die Reliabilität an, wie zuverlässig eine Schätzung ist. Damit können Angaben gemacht werden, wie gut eine Schätzung den tatsächlichen (aber nicht bekannten) Wert, z. B. die Personenfähigkeit, vorhersagen kann. Wie bereits erwähnt, wird die Zuverlässigkeit größer, wenn die Schätzung auf mehreren Antworten beruht.

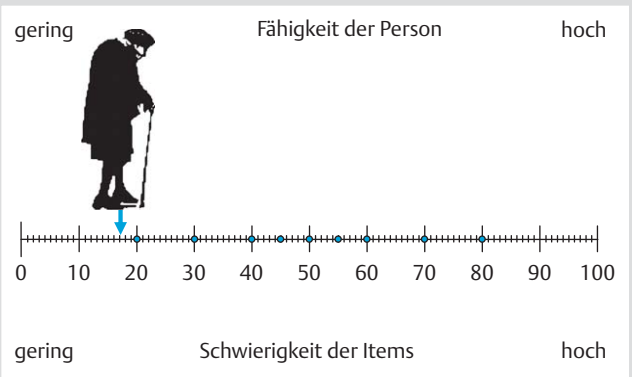

Abb. 7 Fehlende Passung zwischen Personenfähigkeit und Itemschwierigkeit.

Die Reliabilität hängt auch davon ab, welche Items genau für eine Schätzung herangezogen werden. Wird mit einem Test, der besonders die Fähigkeit im mittleren Bereich erfasst (s. ▶ **Abb. 7**), die Funktionsfähigkeit der stark eingeschränkten Beispielpatientin A geschätzt, so kann dies nur mit großer Ungenauigkeit erfolgen. In diesem Test fehlen nämlich Fragen, die für die Patientin mit ihren Einschränkungen relevant sind, die sich auf ihren Fähigkeitsbereich beziehen.

In der KTT wird der Reliabilitätskoeffizient meist für eine Gruppe von Personen angegeben, für die der Test vorgesehen ist, unabhängig davon, welche individuellen Leistungsfähigkeiten vorliegen. Betrachtet man die Reliabilität in der KTT genauer, wird ersichtlich, dass sie nicht in allen Schwierigkeitsbereichen des Tests gleich gut sein kann (vgl. ▶ **Abb. 7**). Dies hängt wesentlich von der Zusammensetzung der Items ab. ▶ **Abb. 8** stellt diesen Zusammenhang dar. Die Kurve verdeutlicht, wie bei einer statischen Skala die Information, welche die Items bzw. der Summenwert liefern, in den Extrembereichen abnimmt. Personen mit einer sehr hohen bzw. sehr geringen Fähigkeit auf dem zu messenden latenten Merkmal können nur wenig reliabel eingestuft werden, die Information der Items ist zu gering. Das in ▶ **Abb. 7** schematisch dargestellte Messinstrument würde so eine ungleichmäßige Informationskurve liefern.

Bei einem adaptiven Test wird dieser Nachteil vermieden: Die Reliabilität hängt hier von der jeweiligen Zusammenstellung von Items in einem Test ab bzw. von der Fähigkeit der Person, die diese bearbeitet. So kann für jeden Test abhängig von den beinhalteten Items und der zu schätzenden Personenfähigkeit ein separater Reliabilitätskoeffizient errechnet werden. Liegen in allen Fähigkeitsbereichen gute Items vor (s. ▶ **Abb. 5**, ideales Messinstrument), so kann für jeden Fähigkeitsbereich einer Person derselbe Reliabilitätskoeffizient garantiert werden (waagrechte Linie in ▶ **Abb. 8**).

Die bestmögliche Zuverlässigkeit eines Messinstrumentes ist dann erreichbar, wenn die Schwierigkeiten der Items exakt zu der Fähigkeit der Rehabilitanden passen [11], wenn also z. B. ein Test mit schwierigen Items einer Person mit hoher Fähigkeit vorgegeben wird.

