Änderungssensitivität

W. Igl[1]
C. Zwingmann[2]
H. Faller[1]

Sensitivity to Change

Zusammenfassung

In der Rehabilitation werden bei evaluativen Studien häufig Patientenfragebogen eingesetzt, um die Effekte von Rehabilitationsmaßnahmen abzubilden. Dazu müssen die verwendeten Fragebogen in der Lage sein, „wahre“ Veränderungen über die Zeit wiederzugeben, d. h. sie müssen änderungssensitiv sein. Im vorliegenden Beitrag wird die Bedeutung des Konzepts „Änderungssensitivität“ für evaluative Instrumente bzw. evaluative Studien am Beispiel der Lebensqualitätsforschung verdeutlicht. Es werden sowohl qualitative Aspekte, z. B. hinsichtlich der Antwortskalierung von Assessmentinstrumenten, als auch quantitative Methoden, d. h. Studiendesigns und Kennwerte, behandelt. Darüber hinaus werden Hinweise zur Interpretation der Kennwerte gegeben.

Schlüsselwörter
Evaluation · Assessmentinstrumente · Änderungssensitivität · Responsivität · gesundheitsbezogene Lebensqualität

Abstract

In rehabilitation research patient questionnaires are widely used for evaluative purposes, i. e. to measure improvements or deteriorations over time. This is only possible if the questionnaires applied appropriately reflect „true“ change over time, i. e. they have to be sensitive to change. The aim of this paper is to point out the importance of the „sensitivity to change“ concept for evaluative assessment tools and evaluative studies, respectively, considering quality of life research as an example. Various qualitative aspects, e. g. scaling of response options of assessment tools, are covered as well as quantitative methods, i. e. study designs and indices. Furthermore, recommendations for interpretation are given.

Key words
Evaluation · assessment instruments · sensitivity to change · responsiveness · health-related quality of life

Institutsangaben
[1] Arbeitsbereich Rehabilitationswissenschaften, Institut für Psychotherapie und Medizinische Psychologie der Universität Würzburg
[2] Verband Deutscher Rentenversicherungsträger, Rehabilitationswissenschaftliche Abteilung, Frankfurt am Main

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

Korrespondenzadresse
Dipl.-Psych. Wilmar Igl · Institut für Psychotherapie und Medizinische Psychologie der Universität Würzburg · Arbeitsbereich Rehabilitationswissenschaften · Marcusstraße 9 – 11 · 97070 Würzburg · E-mail: wilmar.igl@mail.uni-wuerzburg.de

Bibliografie
Rehabilitation 2005; 44: 100 – 106 © Georg Thieme Verlag KG Stuttgart · New York
DOI 10.1055/s-2004-834719
ISSN 0034-3536

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

Einleitung

In der Rehabilitation werden im Rahmen evaluativer Studien häufig Patientenfragebogen zu mehreren Messzeitpunkten eingesetzt, um auf der Grundlage von Prä-Post-Differenzen die Effekte von Rehabilitationsmaßnahmen aus Sicht der Rehabilitanden zu erfassen. Dies kann nur gelingen, wenn die verwendeten Fragebogenskalen in der Lage sind, die aufgetretenen Veränderungen über die Zeit abzubilden, d. h. wenn es sich um änderungssensitive Instrumente handelt. Im vorliegenden Beitrag soll diese Forderung verständlich gemacht und erläutert werden, warum die klassischen Gütekriterien bzw. Kennwerte zu Objektivität (Unabhängigkeit vom Untersucher), Reliabilität (Genauigkeit) sowie Kriteriums- und Konstruktvalidität (Gültigkeit) zur Beurteilung nicht ausreichen, sondern zusätzlich das Konzept der Änderungssensitivität (sensitivity to change)[1] Berücksichtigung finden sollte. Es werden verschiedene qualitative Kriterien, z. B. hinsichtlich der Itemformulierung oder Antwortskalierung, vorgestellt, die bei der Auswahl bzw. bereits bei der Konstruktion eines evaluativen Fragebogens beachtet werden sollten. Zur quantitativen Erfassung der Änderungssensitivität werden verschiedene Methoden, d. h. Studiendesigns und Kennwerte, erläutert. Zudem werden Hinweise zur Interpretation der Kennwerte gegeben. Obwohl das Konzept der Änderungssensitivität grundsätzlich auch auf biomedizinische Messungen (z. B. Blutdruckmessung) angewendet werden kann, beschränken sich die folgenden Ausführungen auf Instrumente zur Erfassung von psychosozialen Konstrukten aus Patientensicht (z. B. subjektive Gesundheit, gesundheitsbezogene Lebensqualität). Der Beitrag zeichnet insofern die Diskussion innerhalb der Lebensqualitätsforschung nach.

