Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen

M. Wirtz

Determining the Quality of Rater Judgements Using Intraclass Correlation, and Enhancing Rater Judgements

Zusammenfassung

Einschätzungen durch Ärzte oder Therapeuten zählen zu den wichtigsten Messmethoden in der klinischen Praxis. Es wird gezeigt, wie die Zuverlässigkeit von Beurteilungen mittels Ratingskalen durch statistische Maßzahlen bestimmt werden sollte. Zudem wird verdeutlicht, welche Ursachen mangelnde Zuverlässigkeit von Beurteilungen haben kann. Das Wissen über diese Ursachen kann die Basis für Beurteilertrainings sein, die zur Sicherstellung der Qualität klinischer Einschätzungen genutzt werden können.

Schlüsselwörter

Beurteilerreliabilität · Ratingskalen · Intraklassenkorrelation · Beurteilungsfehler · Beurteilertraining

Abstract

In clinical practice ratings by physicians and therapists are among the most frequently used assessment procedures. It is shown, which statistical measures should be used to assess the reliability of such ratings. Additionally, potential causes of insufficient reliability are presented. Improvement of rating quality may be achieved by rater training, which is based on an analysis of rating errors.

Key words

Rater reliability · rating scales · intraclass correlations · rating mistakes · rater training

In der klinischen Praxis stellen Beurteilungen oder Einschätzungen durch unabhängige Dritte eine der wichtigsten Erhebungsmethoden dar. Soll beispielsweise ein behandelnder Arzt die Erkrankung eines Patienten diagnostizieren oder die Schwere einer Erkrankung einschätzen, so ist dies als typische Beurteilungssituation aufzufassen: Dem Patienten wird eine Erstdiagnose oder aber eine Zahl zugeordnet, die ökonomisch und zuverlässig den Zustand des Patienten widerspiegeln soll. Beurteilungen dieser Art können als Messungen von Eigenschaften aufgefasst werden [13].[1] In diesem Artikel soll gezeigt werden, welche Metho-

[1] Hierfür muss zusätzlich gewährleistet sein, dass jedem Patienten bei jeder Beurteilung genau *ein* Wert zugeordnet wird: Ein Patient kann nicht

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

Institutsangaben
Methodenzentrum des Rehabilitationswissenschaftlichen Forschungsverbundes Freiburg/Bad Säckingen, Freiburg

Korrespondenzadresse
Dr. phil. Dipl.-Psych. Markus Wirtz · Abteilung für Rehabilitationspsychologie · Institut für Psychologie · Universität Freiburg · Engelbergerstraße 41 · 79085 Freiburg · E-mail: wirtz@psychologie.uni-freiburg.de

Bibliografie
Rehabilitation 2004; 43: 384 – 389 © Georg Thieme Verlag KG Stuttgart · New York
DOI 10.1055/s-2003-814935
ISSN 0034-3536

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

den angewendet werden sollten, um die Güte dieses „Messvorgangs“ zu bestimmen, und welche Probleme für die praktische Anwendung bestehen, wenn diese sich als nicht ausreichend erweist.

Weshalb ist es überhaupt notwendig, sich mit der Güte solcher Beurteilungen auseinander zu setzen? Messungen, bei denen subjektive Eindrücke oder nicht direkt beobachtbare Eigenschaften durch Einschätzungen quantifiziert werden, sind leider nicht fehlerfrei: Dies wird insbesondere dadurch deutlich, dass verschiedene Beurteiler bei der Einschätzung derselben Person nicht immer zu identischen Ergebnissen kommen. Somit spiegeln die Messwerte nicht eindeutig die Eigenschaft der eingeschätzten Person wider. Berechnet man dann beispielsweise die Korrelation des eingeschätzten Merkmals mit einem weiteren Merkmal, muss davon ausgegangen werden, dass der Zusammenhang umso deutlicher unterschätzt wird, je fehlerhafter oder ungenauer das Merkmal selbst eingeschätzt wurde [8].