Daraus lässt sich ableiten, dass ein Test die Schwierigkeit der vorgegebenen Items dem Antwortverhalten des Rehabilitanden anpassen muss, wenn bestmögliche Reliabilität und Ökonomie

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

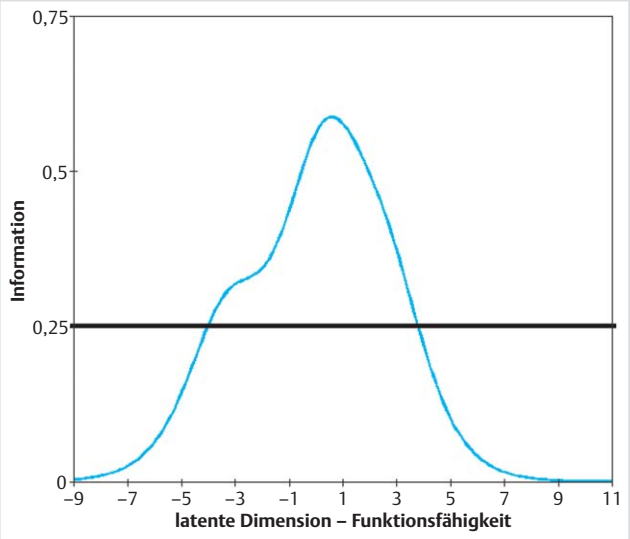

Abb. 8 Informationskurven für einen statischen (Kurve) und einen adaptiven (waagrechte Linie) Test. Je höher der Wert der Informationsfunktion ist, desto genauer und reliabler misst der Test in dem betreffenden Fähigkeitsbereich.

erreicht werden sollen. In diesem Zusammenhang sei darauf hingewiesen, dass auch andere IRT-Modelle als das Rasch-Modell dem adaptiven Testen zugrunde gelegt werden können: Diese lassen dann zu, dass sich auch die Trennschärfe oder andere Merkmale der Items zwischen den Items einer Skala unterscheiden können. Die Itementwicklung kann dann nach schwächeren Selektionskriterien erfolgen. Hierdurch schränkt sich die Interpretierbarkeit der Ergebnisse auf Einzelitemebene ggf. erheblich ein, und die Reliabilität und Ökonomie der Testung wird nicht nur durch die Passgenauigkeit der Itemschwierigkeiten determiniert. Eine differenzierte Diskussion dieser Unterschiede für verschiedene IRT-Modelle würde den Rahmen dieser einführenden Darstellung sprengen (s. hierzu [7, 8, 12]).

Adaptiver Algorithmus

Das Wichtigste beim adaptiven Testen ist es, eine gute Passung zwischen der Fähigkeit der Person und den Schwierigkeiten der Items herzustellen. In allen Schritten des adaptiven Tests liegt das Hauptaugenmerk darauf und damit verbunden auf der Frage, welche Items ausgewählt werden müssen, damit die Fähigkeit der Person am besten geschätzt werden kann.

Es muss demnach ein adaptiver Algorithmus spezifiziert werden (s. ▶ **Abb. 9**), der die Itemauswahl zu Beginn und während des Tests steuert und festlegt, unter welchen Bedingungen ein Test beendet wird.

Beginn der Testung

Sofern keine Vorab-Informationen vorliegen, sollte die erste Frage, die einem Patienten im Rahmen eines adaptiven Tests präsentiert wird, eine mittlere Schwierigkeit aufweisen. Hierfür kann man zufällig ein Item aus den Items mit mittlerem Schwierigkeitsparameter auswählen (s. auch ▶ **Abb. 3**).

Meistens gibt es aber schon vorab weitere Informationen über die Person, die den Test bearbeitet. Diese Vorab-Informationen können helfen, bereits zu Beginn des Tests ein möglichst optimales Item auszuwählen. Der Patient bekommt dann beispielsweise genau das Item vorgelegt, das der durchschnittlichen Fähigkeit einer genauer definierten Subpopulation (z. B. gemäß Alter,