Stellen wir uns zunächst folgende Situation vor: Ein Arzt möchte die Veränderung der gesundheitsbezogenen Lebensqualität seiner Patienten im Verlauf einer Rehabilitationsmaßnahme untersuchen. Er ist dabei besonders an der Sicht der Patienten interessiert und plant daher einen Patientenfragebogen einzusetzen. Nach Durchsicht und Beurteilung verschiedener Instrumente entschließt er sich, die Version 2 des IRES-Fragebogens („Indikatoren des Reha-Status") zu verwenden [1]. Der IRES-Fragebogen bildet mit seiner dreidimensionalen Struktur das Konstrukt „gesundheitsbezogene Lebensqualität" ab, welches nach allgemeiner Auffassung somatische, psychische und funktionale Aspekte umfasst [2, 3]. Auch die Angaben zur Reliabilität sowie Kriteriums- und Konstruktvalidität des Verfahrens erscheinen dem Arzt überzeugend [1, 3]. Sein besonderes Interesse gilt der IRES-Dimension „Funktionaler Status", da für ihn entscheidend ist, ob seine Patienten nach der Behandlung besser in der Lage sind, die Aufgaben ihres (Berufs-)Alltags zu bewältigen. Nach Durchführung der Studie zeigt sich jedoch, dass sich die Werte im funktionalen Bereich im Gegensatz zu den anderen Dimensionen nur geringfügig verbessert haben. Gleichwohl ist der Arzt sowohl von der Sorgfalt bei der Durchführung der Studie als auch von der Wirksamkeit der durchgeführten Behandlung und den aufgetretenen Verbesserungen gerade im funktionalen Bereich nach wie vor überzeugt. Gründe hierfür liefern ihm nicht nur eigene Verhaltensbeobachtungen, sondern auch Äußerungen der Patienten in systematischen mündlichen Befragungen.[2]

Diese Enttäuschung hätte der Arzt möglicherweise vermeiden können, wenn er sich im Vorfeld auch nach der Änderungssensitivität des gewählten Verfahrens erkundigt hätte. Er hätte dann gesehen, dass die Dimensionen der Version 2 des IRES-Fragebogens deutliche Unterschiede in ihrer Eignung, Veränderungen zu messen, aufweisen, wobei der „Funktionale Status" sich als eher wenig änderungssensitiv erweist [3].

Theoretischer Hintergrund

Bereits Mitte der 80er-Jahre stellten Kirshner u. Guyatt [4] ein Rahmenmodell zur Kategorisierung von Messinstrumenten entsprechend deren Anwendungsziel vor und bildeten damit die Grundlage für das Konzept der Änderungssensitivität [4–6]. Sie unterscheiden folgende Messziele:

1. Diskrimination: Ein diskriminativer Einsatz eines Messinstruments liegt vor, wenn die Daten die Grundlage für eine differenzielle Zuweisung bilden, etwa zu verschiedenen Behandlungsmodulen im Rahmen der rehabilitationsspezifischen Diagnostik, zur Rehabilitation überhaupt oder zu Nachsorgemaßnahmen. Darüber hinaus kann von Diskrimination als Messziel gesprochen werden, wenn im Rahmen epidemiologischer Untersuchungen Prävalenz und Inzidenz bestimmter Belastungen erhoben werden.
2. Prädiktion: Bei prädiktiver Verwendung eines Messinstruments wird auf der Grundlage der Daten eine Prognose vorgenommen, etwa hinsichtlich des Rehabilitationserfolgs (z. B. Rückkehr zur Arbeit) oder des späteren Krankheitsverlaufs.
3. Evaluation: Ein evaluativer Einsatz eines Messinstruments liegt vor, wenn Veränderungen über die Zeit abgebildet werden sollen, z. B. Veränderungen nach bzw. durch Behandlungen im Rahmen einer Rehabilitationsmaßnahme.

Obwohl die drei klassischen Gütekriterien Objektivität, Reliabilität und Validität die Grundlage für die Beurteilung der Qualität von Messinstrumenten bilden, erhalten diese Gütekriterien in Abhängigkeit vom Anwendungsziel des Instruments eine andere Wertigkeit [3–8]. Das Gütekriterium der Objektivität, welches die Durchführung, Auswertung und Interpretation eines Erhebungsinstruments betrifft, sollte in jedem Fall durch eine genaue und umfassende Manualisierung sichergestellt werden. Diese gibt u. a. vor, wie die Instruktion der befragten Person vor der Untersuchung und die Berechnung der Skalen- und Summenwerte vorgenommen werden sollte und welche inhaltliche bzw. klinische Bedeutung die Ergebnisse für die jeweilige Person haben. Des Weiteren muss man in der Lage sein, mit dem Instrument zuverlässige bzw. genaue Messungen vorzunehmen (Reliabilität). Bei diskriminativen bzw. prädiktiven Instrumenten kommt es darüber hinaus vor allem darauf an, dass plausible Zusammenhänge mit inhaltlich relevanten Außenvariablen nachgewiesen werden (Kriteriums- bzw. Konstruktvalidität, bei prädiktivem Einsatz vor allem prädiktive Kriteriumsvalidität) und dass gegebenenfalls die postulierte dimensionale Struktur des Instru-

[1] In ähnlicher Bedeutung wird häufig auch der Begriff „Responsivität" (responsiveness) verwendet.

[2] In diesem Beispiel soll das auf der Grundlage von Verhaltensbeobachtungen und des Arzt-Patient-Gesprächs gebildete ärztliche Urteil als (Pseudo-)„Goldstandard" (d. h. als ein valides Außenkriterium) belegen, dass tatsächlich Veränderungen aufgetreten sind.

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

ments belegt wird (z. B. faktorielle Validität als Aspekt der Konstruktvalidität). Im Unterschied dazu muss bei evaluativen Instrumenten der Nachweis der Fähigkeit zur Abbildung von Veränderungen über die Zeit (Änderungssensitivität) unbedingt hinzukommen. Der Nachweis von Reliabilität sowie Kriteriums- und Konstruktvalidität ist bei evaluativem Einsatz des Messinstruments zwar ebenfalls von Belang, besonders wichtig ist jedoch die Forderung nach hoher Änderungssensitivität [4]. In konzeptueller Hinsicht wird Änderungssensitivität von den meisten Autoren nicht als eigenständiges Gütekriterium aufgefasst, sondern als spezieller, nämlich longitudinaler Aspekt der Validität [7].

