Zur Evaluation intraindividueller Veränderung[1]

Rolf Steyer
Institut für Psychologie der Friedrich-Schiller-Universität Jena
Wolfgang Hannöver
Forschungsstelle für Psychotherapie, Stuttgart
Christine Telser
Fachhochschule für Öffentliche Verwaltung, Lautzenhausen
Reinholde Kriebel
Gelderland-Klinik Geldern

Zusammenfassung

Grawe und Braun (1994) haben ein computerisiertes Verfahren zur Erfassung von Therapieeffekten auf der Basis von Differenzwerten vorgeschlagen. Dabei kann jedoch eine Post-Prä-Differenz, die ausschließlich auf Meßfehlern beruht, genauso groß aussehen, wie eine Post-Prä-Differenz, die auf perfekt reliablen Messungen basiert. Daher schlagen wir zwei einfache, alternative Veränderungskenngrößen vor, die die Unreliabilität psychologischer Messungen berücksichtigen: Die erste ermöglicht eine einfache Beurteilung der statistischen Signifikanz der Veränderung, die zweite dient zur Beschreibung des Ausmaßes einer Veränderung, gemessen in der Streuungseinheit des Vortests. Schließlich wird eine einfache Modifikation dieser Veränderungskenngrößen vorgestellt, die nicht nur das Meßfehlerproblem, sondern auch das Problem situationaler Effekte berücksichtigt.

Schlagworte: Evaluation, Qualitätssicherung, Psychotherapieforschung, Veränderungsmessung, Reliabilität, Konsistenz, State-Trait-Modelle

Abstract

Grawe and Braun (1994) proposed a computerized procedure for evaluating therapy effects, which is based on difference scores. However, in this procedure a post-pre-difference which is due only to measurement error may look as big as a post-pre-difference based on perfectly reliable measurements. Therefore, we propose two simple alternative indices which take into account the problem of measurement error. The first allows a simple evaluation of the statistical significance of change, the second serves to describe the degree of change in units of the pretest standard deviation. Finally, we present a simple modification of these indices taking into account not only the problem of measurement error, but also the problem of situational and/or interactional effects.

Keywords: Evaluation, psychotherapy research, measurement of change, reliability, consistency, state-trait models

[1] In Druck: *Zeitschrift für Klinische Psychologie*

1. EINFÜHRUNG

Die Erfassung und Dokumentation *intraindividueller* Veränderungen im klinisch-psychologischen Kontext ist für die Evaluation und Qualitätssicherung der praktisch-therapeutischen Arbeit von besonderer Bedeutsamkeit, da hier eine Veränderung beim Patienten Behandlungsziel ist. Um die Güte der Behandlung abzuschätzen, ist es wünschenswert, Ausmaß und Qualität der Veränderung zu erfassen. Hierzu wurden über die Jahre verschiedene Verfahren vorgeschlagen. Für einen Überblick s. z. B. Collins und Horn (1992), Harris (1963), Huber (1973), Petermann (1978), von Eye (1990a, b), oder Willet (1989).

Prinzipiell unterscheidet man hierbei zwischen *direkten* und *indirekten* Vorgehensweisen (Krauth, 1983). Die *direkten Verfahren* gründen sich auf Überlegungen von Bereiter (1963), der als grundlegendes Dilemma der Veränderungsmessung eine Diskrepanz zwischen subjektivem Erleben von Veränderung und dessen Abbildung auf eine objektive meßbare Größe sieht. Als Lösung schlägt er vor, als Veränderungen nicht die Differenz zweier Messungen zu erheben, sondern direkt die subjektiv erlebte Veränderung. Diese Idee wurde z. B. von Zielke und Kopf-Mehnert (1978) aufgegriffen und bei der Konstruktion des Veränderungsfragebogens zum Verhalten und Erleben (VEV) umgesetzt.

Indirekte Verfahren der Veränderungsmessung beruhen auf der Konstruktion *änderungssensitiver Items* (s. z. B. Krauth, 1983; Webster & Bereiter, 1963). Ein Beispiel liefert wiederum Zielke (1979) mit der Konstruktion der Kieler Änderungssensitiven Liste (KASSL). Dieses und andere Verfahren wurden verschiedentlich kritisiert (Krauth, 1983; Westmeyer, 1983). Allgemein versteht man unter der indirekten Methode der Veränderungsmessung die Betrachtung der Differenz zweier, von derselben Person erhobenen Meßwerte. Üblicherweise erhebt man eine bestimmte Variable am selben Patienten vor (Prätest) und nach der Behandlung (Posttest). (Im folgenden werden wir den Prätest mit X und den Posttest mit Y notieren.) Diese indirekten Verfahren können wir in mindestens drei verschiedene Gruppen einteilen, die sich hinsichtlich ihrer Zielrichtung unterscheiden: Deskriptive Verfahren, inferentiell schätzende Verfahren und inferentiell testende Verfahren.