Deswegen wurden Kriterien definiert und entwickelt, die für jede Messung erfüllt sein sollten, damit das Ausmaß von Fehlereinflüssen auf das Messergebnis vernachlässigbar ist [8]: Messungen müssen tatsächlich das interessierende Merkmal erfassen (Kriterium der Validität), und das Messergebnis sollte die betreffende Merkmalsausprägung möglichst genau abbilden, sodass Personen mit unterschiedlicher Merkmalsausprägung zuverlässig voneinander unterschieden werden können (Kriterium der Reliabilität). Sind diese beiden Gütekriterien erfüllt, so sind die Messergebnisse und die „wahren“ Merkmalsausprägungen hinreichend deckungsgleich.

Eine zufrieden stellende Reliabilität ist die Grundlage für valide Daten:[2] Zunächst einmal muss sichergestellt werden, dass verschiedene Beurteiler identische oder ähnliche Urteile für dieselbe Person abgeben. Nur wenn diese Bedingung erfüllt ist, kann man davon ausgehen, dass Beurteiler prinzipiell *austauschbar* sind und die Beurteilung der Merkmalsausprägung nicht wesentlich von der individuellen Perspektive des jeweiligen Beurteilers bestimmt wird. Um zu identifizieren, *welches Merkmal* von den Beurteilern aber tatsächlich beurteilt wird, muss geklärt werden, welche Eigenschaften der beurteilten Personen für die Einschätzung durch die Beurteiler ausschlaggebend sind: Welche Indikatoren werden von den Beurteilern im Sinne einer hohen vs. niedrigen Merkmalsausprägung gewertet? Welche implizite Operationalisierung besitzen die Beurteiler von dem einzuschätzenden Merkmal [18]? Zur Lösung dieses Problems der Validität von Beurteilungen sind – anders als im Falle psychometrischer Tests – vorwiegend qualitative Methoden geeignet [14,19].

Werden psychometrisch konstruierte Fragebogenskalen (z. B. SCL-90-R, [6]; SF-36, [4]) eingesetzt, so ist es unerlässlich, Maßzahlen für die zentralen Gütekriterien Reliabilität und Validität zu berücksichtigen und bei einer Dokumentation der Ergebnisse mit anzugeben. Werden Experten- oder Beobachterurteile analysiert, so wird die Problematik der Güte der Messung erstaunlicherweise oft ignoriert, obwohl auch hier natürlich nicht davon ausgegangen werden kann, dass der Beurteiler den „wahren“ Zustand des Patienten unverfälscht einschätzen kann.

Im Folgenden soll die Reliabilitätsproblematik behandelt werden: Wie lässt sich die Reliabilität der Beurteilungen bestimmen? Mittels welcher Maßnahmen kann die Reliabilität erhöht werden? Obwohl im Text die Einschätzung von Personen als Beispiel verwendet wird, gelten alle Aussagen für die Beurteilung von Objekten (z. B. auch Kliniken) allgemein.

Welche Methoden sollten zur Bestimmung der Reliabilität verwendet werden?

Um die Beurteilerreliabilität zu bestimmen, muss – anders als in den meisten Anwendungssituationen – dieselbe Person von *mehreren* Beurteilern eingeschätzt werden. Welche statistischen Methoden dann zur Bestimmung der Reliabilität der Beurteilung eingesetzt werden sollten, hängt wesentlich von dem Messniveau der verwendeten Skala ab. Zwei Skalenarten müssen hier unterschieden werden [7]:

1. **Kategoriensysteme:** Es soll eine Klassifikation (z. B. Erstdiagnose nach ICD-10) vorgenommen werden, bei der über die Zugehörigkeit eines Patienten zu genau einer von mehreren vorgegebenen Kategorien entschieden werden soll (nominales oder kategoriales Urteil). Die Entscheidung, ob ein bestimmtes Merkmal vorhanden ist oder nicht, kann als Spezialfall eines solchen kategorialen Urteils mit zwei Kategorien aufgefasst werden.
2. **Rating- oder Einschätzskalen:** Es wird eine graduelle Abstufung einer Merkmalsausprägung erfasst (z. B. „– 3“ = „trifft überhaupt nicht zu“ … „+ 3“ = „trifft genau zu“, vgl. [2,10]).