Im Folgenden sollen zunächst einige qualitative Aspekte, die für die Änderungssensitivität von Bedeutung sind, am Beispiel des Konstrukts „gesundheitsbezogene Lebensqualität" vor und nach einer Rehabilitationsmaßnahme erläutert werden, bevor anschließend verschiedene Methoden zur Quantifizierung der Änderungssensitivität vorgestellt werden. Zu beachten ist dabei, dass aus den qualitativen Aspekten lediglich Empfehlungen, jedoch keine allgemein gültigen, strengen Kriterien zur Auswahl bzw. Entwicklung eines änderungssensitiven Instruments abgeleitet werden können. Bei der Neukonstruktion eines solchen Instruments sollten schließlich nur diejenigen Items bzw. Skalen aufgenommen werden, die die besten quantitativen Kennwerte aufweisen [4, 5].

Qualitative Voraussetzungen für Änderungssensitivität

Als Erstes gilt es zu klären, welche Aspekte des untersuchten Konstrukts – hier der gesundheitsbezogenen Lebensqualität – erfasst werden sollen. Diese Überlegungen sind wichtig, da die Übereinstimmung zwischen denjenigen Bereichen, in denen der Patient Probleme berichtet, den Zielen der Behandlung und den gemessenen Variablen einen entscheidenden Einfluss auf die Abbildung von Veränderungen hat. Eine verbreitete Vorgehensweise ist, möglichst umfassend all jene Bereiche mit Fragen zu erfassen, die für die Lebensqualität von Patienten mit verschiedenen Krankheiten von Bedeutung sein können (krankheitsübergreifendes/generisches Instrument), sodass Veränderungen in vielen Aspekten erfasst werden. Grundsätzlich können auch generische Instrumente eine gute Änderungssensitivität aufweisen und auch aggregierte Werte (z. B. summierte Skalenwerte, Gesamtscore) aufgrund ihrer höheren Reliabilität sich änderungssensitiver als Einzelitems verhalten [9, 10]. Allerdings kann das Problem auftreten, dass bei Patienten mit einem spezifischen Krankheitsbild möglicherweise viele Fragen nicht relevant sind. Aggregierte Werte eines solchen Fragebogens können sich dann als weniger änderungssensitiv als einzelne Antworten erweisen, da sie durch die nicht zutreffenden Fragen „verwässert" werden. Für eine hohe Änderungssensitivität ist es in diesem Fall besser, nicht mit generischen, sondern mit (krankheits-)spezifischen Fragen solche Patientenstichproben zu untersuchen, die hinsichtlich der zu verändernden Aspekte als weitgehend homogen angesehen werden können. Da sich hier die spezifischen Instrumente auf die für eine bestimmte Erkrankung, Personengruppe oder Problematik bedeutsamen gesundheitlichen Aspekte konzentrieren, sind sie – natürlich beschränkt auf die jeweilige homogene Patientengruppe – häufig eher in der Lage, Veränderungen abzubilden [11 – 14]. Wenn dennoch heterogene Stichproben untersucht werden sollen, müssen hinsichtlich der Änderungssensitivität oft Kompromisse eingegangen werden. Es sollte dann möglichst versucht werden, neben generischen Instrumenten für alle Patienten auch spezifische Instrumente für definierte Subgruppen einzusetzen. Zuweilen wird auch versucht, für jeden Patienten Fragen auszuwählen, die sich spezifisch auf die individuellen Problembereiche beziehen, in denen eine Veränderung angestrebt wird (individualspezifisches Instrument) [15 – 17]. Die Auswertung und Interpretation solcher Ansätze kann jedoch mit erheblichen Schwierigkeiten verbunden sein [18]. Bei der Neukonstruktion eines änderungssensitiven Messinstruments zur gesundheitsbezogenen Lebensqualität müssen solche inhaltlichen Bereiche ausgewählt werden, die in der angestrebten Zielgruppe in engem Zusammenhang mit Veränderungen der Lebensqualität stehen und Potenzial für genügend große, möglichst klinisch bedeutsame Veränderungen aufweisen. Dies können z. B. häufige und starke Beschwerden des Patienten sein. Fragen nach Symptomen, die nur für wenige Patienten relevant sind oder bei denen zumeist nur eine geringe Belastung vorliegt, sollten nicht aufgenommen werden [4, 5]. Im Hinblick auf das Antwortformat sollte darauf geachtet werden, dass die Graduierung fein genug ist (4 – 10 Antwortstufen, visuelle Analogskala) und durch die Verbalisierungen der „Pole" ein hinreichend breiter Bereich abdeckt wird [4, 5]. Hierdurch können „Boden-" und „Deckeneffekte" minimiert werden. Diese entstehen, wenn der Bereich des Schweregrads der Erkrankung den Bereich der Antwortoptionen überschreitet und an den „Polen" des Antwortbereichs keine Differenzierung mehr möglich ist. Eine Zusammenfassung der genannten und für die Änderungssensitivität relevanten qualitativen Aspekte findet sich in Tab. **1** in den Zeilen Itemmerkmale und Skalenmerkmale.