Ein Beispiel für *deskriptive Verfahren* haben in jüngster Zeit z. B. Grawe und Braun (1994) vorgestellt, das auf einer normierten Post-Prätest-Differenz beruht. Vom Wert y des Posttests Y wird der Wert x des Prätests X subtrahiert und durch die Streuung $Std(X)$ des Prätests geteilt. Sie lehnen damit ihren Veränderungsindex an die Effektstärkenberechnung an, wie sie z. B. von Smith und Glass (1977) im Kontext der Metaanalyse verwendet wird. Sie benutzen eine z-Metrik, um verschiedener Skalen miteinander vergleichbar zu machen. Die Veränderungsindizes für verschiedene Skalen werden nicht isoliert betrachtet, sondern in Beziehung zu deren Mittelwerten in einer bestimmten Vergleichsgruppe gebracht. Auf diese Weise kann sich der Diagnostiker ein relativ umfassendes Bild des Veränderungsprofils machen.

Obwohl uns das von Grawe und Braun (1994) vorgestellte Verfahren insgesamt als sehr sinnvoll und nützlich erscheint, ist anzumerken, daß dabei das Meßfehlerproblem vernachlässigt wird. Die Konsequenz davon ist, daß eine Post-Prä-Differenz, die ausschließlich auf Meßfehlern beruht, genauso groß aussehen kann, wie eine Post-Prä-Dif-

ferenz, die auf perfekt reliablen Messungen basiert. In anderen Worten, der Veränderungsindex von Grawe und Braun kann nicht zwischen einer zufälligen, meßfehlerbedingten Fluktuation und einer systematischen, tatsächlichen Veränderung unterscheiden.

Bereits in den 50iger und 60iger Jahren wurde dieses Problem erkannt. So haben z. B. Dubois (1957) sowie Manning und Dubois (1962) vorgeschlagen, nicht die Differenz der Meßwerte als Veränderungsvariable zu betrachten, sondern die Differenz zwischen dem *erwarteten Posttestwert* und dem tatsächlichen Posttestwert. Der erwartete Wert des Posttests wird dabei unter der Hypothese aus dem Wert des Prätests berechnet, daß zwischen den beiden Messungen keine Veränderung eingetreten ist. Diese Überlegung greift auch Lord (1963) auf, wenn er schreibt: „ ... a common-sense consideration of the simple difference scores may be misleading. What is needed is to compare the data actually obtained with the data that would have been obtained under the appropriate null hypothesis of no treatment effect“ (Lord, 1963, p. 25). Eine einfache aber wichtige Verbesserung des Verfahrens von Grawe und Braun (1994) bestände also darin, die einfache Post-Prätest-Differenz $Y - X$ durch die Differenz $Y - E_0(Y \mid X)$ zu ersetzen, wobei $E_0(Y \mid X)$ die Regression von Y auf X unter der Nullhypothese bezeichnet, daß, abgesehen von meßfehlerbedingten Fluktuationen, keine Veränderung stattgefunden hat (mehr dazu in den Abschnitten 3 und 4).

Ein Beispiel für *inferentiell schätzende Verfahren* hat Lord (1963) vorgestellt. Ziel ist dabei, die meßfehlerbereinigte Veränderung, d. h. die Differenz zwischen dem wahren Wert der Person zum Zeitpunkt 2 und ihrem wahren Wert zum Zeitpunkt 1 zu schätzen. Lord gibt dafür eine Schätzformel an, die auf einer linearen Regression der wahren Veränderung auf Prä- und Posttestwerten beruht. Cronbach und Furby (1970) haben später darauf hingewiesen, daß man bei einer solchen Schätzung auch andere Variablen als Prä- und Posttest als Regressoren verwenden kann. Sind weitere Variablen mit der zu schätzenden Veränderung korreliert, dann kann deren Einbeziehung als Regressoren zu einer präziseren Schätzung der wahren Veränderung führen.