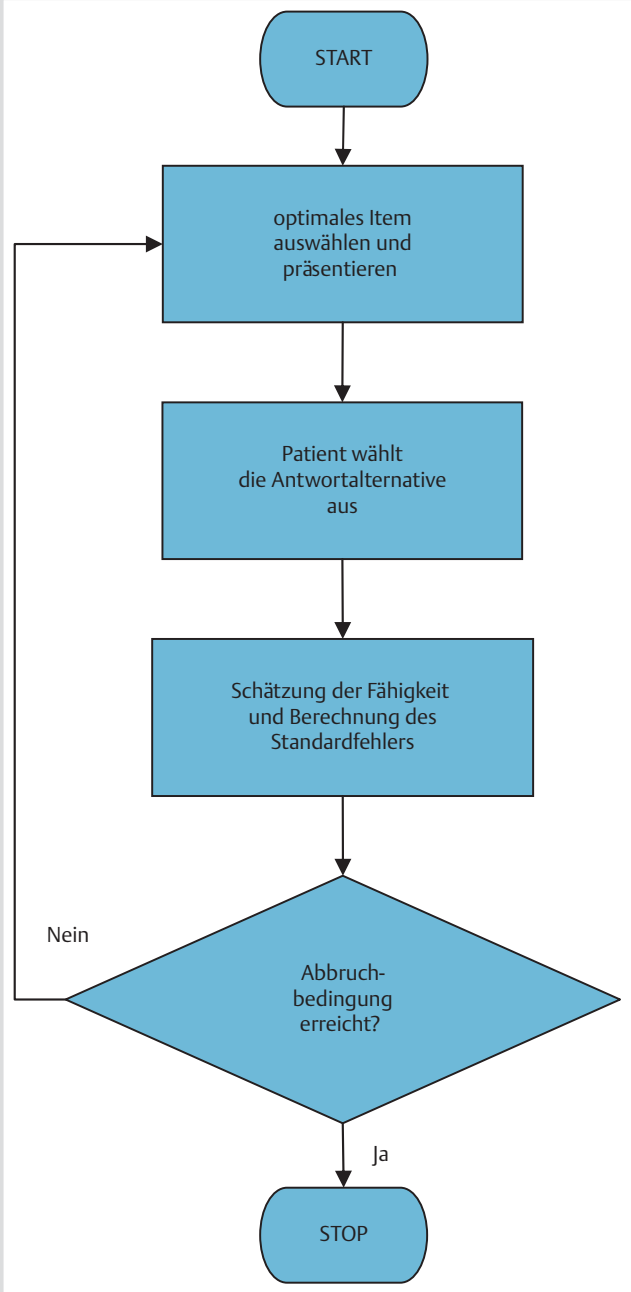

Abb. 9 Algorithmus für die Itemauswahl eines adaptiven Tests (vgl [13]).

Geschlecht, Indikationsgebiet) entspricht. Denkbar ist auch ein kurzer Vortest, durch den eine Vorab-Einstufung des Patienten in eine von mehreren vorher definierten Leistungsstufen vorgenommen wird [14]. So könnte Patient B in die Gruppe der männlichen Patienten mit einer vorausgegangenen Hüft-OP eingeordnet werden. Oder Patientin A wird ein Item mittlerer Schwierigkeit für ältere Orthopädie-Patientinnen vorgelegt.

Verlauf der Testung

Das entscheidende Kriterium für die Itemauswahl während der Testung ist die Schätzung der Fähigkeit der Person auf Grundlage der Angaben des Probanden bei schon beantworteten Items. Vereinfacht kann das dahinterliegende Rational folgendermaßen beschrieben werden:

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

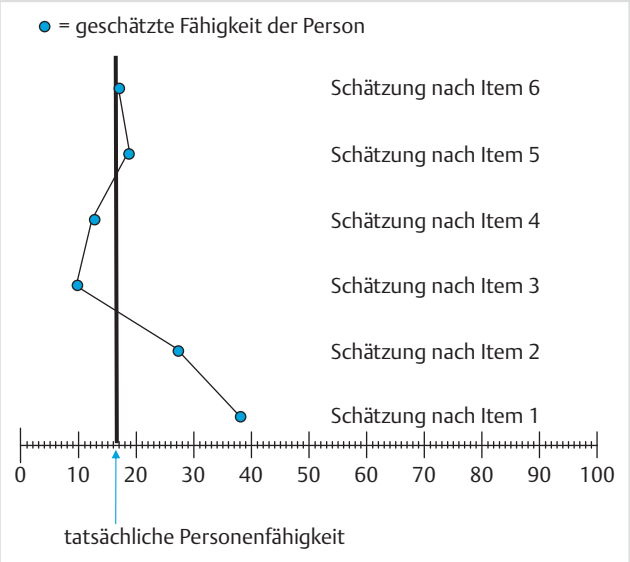

Abb. 10 Annähern der Schätzung des Personenparameters an den tatsächlichen Wert auf einer Skala von 0 (geringe Personenfähigkeit) bis 100 (hohe Fähigkeit).