Quantitative Methoden zur Erfassung der Änderungssensitivität

Die verschiedenen Methoden zur Erfassung der Änderungssensitivität basieren auf der Annahme, dass eine „wahre" Veränderung aufgetreten ist, die durch ein Instrument mehr oder weniger gut abgebildet wird. Streng genommen erfordert dies für die Beurteilung der Änderungssensitivität eines Instruments, die Größe der „wahren" Veränderung und die Größe der gemessenen Veränderung zu kennen und zu vergleichen. Um jedoch die Größe der „wahren" Veränderung erfassen zu können, sind valide Außenkriterien, sog. „Goldstandards", notwendig, welche aber in der Forschungspraxis zumeist nicht zur Verfügung stehen, wie später noch genauer erläutert wird. Das strenge Kriterium der „absoluten" Änderungssensitivität ist allerdings nicht notwendig, wenn es um die Wahl zwischen verschiedenen Instrumenten geht. In diesem Fall kann auch der parallele Einsatz und Vergleich verschiedener Messinstrumente anhand der „relativen" Änderungssensitivität die gewünschten Informationen liefern. Bei diesem Vorgehen setzt man lediglich voraus, dass eine Veränderung aufgetreten ist, deren genaue Höhe jedoch nicht von Bedeutung ist. Soll schließlich nur die Entscheidung darüber getroffen werden, ob ein Instrument grundsätzlich für die Erfassung von Veränderungen geeignet ist, kann bereits die „zufallskritische" Beurteilung der Änderungssensitivität eines einzelnen Instruments ausreichen, indem man die gemessenen Veränderungen zu den Zufallsschwankungen des Instruments

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

ins Verhältnis setzt. Auf Vergleiche mit anderen Messinstrumenten muss jedoch dann – aufgrund von unterschiedlichen Rahmenbedingungen anderer Untersuchungen – zumeist verzichtet werden. Vor diesem Hintergrund ist auch die Aussagekraft der im Folgenden erläuterten Methoden zu betrachten.

Welche Methoden stehen nun zur Verfügung, um die Änderungssensitivität von Messinstrumenten zu bestimmen? Dies hängt von den möglichen Studiendesigns und der Verfügbarkeit von (änderungs-)validen Außenkriterien bzw. alternativen Messinstrumenten ab. Es werden Ein- und Zwei-(bzw. Mehr-) Gruppen-Designs unterschieden, wobei Zwei-(bzw. Mehr-) Gruppen-Designs grundsätzlich über eine höhere Aussagekraft verfügen, aber auch deutlich aufwändiger sind [7,8].

Unter den Ein-Gruppen-Designs (vgl. Abb. **1**) ist zunächst das einfache Vorher-Nachher-Design zu nennen, bei dem zu Beginn und am Ende einer Intervention Messwerte erhoben werden.

Abb. **1** Messzeitpunkte beim einfachen Ein-Gruppen-Design.

Zwar ist dieses Design einfach anzuwenden und wird daher wohl auch am häufigsten eingesetzt, die Interpretation der Ergebnisse gestaltet sich jedoch schwierig. So kann z. B. nicht eindeutig entschieden werden, ob eine fehlende Veränderung in einem Fragebogenwert auf eine mangelnde Wirksamkeit der Intervention (sofern diese nicht schon zweifelsfrei belegt ist) oder eine zu geringe Änderungssensitivität des Instruments zurückzuführen ist. Es lassen sich außerdem keine Aussagen zur Zufallsvariabilität des untersuchten Merkmals unter stabilen Bedingungen treffen.

Durch die zusätzliche Durchführung einer Baselinephase vor der Intervention (vgl. Abb. **2**) kann diese Variabilität erfasst und mit den Veränderungen während der Intervention verglichen werden. Dieses erweiterte Design ist aufgrund der drei Messzeitpunkte natürlich aufwändiger. Schwierigkeiten können bei der Interpretation entstehen, wenn die Baselinephase kürzer als die Interventionsphase ist; dann besteht die Gefahr, dass die Zufallsvariabilität unter stabilen Bedingungen unterschätzt wird. Problematisch ist auch eine sehr lange Baselinephase, denn dann treten möglicherweise „wahre" Veränderungen des untersuchten Merkmals auf.

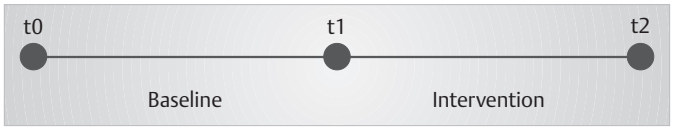

Abb. **2** Messzeitpunkte beim Ein-Gruppen-Design mit Baselinephase.