Hier soll ausschließlich die Reliabilität von Beurteilungen mittels Ratingskalen vertieft werden. Zur Zuverlässigkeit von Zuordnungen zu Kategorien, die zumeist mittels Cohens Kappa überprüft wird, sei auf die entsprechende Literatur verwiesen [5,7,14 – 18].

In der klassischen Testtheorie ist die Reliabilität als der Anteil der Varianz der gemessenen Daten definiert, der durch die „wahren“ Werte erklärt werden kann. Eine Reliabilität von 0,9 bedeutet, dass die Einschätzungen eines Beurteilers zu 90 % von den tatsächlichen, „wahren“ Werten der beurteilten Personen determiniert werden und lediglich 10 % der Varianz der Daten durch Fehlereinflüsse bestimmt sind. Eine statistische Maßzahl ist nur dann ein Reliabilitätskoeffizient, wenn dieser eine erwartungstreue Schätzung des Varianzverhältnisses von „wahrer“ Varianz zu Gesamtvarianz darstellt. Die so genannte Intraklassenkorrelation (ICC) erfüllt genau diese Bedingung [9,12,18].[3] Ein Wert der

mehrere Erstdiagnosen innerhalb desselben Kategoriensystems erhalten. Zudem muss das Kategoriensystem oder die verwendete Ratingskala eine angemessene Abbildung des empirischen Merkmalsspektrums erlauben.

[2] Bei psychometrischen Tests gilt diese Beziehung nur eingeschränkt: Sehr hohe Reliabilität und Validität können nicht gleichzeitig erreicht werden (partielle Inkompatibilität der beiden Kardinalkriterien, [11], S. 39). Dieses Problem besteht für die Beurteilerreliabilität nicht: Hier ist eine möglichst hohe Reliabilität immer günstig für die Validität der Beurteilungen ([18], S. 208).

[3] Alternative Maße wie z. B. Finns r [15] für intervallskalierte Daten werden in der modernen Literatur nicht mehr diskutiert, da diese kein Re-

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

ICC von 0 indiziert vollständig zufälliges Beurteilerverhalten. Ein Wert von 1,0 liegt vor, wenn die Beurteilung perfekt reliabel erfolgt.

Für die Anwendung der ICC zur Bestimmung der Beurteilerreliabilität ist es wichtig zu beachten, dass insgesamt sechs verschiedene Versionen des Intraklassenkorrelationskoeffizienten existieren. Die Eigenschaften dieser in der Literatur und auch in Lehrbüchern selten erwähnten Gütemaße soll im Folgenden erläutert werden. In Abhängigkeit davon, wie die Beurteilerdaten in der Praxis eingesetzt oder interpretiert werden sollen, muss entschieden werden, welcher der Koeffizienten die Reliabilität angemessen quantifiziert. Hierfür sind drei Entscheidungskriterien ausschlaggebend:

Entscheidungskriterium 1: Sollen die von den Beurteilern vergebenen absoluten Messwerte unabhängig von der Person des Beurteilers weiterverwertet werden?
Ist dies der Fall, so ist es nicht nur notwendig, dass eine hohe Korrelation zwischen den Urteilen der Beurteiler besteht, sondern es muss zusätzlich gefordert werden, dass die *absoluten Werte,* die von verschiedenen Beurteilern für dieselbe Person vergeben werden, ähnlich sind.

Ein Beispiel soll diesen Punkt verdeutlichen: Zwei Therapeuten beurteilen die Behandlungsbedürftigkeit von Patienten mittels einer 10-stufigen Ratingskala (Wertebereich 1 – 10). Therapeut A ist sehr konservativ in seiner Einschätzung und vergibt Punktwerte im Bereich von 1 – 5. Therapeut B nutzt hingegen das obere Ende der Skala aus und schätzt die Behandlungsbedürftigkeit jedes Patienten um jeweils genau 5 Punkte höher ein als Therapeut A. Die Korrelation für diesen Datensatz ist genau gleich 1, aber die Übereinstimmung zwischen den vergebenen Punktwerten wäre sehr schlecht. Für die tatsächliche Anwendungssituation, in der jeder Patient nur von einem der beiden Therapeuten beurteilt wird, würde dies bedeuten, dass Patienten, die das Glück oder aber das Pech haben, von Therapeut B beurteilt zu werden, als behandlungsbedürftiger eingeschätzt werden, *unabhängig von ihrem tatsächlichen Zustand.*