- Ist für einen Patienten ein vorgelegtes Item zu schwierig (weil seine Einschränkungen zu groß sind), wird als nächstes ein Item präsentiert, das eine geringere Schwierigkeit aufweist.
- Ist für einen Patienten ein vorgelegtes Item zu einfach (weil seine Einschränkungen vergleichsweise gering sind), wird als nächstes ein Item präsentiert, das eine höhere Schwierigkeit aufweist.

Dieses Rational liegt auch der Itemvorgabe in den **Abb. 3, 4** zugrunde. Ziel ist es dabei, sich der wahren Fähigkeit der Person bzw. dem wahren Personenparameter immer mehr zu nähern (s. **Abb. 10**).

Ende der Testung

Bei Erreichen einer ausreichend genauen Schätzung kann der Test beendet werden. Die Schätzung der Personenfähigkeit ist dann „ausreichend", wenn deren Messgenauigkeit eine vorab definierte Grenze erreicht (z. B. 0,25 * Standardfehler). Wie oben dargestellt, hängt dies wiederum davon ab, wie gut die Schwierigkeit der Fragen die Fähigkeit der Person trifft. Demnach ist ein adaptiver Test umso besser und für Patienten umso kürzer, je sorgfältiger der Test entwickelt worden ist und je besser die verfügbaren Items das gesamte Fähigkeitsspektrum abdecken.

Über dieses gebräuchlichste Kriterium hinaus können auch weitere Bedingungen für einen Testabbruch definiert werden. Diese können sich z. B. an der Anzahl bearbeiteter Items oder an der Testzeit orientieren.

Insgesamt kann das Vorgehen beim adaptiven Testen folgendermaßen zusammengefasst werden:

1. Das erste Item liegt mit seiner Schwierigkeit in einem mittleren Bereich (wenn keine genaueren Informationen über die Person in die adaptive Diagnostik einbezogen werden).
2. Auf Grundlage der Antwort(en) der Person wird ein Maß für die Personenfähigkeit (bzw. Einschränkung) und deren Schätzgüte berechnet.
3. Ist ein vorgelegtes Item für einen Patienten zu schwierig, wird ein leichteres Item vorgelegt. Ist das Item zu einfach, wird ein schwereres Item vorgelegt (hierbei ist noch sicherzustellen, dass ein Patient kein Item doppelt dargeboten bekommt).[1]
4. Die letzten beiden Schritte werden so oft wiederholt, bis die Schätzung des Personenmerkmals ein definiertes Gütekriterium erreicht oder andere Kriterien für den Testabbruch erfüllt sind.

Diskussion: Vor- und Nachteile des adaptiven Testens

Vorteile

Das adaptive Testen hat den entscheidenden Vorteil der Ökonomie. Anders als bei herkömmlichen Messverfahren müssen Patienten nur einen Bruchteil der zur Verfügung stehenden Fragen bearbeiten. Dies wird in den **Abb. 3, 4** deutlich: Obwohl Fragen für den gesamten Fähigkeitsbereich vorhanden sind, werden aus diesen nur sehr wenige für die Patientin A ausgewählt. Dieser Vorteil wiegt insbesondere dann sehr stark, wenn es sich um Konstrukte handelt, die mit klassischen Messverfahren nur durch sehr viele Items erfasst werden könnten, bzw. dann, wenn ein Messinstrument den Anspruch hat, den gesamten Fähigkeitsbereich zu erfassen. Dies ist vor allem dann sinnvoll, wenn ein Instrument eingesetzt werden soll, um sowohl große Einschränkungen in einer akuten Erkrankungsphase als auch weniger starke Einschränkungen im Laufe der Behandlung zu messen.