Eine andere Erweiterungsmöglichkeit des einfachen Vorher-Nachher-Designs besteht darin, ein (änderungs-)valides Außenkriterium („Goldstandard") für das gemessene Konstrukt mit zu erheben (s. o.). Anhand dieses Außenkriteriums können Korrelationen mit den Veränderungswerten des untersuchten Instruments berechnet werden. Des Weiteren können die Patienten in Gruppen mit oder ohne bedeutsame Veränderung eingeteilt werden, um Kennziffern aus der Signal-Entdeckungs-Theorie, wie Sensitivität oder Spezifität, zu ermitteln [19–21]. Die Sensitivität stellt hier einen Kennwert des untersuchten Messinstruments dar, der darüber Auskunft gibt, mit welcher Wahrscheinlichkeit Patienten mit tatsächlicher Veränderung auch als verändert klassifiziert werden. Die Spezifität dagegen ist die Wahrscheinlichkeit, mit der Patienten ohne tatsächliche Veränderung auch als unverändert eingeordnet werden. In der Realität gestaltet es sich jedoch häufig problematisch, ein geeignetes Außenkriterium zu finden. Als „Pseudo-Goldstandard" werden u. a. Globalurteile der erlebten Veränderungen von den Patienten oder von den behandelnden Ärzten erhoben [22]. Derartige Maße können jedoch durch Beurteilungsfehler und Erinnerungseffekte verzerrt sein. Insbesondere wenn durch eine sog. „direkte Veränderungsmessung" (Einpunktmessung: „Hat sich Ihre Lebensqualität seit dem Beginn der Behandlung verändert?") am Ende einer Behandlung das Ausmaß der Veränderung beurteilt werden soll, kann häufig der Ausgangszustand vom Beurteiler nicht mehr genau erinnert werden [23].

Zu den Zwei-Gruppen-Designs (vgl. Abb. **3**) zählen Interventionsstudien mit randomisierter Zuweisung der Patienten in eine Experimental- bzw. Kontrollgruppe, im Angloamerikanischen auch als „randomized controlled trials" (RCTs) bezeichnet.

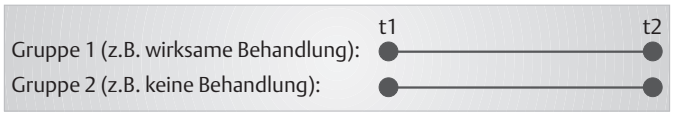

Abb. **3** Messzeitpunkte bei Zwei-Gruppen-Designs.

Im Unterschied zu üblichen RCTs ist hier nicht das Ziel der Untersuchung, die Wirksamkeit einer ungeprüften, neuen Behandlung gegenüber einer bereits geprüften Standardbehandlung zu ermitteln. Vielmehr wird hier als Intervention eine Behandlung mit in früheren Studien nachgewiesener Wirksamkeit, also eine wirksame Standardbehandlung, eingesetzt, um etwaige fehlende Veränderungen in den Messwerten der Interventionsgruppe auf das Instrument, nicht jedoch auf die fehlende Wirksamkeit der Behandlung zurückführen zu können. In der Kontrollgruppe wird dagegen keine (wirksame) Intervention eingesetzt, da hier die Erhebung der Zufallsvariabilität unter stabilen Bedingungen erfolgen soll. Dies spricht u. U. sogar gegen den Einsatz eines Plazebos in der Kontrollgruppe, da auch dadurch „wahre" Veränderungen verursacht werden können. Die Kontrollgruppe wird des Weiteren zeitlich parallel durchgeführt, wodurch sich Zeiteinflüsse verschiedenster Art gleichermaßen auswirken und damit nicht für Mittelwertdifferenzen zwischen den Gruppen verantwortlich sein können.

Die so erhaltenen Daten der Prä-Post-Messungen und gegebenenfalls einer Vorerhebung ermöglichen nun die Berechnung verschiedener Arten von Kennwerten zur Beurteilung der Änderungssensitivität [7,8]. Im Folgenden werden verschiedene

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

Kennwerte zur Bestimmung von Änderungssensitivität vorgestellt. Am häufigsten werden zurzeit verschiedene Varianten von Effektstärken/Effektgrößen[3] verwendet, welche daher ausführlicher dargestellt werden [7, 8, 24 – 28].

Den hier verwendeten Effektstärkevarianten (ES) liegt das Prinzip zu Grunde, dass eine Differenz zwischen zwei Mittelwerten M_{t1} und M_{t2} an einem Streuungsmaß SD standardisiert wird (vgl. Formel 1); die einzelnen Effektstärkemaße unterscheiden sich dabei nur in der verwendeten Standardabweichung SD im Nenner [28]:

$$ES = (M_{t2} - M_{t1})/SD \qquad \text{(Formel 1)}$$

Die Standardized Effect Size (SES, in angloamerikanischen Publikationen oft nur als „the effect size" bezeichnet) verwendet als Streuungsmaß die Standardabweichung SD_{t1} der Messwerte zum Prä-Messzeitpunkt. Dies hat zur Folge, dass bei Stichproben, deren Messwerte zum Prä-Messzeitpunkt homogen sind, größere Effektstärken resultieren. Die SES wird als „klassische" Variante der Effektstärken betrachtet und auch am häufigsten angewendet [7].

Der Standardized Response Mean (SRM) hingegen setzt die Mittelwertdifferenz $M_{t2} - M_{t1}$ in Bezug zur Standardabweichung der Messwertdifferenzen (SD_{t2-t1}), weshalb der SRM von der Homogenität der Veränderungswerte abhängig ist. So ergeben sich für den SRM – vereinfacht gesagt – höhere Effektstärken, wenn sich die Messwerte der Patienten in ähnlicher Weise zwischen zwei Messzeitpunkten verändern.[4]