In den meisten Anwendungssituationen muss verlangt werden, dass verschiedene Beurteiler *ähnliche absolute Werte* vergeben. Neben einer hohen Korrelation der Urteile müssen hierfür die Mittelwerte der Beurteiler ähnlich sein. Die so genannten *unjustierten* (nicht mittelwertkorrigierten) ICCs berücksichtigen diese *beiden* Informationsaspekte und quantifizieren somit die Beurteilerreliabilität bezüglich der absoluten Werte angemessen. In der Literatur ist auch die Bezeichnung ICC mit absoluter Übereinstimmungsdefinition gebräuchlich [9]. Im genannten Beispiel wäre – unter der Annahme, dass jeder Wert im Bereich von 1 – 5 von Therapeut A bzw. im Bereich 6 – 10 von Therapeut B gleich häufig vergeben wird – die ICC_{unjust} (zur Notation der verschiedenen ICCs s. ANHANG) gleich 0,17, was eine sehr schlechte Reliabilität indiziert.

Tab. **1** Beispiel für die Abhängigkeit der drei ICCs vom Mittelwertsunterschied zwischen zwei Beurteilern (übernommen aus [18]).

	Beurteiler 1	*Beurteiler 2*					
		$\bar{x}_{diff}$	*0*	*1*	*2*	*3*	*4*	*5*
Person 1	1	2	3	4	5	6	7
Person 2	2	1	2	3	4	5	6
Person 3	3	4	5	6	7	8	9
Person 4	4	5	6	7	8	9	10
Person 5	5	3	4	5	6	7	8
Person 6	6	6	7	8	9	10	11
Person 7	7	7	8	9	10	11	12
Person 8	8	9	10	11	12	13	14
Person 9	9	8	9	10	11	12	13
Person 10	10	10	11	12	13	14	15
$ICC_{unjust,einfakt}$		0,94	0,89	0,75	0,56	0,35	0,16
ICC_{unjust}		0,94	0,89	0,77	0,63	0,50	0,39
$r = ICC_{just}$		0,94	0,94	0,94	0,94	0,94	0,94

Justierte (mittelwertkorrigierte) Reliabilitätsmaße haben ähnliche Eigenschaften wie reine Korrelationsmaße, da ihr Wert unabhängig von Mittelwertsunterschieden zwischen Beurteilern ist. Justierte Maße nehmen systematisch höhere Werte an als unjustierte Maße, wenn der Mittelwert der von den Beurteilern vergebenen Werte – wie im genannten Beispiel – differiert. Die ICC_{just} darf jedoch nur als Reliabilitätsmaß verwendet werden, wenn bei der Interpretation der Daten nur die Position der beurteilten Personen in der *individuellen Rangreihe der Beurteiler* entscheidend ist. Für dieses Maß wird auch die Bezeichnung „ICC mit Konsistenzdefinition" verwendet [9]. In dem Beispiel ist der Wert der ICC_{just} gleich 1, weil von beiden Ratern dieselben Personen als relativ hoch bzw. niedrig behandlungsbedürftig eingeschätzt werden, obwohl die absoluten Werte deutlich differieren. Nur wenn das unterschiedliche Grundniveau der Beurteiler in der Punktvergabe bei der Interpretation der Daten in Rechnung gestellt wird, ist die ICC_{just} das angemessene Reliabilitätsmaß. Da dies in der Praxis eher selten vorkommt, sei an dieser Stelle auf den Artikel von Shrout u. Fleiss [12] verwiesen: Sie diskutieren Situationen, in denen eine solche Interpretation angemessen ist (s. auch [9,18]).