Bei klassischen Testverfahren sind Ökonomie und Messpräzision gegenläufig: Um ein besonders realiables Messinstrument zu erhalten, müssen im Bereich der Ökonomie Abstriche gemacht werden, oder aber umgekehrt. Das adaptive Testen schafft beides. Neben der aufgeführten Ökonomie kann auch eine sehr gute Messpräzision erreicht werden. Insbesondere kann eine durchgängig gute Reliabilität in allen Schwierigkeitsbereichen der zu messenden Eigenschaft sichergestellt werden (s. **Abb. 8**).

Ein adaptives Messverfahren kann, ohne unökonomisch zu sein, den gesamten Schwierigkeitsbereich eines Merkmals abdecken. Dies hat ebenfalls den Vorteil, dass dadurch weniger Decken- bzw. Bodeneffekte auftreten im Vergleich zu einem klassischen Instrument. Dadurch ist ein adaptives Vorgehen ideal zur Veränderungsmessung: Sowohl große Einschränkungen in einer akuten Erkrankungsphase als auch mildere Einschränkungen im Laufe der Behandlung können mit demselben Messinstrument erhoben werden. Bei herkömmlichen Erhebungsmethoden sind oftmals unterschiedliche Instrumente zur Erfassung von unterschiedlichen Einschränkungsbereichen erforderlich, dadurch ist aber eine Vergleichbarkeit der Messwerte infrage gestellt. Die Vergleichbarkeit ist beim adaptiven Testen gegeben.

Das adaptive Testen kann außerdem zu einer höheren Akzeptanz der Patienten beitragen: Beim adaptiven Testen wird der Aufwand für die Patienten auf ein Minimum reduziert. Die Items, die den Patienten vorgelegt werden, sind für die Situation des Patienten relevant. Bei umfangreichen Befragungen nach der klassischen Testtheorie sind hingegen eine Vielzahl von Fragen zu beantworten, die für den Patienten entweder zu einfach oder

[1] Diese Darstellung des Vorgehens ist vereinfacht. Genauer gesagt werden aufgrund der Schätzung der Personenfähigkeit (auf Grundlage der vorhandenen Antworten) Items ausgewählt, die die maximale zusätzliche Information liefern. Dies entspricht oft dem geschilderten Vorgehen. Eine Person kann aber auch ein eigentlich zu schwieriges Item mit einer gewissen Wahrscheinlichkeit lösen; dann sollte trotzdem sichergestellt werden, dass der Algorithmus die Fähigkeit richtig (nämlich unterhalb der Schwierigkeit des gelösten Items liegend) einschätzt.

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

zu schwierig sind. Dies führt oft zu Unmut: Die Fragen werden als unangemessen wahrgenommen und die Kooperationsbereitschaft des Patienten sinkt. Genau dieser Effekt wird durch das adaptive Testen vermieden. Hinzu kommt noch die Testökonomie, die ebenfalls positive Auswirkungen auf die Akzeptanz des Patienten hat.

Nachteile

Wie bei den Vorteilen des adaptiven Testens erwähnt, gibt es hauptsächlich Fragen, die von ihrer Schwierigkeit her zu der Fähigkeit der Personen passen. Diese Passung führt zu weniger frustrierenden Erlebnissen durch zu schwierige Fragen insbesondere bei Leistungstestungen. Auf der anderen Seite kann daraus aber gelegentlich ein spezifischer Nachteil entstehen: Motivierende Erfolgserlebnisse durch relativ einfache Items finden nämlich ebenfalls weniger statt. Deshalb können adaptive Testprozeduren im Vergleich zu klassischen Testverfahren u.U. auch eine eher motivationshemmende Wirkung haben, insbesondere bei Personen, die eine Fähigkeit im oberen Bereich aufweisen.

Schon bei der Planung eines adaptiven Tests müssen die oben dargestellten Voraussetzungen der IRT beachtet werden. So können für einen Test nur Fragen aufgenommen werden, welche dieselbe Problematik beim Patienten erfassen. Sollen mehrere Konstrukte erfasst werden, müssen diese mit unterschiedlichen adaptiven Tests entwickelt und erhoben werden. Aber selbstverständlich kann die Erfassung mehrerer Konstrukte innerhalb eines Computerprogramms oder in Form verzweigter Tests umgesetzt werden.