Als *inferentiell testendes Verfahren* kann die *kritische Differenz* angesehen werden, die z. B. von Lienert (1961, S. 454) oder von Huber (1973, S. 131) dargestellt wird. Ziel dieses Verfahrens ist es, eine Differenz zwischen Post- und Prätest anzugeben, von der an eine intraindividuelle Veränderung auf einem zu wählenden Niveau α (z. B. $\alpha = 0.05$) signifikant ist. Dabei geht man von der Nullhypothese unveränderter *wahrer Werte* der Person aus, berechnet (unter der Annahme unkorrelierter Meßfehler und für alle Personen gleicher Meßfehlervarianz) die Streuung der Post-Prätestdifferenz $Y - X$ und teilt diese Post-Prätestdifferenz durch ihre Streuung. Unter der Annahme der Gültigkeit der Nullhypothese und der anderen beiden genannten Voraussetzungen resultiert ein z-Wert, der dann unter der Voraussetzung einer Normalverteilung zur Prüfung der Nullhypothese „keine intraindividuelle Veränderung“ verwendet werden kann. Dieses Verfahren findet sich auch unter der Bezeichnung *clinical significance* bei Jacobson und Truax (1991) oder *clinically significant change* bei Speer (1992) wieder.

In diesem Artikel werden wir sowohl ein weiteres *deskriptives* als auch ein weiteres *inferentiell testendes* Verfahren vorstellen. Während ersteres den von Grawe und Braun (1994) vorgeschlagenen Effektstärkeindex verbessern soll, kann man letzteres als Alternative zur kritischen Differenz betrachten, das auch zur zufallskritischen Ab-

sicherung des deskriptiven Kennwerts dient. Anstelle der einfachen Post-Prätest-Differenz werden wir in beiden Verfahren die Differenz zwischen dem tatsächlichen Posttestwert und dem (unter der Nullhypothese „keine wahre intraindividuelle Veränderung“) erwarteten Posttestwert betrachten und *z*-transformieren, so daß, wie bei der kritischen Differenz, wieder eine einfache *z*-Statistik entsteht, bei der allerdings bei der Konstruktion dieser Statistik das Problem der Regression zur Mitte berücksichtigt wird. Außerdem stellen wir einige Konsequenzen für den Fall dar, wenn man nicht nur berücksichtigen will, daß psychologische Messungen meßfehlerbehaftet sind, sondern auch situativen und/oder interaktiven Effekten unterliegen.

2. Regression zur Mitte

Um die Betrachtung der Differenz zwischen dem tatsächlichen Posttestwert und dem unter der Nullhypothese „keine *wahre* (d. h. in den Werten der True-score-Variablen) Veränderung“ erwarteten Posttestwert anstelle der einfachen Post-Prätest-Differenz zu motivieren, stellen wir zunächst ein einfaches illustratives Beispiel dar, wo es plausibler ist, bei denjenigen Personen eine Veränderung anzunehmen, für die Prä- und Posttestwerte jeweils gleich sind, als bei denjenigen Personen, für die Prä- und Posttestwerte unterschiedlich sind. Damit soll verdeutlicht werden, daß man einfachen Post-Prä-Differenzwerten nicht ansehen kann, ob ihnen eine wahre Veränderung zugrunde liegt oder nur meßfehlerbedingte Fluktuationen.

Angenommen zwei Personen würden hinsichtlich eines Ratings mit den Stufen 0 bis 4, das vor und nach einer Therapie erhoben wurde, miteinander verglichen. Patient A habe vor der Therapie ein Rating von 4 und nach der Therapie ein Rating von 4, Patient B habe vor der Therapie ebenfalls ein Rating 4 und nachher dagegen ein Rating von 3. Ein Vergleich der Post-Prä-Differenzen würde dem Patienten B eine Veränderung zuschreiben, dem Patienten A hingegen nicht. Diese Interpretation ist jedoch äußerst fragwürdig, wenn wir davon ausgehen, daß das Rating fehlerbehaftet ist. Im folgenden Beispiel zeigen wir, daß die Annahme viel plausibler ist, daß sich Patient A (mit dem Prätestwert 4 und dem Posttestwert 4), nicht aber Patient B (mit dem Prätestwert 4 und dem Posttestwert 3) verändert hat.