In Tab. **1** ist ein Datenbeispiel angegeben, das den Unterschied zwischen justierten und unjustierten Maßen nochmals anschaulich verdeutlicht. Beurteiler 1 hat für die 10 zu beurteilenden Personen die Werte in Spalte 2 vergeben. In Spalte 3 sind für einen zweiten Beurteiler Messwerte angegeben, die denselben Mittelwert und dieselbe Varianz wie die Urteile von Beurteiler 1 besitzen. Da die Mittelwerte der beiden Beurteiler gleich sind, nehmen unjustierte (zweit- und drittletzte Zeile) und justierte (letzte Zeile) Maße denselben Wert an. In den Spalten 4 – 8 ist die Wertereihe für Beurteiler 2 jeweils sukzessive um den Wert 1 verschoben. Je stärker sich die Mittelwerte der beiden Urteiler unterscheiden, desto niedriger ist der Wert der unjustierten

liabilitätsmaß darstellen und gezeigt werden kann, dass trotz unreliabler Beurteilungen hohe Koeffizientenausprägungen resultieren können. Lediglich moderne Latent-trait- oder Latent-class-Ansätze für ordinalskalierte Daten stellen eine Alternative zur ICC als Reliabilitätsmaß dar [1,17]. Das häufig angewendete Kendalls W für ordinalskalierte Daten ist ein Zusammenhangsmaß, aber kein Reliabilitätsmaß im eigentlichen Sinne [18].

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

ICCs. Hingegen bleibt die justierte ICC unverändert, da diese definitionsgemäß Mittelwertsunterschiede nicht als Indikator für mangelhafte Reliabilität wertet bzw. keine Ähnlichkeit der absoluten Werte verlangt wird. Für alle Daten ist zudem die justierte ICC gleich der Produkt-Moment-Korrelation, da die Varianzen der beiden Beurteiler gleich sind.

Entscheidungskriterium 2: Soll die Reliabilität des Urteils eines Beurteilers oder des Mittelwerts mehrerer Beurteiler bestimmt werden?

Hat man die Möglichkeit Einschätzungen mehrerer Beurteiler für dieselbe Person einzuholen, so besitzt der Mittelwert über die Einschätzungen mehrerer unabhängiger Beurteiler stets eine höhere Reliabilität als das Urteil eines einzelnen Beurteilers. Sowohl für die unjustierte als auch die justierte ICC kann nicht nur die Reliabilität des Urteils eines beliebigen Beurteilers berechnet werden, sondern auch die Reliabilität des Mittelwertes von mehreren Urteilern ($ICC_{unjust,MW}$, $ICC_{just,MW}$). Die Verwendung der Reliabilität der Mittelwerte als Gütemaß ist jedoch nur zulässig, wenn nicht nur in der Reliabilitätsstudie, sondern auch in der praktischen Anwendung stets die Daten von mehreren Beurteilern pro Person vorliegen und der entsprechende Mittelwert für jede Person als Beurteilungsdatum gebildet werden kann.

Entscheidungskriterium 3: Wurden alle Personen von denselben Beurteilern eingeschätzt?

Sind Personen jeweils von einer anderen Beurteilergruppe beurteilt worden, können lediglich die unjustierten Maße $ICC_{unjust,einfakt}$ bzw. $ICC_{unjust,einfakt,MW}$ als Reliabilitätsmaße berechnet werden.[4] Dies wäre beispielsweise dann der Fall, wenn an verschiedenen Kliniken jeweils eine andere, klinikinterne Gruppe von Ärzten die Patienten beurteilt: Wenn die Daten von allen Kliniken zusammengefasst oder gemeinsam analysiert werden sollen, könnten zur Bestimmung der Reliabilität lediglich diese beiden Maße zum Einsatz kommen. Wenn es möglich ist, sollte man jedoch auch dann, wenn ein unjustiertes Reliabilitätsmaß berechnet werden soll, darauf achten, dass alle Personen von denselben Beurteilern eingeschätzt werden, da die ICC_{unjust} eine genauere Bestimmung der Reliabilität ermöglicht als die $ICC_{unjust,einfakt}$. Justierte Reliabilitätsmaße können nur berechnet werden, wenn alle Personen von denselben Beurteilern eingeschätzt wurden.