Außerdem verlangt ein guter adaptiver Test, dass hinreichend viele Fragen vorhanden sind, die den gesamten Fähigkeitsbereich abdecken (sog. Itembank). Die Vorteile adaptiven Testens kommen erst dann zum Tragen, wenn hinreichend viele Fragen im gesamten Schwierigkeitsspektrum des zu messenden Merkmals vorhanden sind. Die Entwicklung eines adaptiven Tests erfordert deshalb einen höheren zeitlichen Aufwand, dieser besteht neben dem Konstruktionsaufwand für die Itembank auch in der empirischen Überprüfung der Modellgeltung. Daneben ist die Umsetzung des dargestellten Algorithmus in ein Computerprogramm derzeit noch häufig aufwändige Pionierarbeit.

Das Entwickeln eines adaptiven Tests bedeutet meistens eine Instrumenten-Neuentwicklung. Dadurch ist die Vergleichbarkeit zu vielen bestehenden Instrumenten der KTT nicht mehr uneingeschränkt gegeben.

Darüber hinaus bringt die Durchführung eines adaptiven Tests, wie er hier dargestellt wurde, vor Ort einige Schwierigkeiten mit sich. In den Rehabilitationskliniken müssen zumeist erst einmal die Voraussetzungen geschaffen werden. Jeder Rehabilitand sollte Zugang zu einem Computer haben, der zur Testdurchführung notwendig ist. Zumindest sollte dies zu vorher definierten Messzeitpunkten (z.B. Aufnahme und Entlassung) den Patienten ermöglicht werden.

Schließlich kann besonders bei Rehabilitanden, die bisher wenig Berührung mit moderner EDV hatten, eine gewisse Scheu vor der Fragebogenbearbeitung am Computer beobachtet werden, was als erste Hürde überwunden werden müsste. Hierbei ist jedoch zu beachten, dass bei moderner Umsetzung die Patienten nur noch Items und Ankreuzoptionen auf dem Bildschirm sehen und keine weiteren Bedienschritte außer dem Anklicken des Antwortkästchens notwendig sind (z.B. [8]). Die früher insbesondere bei der Testung von älteren Patienten geäußerten Bedenken gegen Computeranwendungen [12] erscheinen vor diesem Hintergrund als weniger begründet. Insbesondere wenn ältere Menschen bei der Bedienung assistiert werden, sollte der Vorteil der Ökonomie Probleme beim Ausfüllen von Langfragebogen durch ältere Patienten [15] überwiegen.

Aktueller Stand des adaptiven Testens in der Rehabilitation

Aktuell gibt es in der Rehabilitation nur wenige Beispiele für IRT-gestützte adaptive Testverfahren. In einigen rehabilitativen Teilbereichen wurde das adaptive Testen als Computertest umgesetzt. Beispielhaft aufgeführt seien hier aus dem deutschen Sprachraum adaptive Testmodelle zur Messung von Angst [16] und Depression [17–19] und zur Erfassung von Stresserleben [20]. International ist das Patient-Reported Outcomes Measurement Information System (PROMIS) der National Institutes of Health (NIH) zu nennen. Im Rahmen des PROMIS-Projekts [21–23] wurde ein adaptiver Test entwickelt, der zum Ziel hat, für Patienten mit chronischen Erkrankungen die Bereiche psychische, physische und soziale Gesundheit abzudecken.

In einem Kooperationsprojekt der Pädagogischen Hochschule Freiburg mit der Universität Freiburg, der Rheinisch-Westfälischen Technischen Universität Aachen und des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) in Frankfurt am Main werden aktuell mehrere adaptive Testverfahren speziell für die Rehabilitation entwickelt (Projekt RehaCAT).* Es werden adaptive Computertests für die neurologische, orthopädische und kardiologische Rehabilitation umgesetzt. Insbesondere werden hier Itembanken und adaptive Erhebungsprozeduren für die Konstrukte „Funktionsfähigkeit im Alltag“, „Funktionsfähigkeit im Beruf“, „Angst“ und „Depression“ [24] entwickelt und schließlich für den klinischen Einsatz erprobt und optimiert. Diese Bereiche orientieren sich an den Vorgaben der International Classification of Functioning, Disability and Health (ICF) [25], insbesondere an den Kategorien Aktivität, Partizipation und dem Konstrukt der Schädigungen.