Mit Hilfe eines PC-Programms haben wir zwei kontinuierliche Variablen X^* und Y^* mit einer bivariaten Normalverteilung mit der Korrelation 0.63 erzeugt. Danach wurde jede dieser kontinuierlichen Variablen in fünf Stufen aufgeteilt, so daß zwei fünfstufige Variablen X und Y mit den Werten 0 bis 4 entstanden, wie sie bei Ratings (z. B. des Befindens) vor und nach einer Therapie vorliegen könnten. Die Randverteilungen der beiden Variablen X und Y wurden durch folgende Wahrscheinlichkeiten festgelegt:

$$P(X = 0) = P(Y = 0) = 0{,}0228$$
$$P(X = 1) = P(Y = 1) = 0{,}1359$$
$$P(X = 2) = P(Y = 2) = 0{,}6826$$
$$P(X = 3) = P(Y = 3) = 0{,}1359$$

$$P(X = 4) = P(Y = 4) = 0{,}0228.$$

Die Korrelation zwischen X und Y beträgt dann 0.50. Nehmen wir an, daß diese Variablen *parallel* im Sinne der Klassischen Testtheorie sind (d. h. X und Y haben die gleiche True-score-Variable, unkorrelierte Meßfehler und gleiche Fehlervarianzen), dann ist deren Korrelation zugleich auch ihre Reliabilität. (Die Höhe der Korrelation ist also durchaus realistisch gewählt.) Die bedingten Wahrscheinlichkeiten für die Y-Variable bei gegebenem Wert der X-Variable berechnen sich dann wie in Tabelle 1 aufgeführt.

Tabelle 1. *Bedingte Wahrscheinlichkeiten für die Posttestwerte bei gegebenen Prätestwerten*

		Posttest Y				
		0	1	2	3	4
Prätest X	0	0.26	0.47	0.27	0.00	0.00
	1	0.08	0.36	0.56	0.01	0.00
	2	0.01	0.11	0.76	0.11	0.01
	3	0.00	0.01	0.56	0.36	0.08
	4	0.00	0.00	0.27	0.47	0.26

Die Tabelle 1 veranschaulicht, daß der erwartete Posttestwert zur Mitte „regrediert."[2] So beträgt z. B. die (bedingte) Wahrscheinlichkeit, daß der Posttest Y den Wert 4 annimmt, falls der Prätest X ebenfalls 4 ist, nur 0.26, wohingegen die Wahrscheinlichkeit, daß Y den Wert 3 annimmt, wenn X gleich 4 ist, 0.47 beträgt, also fast doppelt so groß ist. (Tatsächlich ist 3 auch der bedingte Erwartungswert des Posttests bei gegebenem Vortestwert 4.) Dieses Beispiel zeigt, daß in der Regel allein durch die Unreliabilität der Messungen, Veränderungen in der Ausprägung der Meßwerte zu erwarten sind, auch wenn keine tatsächliche Veränderung stattgefunden hat. Die diesem Beispiel zugrundegelegte Annahme der Parallelität im Sinn der Klassischen Testtheorie beinhaltet ja unter anderem, daß keine wahre (d. h. in den wahren Werten) Veränderung zwischen Prä- und Posttest stattgefunden hat. In diesem Fall ist es also tatsächlich plausibler, bei denjenigen Personen eine Veränderung anzunehmen, bei denen die Prä- und Posttest-

[2] Allgemein versteht man unter *Regression zur Mitte* den durch die folgende Formel beschriebenen Sachverhalt: $|x - E(X)| / \mathrm{Std}(X) > |[E(Y|X{=}x) - E(Y)]| / \mathrm{Std}(Y)$, d. h. der Betrag der Differenz zwischen dem Wert x eines Regressors X und seinem Erwartungswert $E(X)$ geteilt durch die Standardabweichung von X ist größer als der Betrag der Differenz zwischen dem bedingten Erwartungswert des Regressanden gegeben $X = x$ und dem unbedingten Erwartungswert geteilt durch die Standardabweichung von Y. Man kann zeigen, daß diese Regression zur Mitte *immer* vorliegt, wenn $x \neq E(X)$ und die Regression $E(Y|X)$ nicht perfekt ist, d. h. wenn $E(Y|X) \neq Y$.

werte jeweils gleich 4 sind, als bei den Personen mit Prätestwert 4 und Posttestwert 3.