In der Praxis ist man zumeist daran interessiert, wie zuverlässig ein beliebiger Beurteiler die Merkmalsausprägungen von Personen einschätzt. Die resultierende Einschätzung – unabhängig vom Beurteiler – wird dann weiterverwendet: Entsprechend ist in den meisten Fällen die ICC_{unjust} oder die $ICC_{unjust,einfakt}$ das angemessene Reliabilitätsmaß.

Die Berechnungsformeln für die genannten ICCs und Anleitungen zur einfachen Berechnung mit der SPSS-Funktion „Reliability" finden sich in der Literatur [18]. Dort werden auch die Voraussetzungen diskutiert, die für die Berechnung der ICCs erfüllt sein müssen (s. hierzu auch [9,12,15]). Eine Übersicht über die Entscheidungskriterien zur Selektion der korrekten ICC findet sich im Anhang.

Eine Beurteilerreliabilität ab 0,7 wird allgemein als „gute" Reliabilität klassifiziert, eine ICC unter 0,5 gilt als schwach oder nicht akzeptabel. Ob die Reliabilität einer Beurteilung zufrieden stellend ist, sollte aber auch immer von dem jeweils zu beurteilenden Merkmal abhängig gemacht werden. Für ein schwer zu erfassendes, komplexes Merkmal sind u.U. niedrigere Reliabilitäten akzeptabel als für ein Merkmal, das mittels alternativer Methoden unproblematisch zu erfassen ist.

In diesem Zusammenhang sei auf ein Problem hingewiesen, das bei der Interpretation von Reliabilitätskoeffizienten oft nicht beachtet wird: Die Reliabilität einer Messung verringert sich unter sonst gleichen Bedingungen, wenn die Merkmalsvarianz in der untersuchten Stichprobe abnimmt [8]: Beurteiler mögen beispielsweise in der Lage sein, hoch begabte von wenig begabten Personen zuverlässig zu unterscheiden, ohne dass sie notwendigerweise die Fähigkeit besitzen müssen, in einer homogenen Gruppe der Hochbegabten zuverlässige Differenzierungen vorzunehmen. Erst wenn Beurteiler zuverlässig die beurteilten Personen hinsichtlich der Merkmalsausprägung voneinander differenzieren können – was eine genügend hohe Präzision der Beurteilung *im Verhältnis* zur Merkmalsvariabilität in der untersuchten Personenstichprobe voraussetzt – liegt eine zufrieden stellende Reliabilität vor. Deswegen ist je nach Anwendungssituation dieselbe Präzision des Urteils unterschiedlich zu bewerten: Zufrieden stellende Reliabilität bedeutet, dass die Präzision einer Messung ausreicht, um die *vorhandenen* Unterschiede in der untersuchten Stichprobe zuverlässig zu entdecken. Hieraus folgt auch, dass eine noch so präzise Beurteilung niemals reliabel sein kann, wenn eine perfekt homogene Stichprobe vorliegt, in der die wahren Merkmalsausprägungen der beurteilten Personen nicht variieren.

Wie kann die Reliabilität von Beurteilungen verbessert werden?

Erweist sich die Reliabilität von Beurteilungen als nicht zufrieden stellend, so ist es notwendig, die Ursache von Unstimmigkeiten zwischen den Beurteilern zu identifizieren und in einem Beurteilertraining die Zuverlässigkeit der Beurteilungen gezielt zu verbessern. Um empirische Ansatzpunkte für ein Training zu erhalten, sollte deshalb neben der ICC die Korrelation jedes einzelnen Beurteilers mit dem Mittelwert der übrigen Beurteiler bestimmt werden (korrigierte Trennschärfen, [8]): Korreliert das Urteil eines oder weniger Beurteiler mit dem Mittelwert der übrigen Beurteiler gering, sind die korrigierten Trennschärfen der übrigen Beurteiler jedoch deutlich höher, so deutet dies darauf hin, dass die Mehrzahl der Beurteiler reliabel urteilt. Eine schwach ausfallende ICC kann dann durch die geringe Qualität des Urteils eines oder weniger Beurteiler bedingt sein. In einem Beurteilertraining müsste in diesem Fall nur die Leistung der betreffenden Urteiler verbessert werden: Hat das Training nicht den gewünschten Erfolg, so wäre es sinnvoll, den/die entsprechenden Beurteiler aus der Raterstichprobe auszuschließen.