Der Guyatt's Responsiveness Index (GRI) unterscheidet sich von den beiden vorherigen Kennwerten dadurch, dass er die gemessene Veränderung ins Verhältnis zu zufälligen Schwankungen unter stabilen Bedingungen setzt. Der GRI kann dabei sowohl auf den Ein-Gruppen-Fall als auch auf den Zwei-Gruppen-Fall angewendet werden. Im Ein-Gruppen-Fall wird die Mittelwertdifferenz $M_{t2} - M_{t1}$ der Interventionsphase durch die Standardabweichung SD_{t1-t0} der Differenzwerte der Baselinephase geteilt (vgl. Abb. **2**), um den GRI_{1Grp} zu berechnen. Im Zwei-Gruppen-Fall wird der GRI_{2Grp} aus der Mittelwertdifferenz $M_{t2} - M_{t1}$ der Interventionsgruppe und der Standardabweichung SD_{t2-t1} der Differenzwerte der Kontrollgruppe gebildet, die hier zeitlich parallel durchgeführt wird (vgl. Abb. **3**). Diese Koeffizienten werden umso größer, je geringer die Streuung der Messwertdifferenzen in der Baselinephase bzw. in der Kontrollgruppe ist.[5]

Die beschriebenen Effektstärken werden häufig zur Quantifizierung der Änderungssensitivität von Messinstrumenten verwendet, da sie relativ einfach zu berechnen sind und in Anlehnung an die Effektstärken nach Cohen [26] abgeleitet wurden, was eine einfache Interpretation suggeriert. Es haben sich jedoch noch keine Richtgrößen oder Mindestwerte zur Beurteilung von Effektstärken der Änderungssensitivität etabliert. Die Faustregel von Cohen, wonach ein Wert von 0,2 einem „kleinen", 0,5 einem „mittleren" und 0,8 einem „großen" Effekt entspricht, kann nicht ohne weiteres zur Beurteilung der Höhe der Änderungssensitivität herangezogen werden, da diese Einteilung ursprünglich zur Bewertung der Ergebnisse von Wirksamkeitsstudien gedacht war [7, 26]. In vielen Fällen werden die verschiedenen Effektstärkevarianten zu denselben Schlussfolgerungen führen [10]. Wie bereits erläutert, können sie jedoch auch, abhängig vom verwendeten Streuungsmaß, unterschiedliche Aussagen zur Änderungssensitivität liefern, was die Interpretation gegebenenfalls zusätzlich erschwert. Aus diesem Grund empfiehlt es sich stets, verschiedene Kennziffern zu berechnen und diese vergleichend zu betrachten. Weitere Faktoren, die nicht mit dem Instrument zusammenhängen, z. B. Wirksamkeit der Intervention, Studiendesign, Art der Patientenstichprobe und Stichprobengröße, beeinflussen diese Kennwerte. Daraus ergibt sich auch, dass Ergebnisse unterschiedlicher Studien oft nicht vergleichbar sind. Die Stichprobenspezifität von Effektstärken kann außerdem dazu führen, dass sich diese deutlich zwischen verschiedenen Stichproben unterscheiden [24]. Untersuchungen zur Prüfung der Änderungssensitivität eines Fragebogens sollten daher möglichst an solchen Stichproben durchgeführt werden, die repräsentativ sind für jene Population, in der das Instrument später eingesetzt werden soll.

Nachdem sowohl Gemeinsamkeiten als auch Unterschiede der Effektstärkevarianten beschrieben wurden, stellt sich die Frage, welcher Index nun am besten geeignet ist zur Beurteilung der Änderungssensitivität von evaluativen Messinstrumenten. Diese Frage wird in der Literatur kontrovers diskutiert. Manche Autoren bevorzugen die SES gegenüber dem SRM, welcher die Standardabweichung von Differenzwerten verwendet. Sie versuchen dadurch das Konzept der Effektstärken gegenüber denjenigen inferenzstatistischen Methoden stärker abzugrenzen, die auf der Verteilung von Differenzwerten beruhen (z. B. t-Test für abhängige Stichproben) [27]. In anderen Veröffentlichungen wird jedoch der SRM aus diesem Grunde als „best index of responsiveness" bezeichnet, da er explizit die Verteilung der Differenzwerte, d. h. der Veränderungen, mit einbezieht [25]. Unter dem Gesichtspunkt, dass der GRI die Verteilung der Differenzwerte darüber hinaus unter stabilen Bedingungen berücksichtigt, wird auch dieser Index zumindest bei Ein-Gruppen-Designs empfohlen [7].

Alternative Ansätze

In der Literatur werden gelegentlich auch inferenzstatistische Kennwerte diskutiert (z. B. t-Tests) [29]. Wegen ihrer ausgeprägten Abhängigkeit von der Stichprobengröße sollten sie jedoch nicht als alleinige Kennziffern zur Beurteilung der Änderungssensitivität benutzt werden. Auch die sog. „smallest real difference" (SRD) wurde zur Beurteilung der Änderungssensitivität vorgeschlagen [30]. Diese entspricht konzeptionell dem Reliable Change Index (RCI), der von Jacobson u. Truax [31] als ein Teilkriterium zur Beurteilung von klinisch bedeutsamen Veränderungen eingeführt wurde. Der RCI setzt die individuelle Messwert-

[3] Die korrekte Übersetzung des englischen Begriffs „effect size" lautet „Effektgröße". Häufig wird jedoch im Deutschen auch der Ausdruck „Effektstärke" verwendet. Da sich letzterer wie im Englischen mit „ES" abkürzen lässt, wird hier der Ausdruck „Effektstärke" verwendet.

[4] Die Standardabweichung der Messwertdifferenzen wird indirekt auch von der Homogenität der Stichprobe zum Prä- und Post-Messzeitpunkt und der Korrelation zwischen den Messzeitpunkten beeinflusst, was jedoch hier nicht weiter erörtert werden soll.