Eine einfachere, nicht IRT-gestützte Möglichkeit, maßgeschneidertes Testen umzusetzen, kann durch die Implementierung fester Verzweigungen in Form von Sprungbefehlen in einem Papier-Fragebogen erfolgen. Dadurch können die Vorteile des adaptiven Testens zumindest ansatzweise genutzt werden. Aus dem Bereich der Rehabilitation ist hierfür der MOSES-Fragebogen [26,27] ein Umsetzungsbeispiel. Dieser Fragebogen orientiert sich an den Grundstrukturen der ICF und erfasst Bereiche der Aktivität und Teilhabe, in erster Linie Mobilität und Selbstversorgung. Das maßgeschneiderte Testen ist hier dadurch realisiert, dass für jeden Bereich eine Eingangsfrage gestellt wird, aus deren Beantwortung hervorgeht, ob von einem Patienten in diesem Bereich Angaben erforderlich sind oder nicht.

Ausblick

Wie bereits erwähnt, muss sich eine adaptive Teststrategie nicht auf nur ein eindimensionales Merkmal beschränken, sondern es kann auch ein sukzessives Erfassen mehrerer eindimensionaler Merkmale umgesetzt werden. In neueren Entwicklungen ist sogar auch ein multidimensionales adaptives Testen möglich. Dieses ist dann eine Alternative, wenn hoch korrelierte Konstrukte (z.B. Angst und Depression; Mobilität und Selbstversor-

* Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

gung) diagnostiziert werden und ein Item einen Beitrag zur Schätzung mehrerer Konstruktausprägungen leisten kann. Dabei wird die Zusammenhangsstruktur der gemessenen Dimensionen bereits für die Itemauswahl und die Schätzung der Merkmalsausprägung berücksichtigt [14]. Dies bedeutet, dass – anders als beim sukzessiven, mehrmaligen eindimensionalen Testen – ein Item Auswirkungen auf den Algorithmus und die Schätzung in mehreren Merkmalen hat, da ein- und dasselbe Item zu mehreren Dimensionen einen Beitrag zur Schätzung leisten kann. Diese neuere Entwicklung zeigt beispielhaft, dass die dem adaptiven Testen inhärenten Potenziale zunehmend erschlossen werden.

Adaptives Testen bringt viele Vorteile mit sich, die eine Zunahme in der Umsetzung dieser Art von Datenerhebung erhoffen lassen. Je mehr Umsetzungen es vor allem im Bereich des Computer-adaptiven Testens gibt, desto einfacher wird es, Folgeprojekte auf bereits bekannten und erprobten Algorithmen aufzubauen. Adaptives Testen bietet viele Möglichkeiten, es ist aber noch umfangreiche Entwicklungserfahrung notwendig, um einen effizienten Einsatz als Standardmethode zu gewährleisten.

Kernbotschaft

Das adaptive Testen stellt eine alternative Herangehensweise an Diagnostik und Evaluation dar, die es Klinikern erlaubt, mit geringem Aufwand eine sehr präzise Einschätzung von interessierenden Patientenmerkmalen zu erhalten. Diese moderne Art des Testens wird auch in der Rehabilitation immer mehr eingesetzt und kann durch seine spezifischen Vorteile das Spektrum an diagnostischen Methoden und Strategien in der medizinischen Rehabilitation ergänzen und bereichern.