[4] Die Bezeichnung „einfaktoriell" ist dadurch begründet, dass varianzanalytisch nur die Unterschiede zwischen den Personen und nicht zwischen den Beurteilern modelliert werden können. Da unterschiedliche Beurteiler die Personen geratet haben, können die Unterschiede zwischen den Beurteilern nicht fehlerfrei bestimmt werden.

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

Sind die korrigierten Trennschärfen für alle Beurteiler schwach, so muss in einem Beurteilertraining die Beurteilungsqualität aller Rater verbessert werden. Es wird durch die Beurteiler offensichtlich zu wenig Gemeinsames erfasst. Die Ursache hierfür kann darin liegen, dass kein Konsens entweder a) hinsichtlich des zu beurteilenden Merkmals oder b) hinsichtlich der entsprechenden Merkmalsausprägungen besteht. Ein Beispiel soll Punkt a verdeutlichen: Beurteilen Ärzte die Behandlungsbedürftigkeit, so mag Arzt A sich bei seinem Urteil eher an medizinischen Kriterien orientieren, für Arzt B ist der subjektiv geäußerte Leidensdruck des Patienten stärker ausschlaggebend. Dann wäre eine höhere Reliabilität u. U. dadurch zu erreichen, dass beide Ärzte sich auf eine Subkomponente als Beurteilungskriterium verständigen (z. B. Behandlungsbedürftigkeit = subjektiver Leidensdruck). Eine andere Möglichkeit bestünde darin, die beiden Unteraspekte explizit getrennt einzuschätzen, sodass von jedem Beurteiler eine Einschätzung des subjektiven Leidensdrucks *und zusätzlich* der medizinischen Kriterien vorgenommen wird.

Das wichtigste Instrument zur Verbesserung des Urteilerverhaltens ist also ein Feedback zwischen den Beurteilern, durch das die impliziten Operationalisierungen der Beurteiler transparent werden. Hierbei ist es hilfreich, beurteilte Personen zu identifizieren, für die entweder besonders starke oder aber sehr geringe Unterschiede in der Beurteilung vorliegen: Wenn geklärt ist, in welchen Fällen Unstimmigkeiten auftreten und worin diese Fälle sich von unproblematischen unterscheiden, kann die Ursache mangelnder Reliabilität zumeist sehr gut eingegrenzt werden.

Um den Konsens bezüglich der Einschätzung von Merkmalsausprägungen (s. o., Punkt b) zu verbessern, bietet es sich an, prototypische Fälle zu beschreiben, bei denen eine hohe vs. niedrige Merkmalsausprägung vorliegt. Ist dies nicht ausreichend, so können u. U. Subpopulationen (z. B. Diagnosegruppen) definiert werden, für die von reliabler vs. nicht reliabler Beurteilung ausgegangen werden kann. Für letztere ist dann eine intensivierte Diagnostik indiziert. Sind genügend Ressourcen vorhanden, so kann die Reliabilität auch gezielt erhöht werden, indem der Mittelwert der Einschätzungen mehrerer Beurteiler bestimmt wird.

Allgemein lässt sich festhalten, dass eine hohe Transparenz des zu beurteilenden Merkmals und der notwendigen Informationsverarbeitungsprozesse (z. B. Einschränkung des Interpretationsspielraums), eine an Beispielen orientierte Klärung typischer Merkmalsausprägungen, eine möglichst homogene Raterstichprobe (z. B. ähnliches Vorwissen) sowie eine heterogene Stichprobe der zu beurteilenden Personen günstig für die Reliabilität des Urteils sind. Das Urteil der Rater sollte zudem nicht durch Beurteilungstendenzen (z. B. Halo-Effekt, Tendenz zur Mitte, Tendenz zu den Extremen) verzerrt werden [3, 7, 18].

Nach jedem Beurteilertraining sollte an einer neuen Personenstichprobe die gewünschte Verbesserung überprüft werden: Wenn die betreffende ICC zufrieden stellend ist und zudem die Trennschärfen sowie die Mittelwerte der Rater homogen[5] sind, kann für die Anwendung von einer zuverlässigen Einschätzung durch die Beurteiler ausgegangen werden, die die Unterschiede zwischen den beurteilten Personen hinreichend genau widerspiegelt.