[5] Siehe Fußnote [4].

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

differenz $X_{t2} - X_{t1}$ ins Verhältnis zum Standardfehler der Differenzwerte SE_{diff}, wobei RCI-Werte > 1,96 als überzufällig (bei $p < 0,05$) interpretiert werden. In ihrer ursprünglichen Fassung ist die SRD zur Erfassung von Änderungssensitivität jedoch ungeeignet, da sie durch Transformation der Skalenwerte beeinflussbar und somit letztlich nur ein Maß für die Mess(un-)genauigkeit ist [24]. Allerdings ist es möglich, den Anteil von Patienten mit reliabler Veränderung, d. h. diejenigen, die eine größere Veränderung als ein Schwellenkriterium C (z. B. C = 1,96) aufweisen, als Maß für die Änderungssensitivität anzusehen. Ein Vorteil besteht darin, dass dadurch – im Gegensatz zu anderen Kennwerten – auch die unterschiedlichen Reliabilitäten der verschiedenen Instrumente Berücksichtigung finden [24].

Im Zwei-Gruppen-Fall können auch die von Norman [32] vorgeschlagenen Koeffizienten, Norman's S_{ANOVA} und Norman's S_{ANCOVA}, berechnet werden. Zur Berechnung dieser Koeffizienten wird ein Quotient aus der Varianz, die durch die Veränderung hervorgerufen wurde, und der Summe dieser Varianz und der Fehlervarianz gebildet. Diese Koeffizienten basieren auf der Grundidee der Varianzanalyse (ANOVA) bzw. der Kovarianzanalyse (ANCOVA) mit Messwiederholung. Bei der Kovarianzanalyse können auch Messwerte der Ausgangslage als Kovariaten berücksichtigt werden. Bisher sind allerdings solche Designs bei Änderungssensitivitätsanalysen noch wenig verbreitet [7].

Hingewiesen sei schließlich auch auf Definitionen von Änderungssensitivität, bei denen ein Mindestkriterium für eine „minimal clinically important difference" (MCID) einbezogen wird [33, 34]. Änderungssensitivität liegt nach diesen Ansätzen nur dann vor, wenn die Veränderungen eine als klinisch relevant bestimmte Größenordnung erreichen.

Tab. **1** Übersicht über Faktoren, die die Änderungssensitivität eines Messinstruments und ihre Beurteilung beeinflussen können

Faktor		*Erläuterung*
Fragestellung		Beurteilung der „absoluten", „relativen" oder „zufallskritischen" Änderungssensitivität (s. Text)
Instrument	Item-merkmale	Frageformulierung (z. B. Beurteilungszeitraum, Häufigkeit der gefragten Ereignisse)
		Antwortformat (z. B. Abstufung, Pole)
		Verteilung der Antworten (z. B. Boden- und Deckeneffekte)
	Skalen-merkmale	Verteilung der Skalenwerte (z. B. Boden- und Deckeneffekte)
	Gütekriterien	Objektivität, Reliabilität, Validität
Untersuchungs-methodik	Stichprobe	Homogenität der Messwerte zu einem Zeitpunkt bzw. der Messwertdifferenzen zwischen zwei Zeitpunkten
	Design	eine Gruppe vs. mehrere Gruppen (mit stabiler Kontrollbedingung)
	Intervention	Wirksamkeit der Behandlung
	Mess-instrumente	Untersuchung eines Instruments vs. mehrerer Instrumente
		Erhebung eines (änderungs-)validen Außenkriteriums
		Übereinstimmung zwischen Störung, Behandlung und gemessenen (Erfolgs-)Kriterien
	Mess-zeitraum	Angemessenheit bezogen auf den beobachteten Veränderungsprozess
	Kennwerte	Art des Kennwertes (SES, SRM, GRI u. a.)

Anm.: Effektstärken der Änderungssensitivität: SES = Standardized Effect Size, SRM = Standardized Response Mean, GRI = Guyatt's Responsiveness Index

Zusammenfassende Schlussfolgerungen

Zusammenfassend soll festgehalten werden, dass zur Beurteilung der Qualität von Messinstrumenten für die Evaluation von Veränderungen das Konzept der Änderungssensitivität unbedingt berücksichtigt werden sollte. Die bisherige Forschung bietet verschiedene Vorschläge zur Konstruktion von änderungssensitiven Instrumenten und Methoden zur Erfassung der Änderungssensitivität an, ohne jedoch verbindliche Vorgaben liefern zu können, welches Design bzw. welcher Kennwert am besten geeignet ist. Tab. **1** gibt eine Übersicht über die verschiedenen Faktoren, die die Beurteilung der Änderungssensitivität eines Messinstruments beeinflussen können. Auf dem aktuellen Stand der Forschung wird daher vorgeschlagen, die genannten qualitativen Empfehlungen zur Konstruktion von evaluativen Instrumenten zu berücksichtigen. Bei der Ermittelung bzw. Beurteilung von quantitativen Kennwerten der Änderungssensitivität sollten des Weiteren aussagekräftige Studiendesigns gewählt und stets mehrere Kennwerte (z. B. verschiedene Effektstärken) berichtet bzw. herangezogen werden, um möglichst sichere Schlussfolgerungen zu gewährleisten.