Literatur

1 *Andresen EM*. Criteria for assessing the tools of disability outcomes research. Archives of Physical Medicine and Rehabilitation 2000; 81 (Suppl 2): S15–S20
2 *Keith RA*, *Granger CV*, *Hamilton BB*, *Sherwin FS*. The Functional Independence Measure: A new tool for rehabilitation. In: *Eisenberg MG, Grzesiak RC*, Hrsg. Advances in clinical rehabilitation. New York: Springer; 1987: 6–18
3 *Mahoney FI*, *Barthel DW*. Functional Evaluation: The Barthel Index. Maryland State Medical Journal 1965; 21: 61–65
4 *Lübke N*, *Grassl A*, *Kundy M et al*. Hamburger Einstufungsmanual zum Barthel-Index. Geriatrie Journal 2001; (1–2): 41–46
5 *Gauggel S*, *Böcker M*, *Zimmermann P et al*. Item-response-Theorie und deren Anwendung in der Neurologie. Nervenarzt 2004; 75: 1179–1186
6 *Wirtz M*, *Böcker M*. Eigenschaften und Nutzen des Rasch-Modells in der klinischen Diagnostik. Rehabilitation 2007; 46: 238–245
7 *Choi SW*, *Reise SP*, *Pilkonis PA et al*. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Quality of Life Research 2010; 19: 125–136
8 *Fries JF*, *Cella D*, *Rose M et al*. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. Journal of Rheumatology 2009; 36 (9): 2061–2066
9 *Bengel J*, *Wirtz M*, *Zwingmann C*, Hrsg. Diagnostische Verfahren in der Rehabilitation. Göttingen: Hogrefe; 2008
10 *Tesio L*. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of Rehabilitation Medicine 2003; 35: 105–115
11 *Wright BD*, *Stone M*. Best test design. Chicago: MESA Press; 1979
12 *McHorney CA*. Ten recommendations for advancing patient-centered outcomes measurement for older persons. Annals of Internal Medicine 2003; 139: 403–409
13 *Ware JE*, *Gandek B*, *Sinclair SJ*, *Bjorner JB*. Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. Rehabilitation Psychology 2005; 50 (1): 71–78
14 *Frey A*. Adaptives Testen. In: *Moosbrugger H, Keleva A*, Hrsg. Testtheorie und Fragebogenkonstruktion. Springer: Berlin; 2007: 261–278
15 *Gauggel S*, *Birkner B*. Diagnostik depressiver Störungen bei älteren Menschen: Eine Übersicht über Entwicklung und Evaluation der „Geriatric Depression Scale“ (GDS). Zeitschrift für Gerontopsychologie und -psychiatrie 1998; 11: 159–171
16 *Becker J*, *Walter OB*, *Fliege H et al*. Using the item response theory to develop a computer adaptive test for anxiety. Quality of Life Research 2002; 11: 670
17 *Fliege H*, *Becker J*, *Walter OB et al*. Development of a computer-adaptive test for depression (D-CAT). Quality of Life Research 2005; 14: 2277–2291
18 *Fliege H*, *Becker J*, *Walter OB et al*. Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. International Journal of Methods in Psychiatric Research 2009; 18: 23–36
19 *Walter OB*. Adaptive Tests for Measuring Anxiety and Depression. In: *van der Linden W, Glas CAW*, Hrsg. Elements of adaptive testing. New York: Springer; 2010: 123–136
20 *Kocalevent RD*. Stress-CAT – Entwicklung eines computeradaptiven Tests zur Erfassung von Stresserleben. Saarbrücken: VDM; 2007
21 *DeWalt DA*, *Rothrock N*, *Yount S*, *Stone AA*. Evaluation of item candidates: the PROMIS qualitative item review. Medical Care 2007; 45 (5 Suppl 1): S12–S21
22 *Reeve BB*, *Hays RD*, *Bjorner JB et al*. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care 2007; 45 (5 Suppl 1): S22–S31
23 *Hays RD*, *Liu H*, *Spitzer K*, *Cella D*. Item response theory analyses of physical functioning items in the medical outcomes study. Medical Care 2007; 45 (5 Suppl 1): S32–S38
24 *Forkmann T*, *Boecker M*, *Norra C et al*. Development of an item bank for the assessment of depression in persons with mental illnesses and physical diseases using Rasch analysis. Rehabilitation Psychology 2009; 54: 186–197
25 *World Health Organisation (WHO)*. International Classification of Functioning, Disability and Health (ICF). Geneva: WHO; 2001
26 *Farin E*, *Fleitz A*, *Follert P*. Entwicklung eines ICF-orientierten Patientenfragebogens zur Erfassung von Mobilität und Selbstversorgung. Physikalische Medizin Rehabilitationsmedizin Kurortmedizin 2006; 16: 197–211
27 *Farin E*, *Fleitz A*, *Frey C*. Psychometric properties of an International Classification of Functioning, Disability and Health (ICF)-oriented, adaptive questionnaire for the assessment of mobility, self-care and domestic life. Journal of Rehabilitation Medicine 2007; 39: 537–546

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.