[5] Die Homogenität der Mittelwerte ist nicht notwendig, wenn eine der justierten ICCs die angemessene Reliabilitätskennziffer ist.

Literatur

[1] Agresti A. Modelling patterns of agreement and disagreement. Statistical Methods in Medical Research 1992; 1: 201 – 218
[2] Bortz J, Döring N. Forschungsmethoden und Evaluation. 3. Aufl. Berlin: Springer, 2001
[3] Bortz J, Lienert GA. Kurzgefasste Statistik für die klinische Forschung. Berlin: Springer, 2002
[4] Bullinger M, Kirchberger I. SF-36 Fragebogen zum Gesundheitszustand (SF-36). Handbuch für die deutschsprachige Fragebogenversion. Göttingen: Hogrefe, 1998
[5] Fleiss JL. Statistical methods for rates and proportions. New York: Wiley, 1981
[6] Franke GH. SCL-90-R – Die Symptom-Checkliste von L. R. Derogatis. Göttingen: Beltz-Test, 2002
[7] Greve W, Wentura D. Wissenschaftliche Beobachtungen. Eine Einführung. Weinheim: Psychologie-Verlags-Union, 1997
[8] Lienert G, Raatz U. Testaufbau und Testanalyse. 5. Aufl. Weinheim: Beltz, PVU, 1998
[9] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1996; 1: 31 – 43
[10] Rohrmann B. Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie 1978; 9: 222 – 245
[11] Rost J. Lehrbuch Testtheorie Testkonstruktion. Bern: Huber, 1996
[12] Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 1979; 86 (2): 420 – 428
[13] Stevens SS. Measurement, psychophysics, and utility. In: Churchman CW, Ratoosh P (eds). Measurement: definitions and theory. New York: Wiley, 1959
[14] Suen HK, Ary D. Analyzing quantitative behavioral observation data. Hillsdale, NJ: Erlbaum, 1989
[15] Tinsley HEA, Weiss DJ. Interrater reliability and agreement of subjective judgements. Journal of Counseling Psychology 1975; 22 (4): 358 – 376
[16] Uebersax JS. A review of modeling approaches for the analysis of observer agreement. Investigative Radiology 1992; 27: 738 – 743
[17] Uebersax JS. Statistical methods for rater agreement [Internet]. Verfügbar unter: http://ourworld.compuserve.com/homepages/jsuebersax, 2001, last updated May 2003
[18] Wirtz M, Caspar F. Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen: Hogrefe, 2002
[19] Wolf M. Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. Journal of Applied Behavior Analysis 1978; 1: 203 – 214

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

ANHANG

Notation zu den verschiedenen Intraklassenkorrelationskoeffizienten

Bedeutung der Subskripte:
unjust = unjustierte ICC. Die absoluten Messwerte werden unabhängig vom jeweiligen Beurteiler interpretiert oder weiter verwertet.

just = justierte ICC. Das individuelle Mittelwertsniveau der Beurteiler wird bei der Interpretation oder Weiterverwertung der Messwerte berücksichtigt.

einfakt = Die ICC basiert auf dem einfaktoriellen varianzanalytischen Modell.

MW = Es wird die Reliabilität des Mittelwerts aller untersuchten Beurteiler geschätzt.

Eigenschaften der 6 Formen der Intraklassenkorrelation:

	Kontrolle von Mittelwertsunterschieden der Beurteiler		*varianzanalytisches Design*		*Aggregation der Urteile*	
	nein unjustiert	*ja justiert*	*einfaktorielles Design*	*zweifaktorielles Design*	*nein Einzelwerte*	*ja Mittelwerte*
$ICC_{unjust,einfakt}$	ja		ja		ja	
$ICC_{unjust,einfakt,MW}$	ja		ja			ja
ICC_{unjust}	ja			ja	ja	
$ICC_{unjust,MW}$	ja			ja		ja
ICC_{just}		ja		ja	ja	
$ICC_{just,MW}$		ja		ja		ja

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