Literatur

1 Gerdes N, Jäckel WH. Der IRES-Fragebogen für Klinik und Forschung. Rehabilitation 1995; 34: XIII – XXIV
2 Testa MA, Simonson DC. Assessment of quality-of-life outcomes. New England Journal of Medicine 1996; 334: 835 – 840
3 Zwingmann C. Der IRES-Patientenfragebogen. Psychometrische Reanalysen an einem rehabilitationsspezifischen Assessmentinstrument. Regensburg: Roderer, 2002
4 Kirshner B, Guyatt G. A methodological framework for assessing health indices. Journal of Chronic Disease 1985; 38: 27 – 36
5 Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: What are the necessary measurement properties? Journal of Clinical Epidemiology 1992; 45: 1341 – 1345
6 Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. Journal of Clinical Epidemiology 2001; 54: 1204 – 1217
7 Schuck P. Designs und Kennziffern zur Ermittlung der Änderungssensitivität von Fragebogen in der gesundheitsbezogenen Lebensqualitätsforschung. Zeitschrift für Medizinische Psychologie 2000; 9: 125 – 130
8 Stratford PW, Binkley JM, Riddle DL. Health status measures: Strategies and analytic methods for assessing change scores. Physical Therapy 1996; 76: 1109 – 1123
9 Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C. Assessing the clinical significance of single items relative to summated scores. Mayo Clinic Proceedings 2002; 77: 479 – 487
10 Igl W, Zwingmann C, Faller H. Erste Ergebnisse zur Änderungssensitivität des IRES-3-Fragebogens. In: Leonhart R, Gerdes N (Hrsg): Der IRES-Fragebogen in Theorie und Praxis. Regensburg: Roderer, im Druck
11 Sangha O, Stucki G. Patientenzentrierte Evaluation der Krankheitsauswirkungen bei muskuloskeletalen Erkrankungen: Übersicht über die

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

wichtigsten Outcome-Instrumente. Zeitschrift für Rheumatologie 1997; 56: 322 – 333
[12] Sampath-Kumar D, Bitzer EM, Dörning H. Das Nottingham Health Profile (NHP) und der MacNew Heart Disease Health-related Quality of Life Questionnaire in der kardiologischen Rehabilitation. Gesundheitswesen 2004; 66: 591
[13] Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Annals of Internal Medicine 1993; 118: 622 – 629
[14] Jackowski D, Guyatt GH. A guide to health measurement. Clinical Orthopaedics and Related Research 2003: 80 – 93
[15] Kiresuk TJ, Smith A, Cardillo JE (Hrsg). Goal attainment scaling: Applications, theory, and measurement. Hillsdale, NJ: Erlbaum, 1994
[16] Malec JF. Goal attainment scaling in rehabilitation. Neuropsychological Rehabilitation 1999; 9: 253 – 275
[17] Gerdes N. Rehabilitationseffekte bei „zielorientierter Ergebnismessung“. Ergebnisse der IRES-ZOE-Studie 1996/97. Deutsche Rentenversicherung 1998; (3 – 4): 217 – 238
[18] Zwingmann C. Zielorientierte Ergebnismessung (ZOE) mit dem IRES-Patientenfragebogen: Eine kritische Zwischenbilanz. Rehabilitation 2003; 42: 226 – 235
[19] Kraemer HC. Evaluating medical tests. Newbury Park: Sage, 1992
[20] Bland M. An introduction to medical statistics. 3. Aufl. Oxford: University Press, 2000
[21] Faller H. Sensitivität, Spezifität, positiver und negativer Vorhersagewert. Rehabilitation 2005; 44: 44 – 49
[22] Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. Journal of Clinical Epidemiology 2002; 55: 900 – 908
[23] Kohlmann T, Raspe H. Zur Messung patientennaher Erfolgskriterien in der medizinischen Rehabilitation: Wie gut stimmen „indirekte“ und „direkte“ Methoden der Veränderungsmessung überein? Rehabilitation 1998; 37, Suppl 1: S30 – S37
[24] Schuck P, Zwingmann C. The „smallest real difference“ as a measure of sensitivity to change: A critical analysis. International Journal of Rehabilitation Research 2003; 26: 85 – 91
[25] Hevey D, McGee HM. The effect size statistic: Useful in health outcomes research? Journal of Health Psychology 1998; 3: 163 – 170
[26] Cohen J. Statistical power analysis for the behavioral sciences. 2. Aufl. Hillsdale, NJ: Erlbaum, 1988
[27] Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Medical Care 1989; 27: 178 – 189
[28] Leonhart R. Effektgrößenberechnung bei Interventionsstudien. Rehabilitation 2004; 43: 241 – 246
[29] Martin DP, Engelberger R, Agel J, Swiontkowski MF. Comparison of the Musculoskeletal Function Assessment questionnaire with the Short Form-36, the Western Ontario and McMaster Universities Osteoarthritis Index, and the Sickness Impact Profile health status measures. Journal of Bone and Joint Surgery 1997; 79: 1323 – 1335
[30] Pfennings LE, Ploeg HM von der, Cohen L, Polman CH. A comparison of responsiveness indices in multiple sclerosis patients. Quality of Life Research 1999; 8: 481 – 489
[31] Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology 1991; 59: 12 – 19
[32] Norman GR. Issues in the use of change scores in randomized trials. Journal of Clinical Epidemiology 1989; 42: 1097 – 1105
[33] Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: Is there a connection? Journal of Biopharmaceutical Statistics 2004; 14: 97 – 110
[34] Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): A literature review and directions for future research. Current Opinion in Rheumatology 2002; 14: 109 – 114

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