Das Fazit aus den bisherigen Ausführungen lautet also: Einer Differenz zwischen zwei Ratings oder zwei Tests (jeweils vor und nach einer Therapie erhoben) kann man nicht ansehen, ob sie eine tatsächliche Veränderung wiedergibt oder nur eine zufällige, meßfehlerbedingte Schwankung. Das ändert sich übrigens auch dann nicht, wenn eine solche Post-Prätest-Differenz durch die Standardabweichung des Prätests geteilt wird, wie dies Grawe und Braun (1994) vorgeschlagen haben.

3. Eine inferentielle Veränderungskenngrösse

Anstelle der einfachen Post-Prätest-Differenz $Y - X$ schlagen wir also, wie auch schon Manning und Dubois (1962), vor, die Differenz zwischen dem tatsächlichen Posttestwert und dem (unter der Nullhypothese „keine wahre intraindividuelle Veränderung“) erwarteten Posttestwert zu betrachten. Teilen wir dann diese Differenz durch ihre Standardabweichung, dann liegt eine einfach interpretierbare *z*-Statistik vor, aus der man sofort ablesen kann, ob eine signifikante intraindividuelle Veränderung angenommen werden kann, wenn man die Annahme der Normalverteilung hinzufügt. (Falls keine Normalverteilung vorliegt, hat man u. E. trotzdem noch einen brauchbaren Anhaltspunkt, um eine Veränderung zu diagnostizieren, auch wenn dann, strenggenommen, kein valider Signifikanztest mehr vorliegt.)

Diese *inferentielle Veränderungskenngröße* ist also wie folgt definiert:

$$V_{\text{infer}} := \frac{Y - E_0(Y|X)}{Std[Y - E_0(Y|X)]}, \tag{1}$$

wobei $E_0(Y \mid X)$ die Regression des Posttests auf den Prätest unter der Nullhypothese „keine wahre Veränderung“ darstellt. Präzisieren wir diese Nullhypothese durch die Annahme, daß X und Y parallel im Sinn der Klassischen Testtheorie sind (d. h. X und Y haben die gleiche True-score-Variable, unkorrelierte Meßfehler und gleiche Fehlervarianzen), und daß die Regression $E_0(Y \mid X)$ linear ist, daß also

$$E_0(Y \mid X) = \alpha_0 + \alpha_1 X, \tag{2}$$

gilt, dann kann man diese Veränderungskenngröße folgendermaßen berechnen:

$$V_{\text{infer}} = \frac{[Y - E(X)] - Rel(X) \cdot [X - E(X)]}{Std(X)\sqrt{1 - Rel(X)^2}}, \tag{3}$$

wobei *Rel*(*X*) die Reliabilität von *X* bezeichnet. (Zum Beweis siehe den Anhang.) Dabei beachte man, daß die Gleichung 3 auf der Nullhypothese „keine wahre Veränderung" basiert, wie sie mit den o. g. Annahmen paralleler Tests präzisiert ist. Unter einer anderen Präzisierung dieser Nullhypothese würde sich auch eine von Gleichung 3 verschiedene Formel ergeben (s. z. B. Foerster, 1995). Fügt man die Normalverteilungsannahme hinzu, ist diese Veränderungskenngröße standardnormalverteilt. Bei einem Wert $V_{\text{infer}} \geq 1.96$ kann man dann die Nullhypothese, daß keine wahre intraindividuelle Veränderung -- egal in welcher Richtung -- stattgefunden hat, auf dem 5%-Niveau verwerfen. Bei einer einseitigen Fragestellung, in der es z. B. nur um die *Verbesserung* der betrachteten Testwertvariablen geht, würde man schon ab einem Wert von $V_{\text{infer}} \geq 1.65$ von einer signifikanten Veränderung auf dem 5%-Niveau sprechen können.

In der Regel wird man den Erwartungswert, die Varianz und die Reliabilität nur über eine Stichprobe schätzen können. Bei hinreichend großen Stichproben sollte die o. g. Entscheidungsregel jedoch zu keinen nennenswerten Verfälschungen führen.

4. Eine deskriptive Veränderungskenngrösse

Natürlich ist auch das von Grawe und Braun (1994) anvisierte Ziel einer intraindividuellen Effektstärke für deskriptive Zwecke sinnvoll. Wie bereits in der Einleitung angemerkt, wird jedoch bei deren Vorgehen das Meßfehlerproblem mit der Konsequenz vernachlässigt, daß eine Post-Prä-Differenz, die ausschließlich auf Meßfehlern beruht, genauso groß aussehen kann, wie eine Post-Prä-Differenz, die auf perfekt reliablen Messungen basiert. Daher kann der Veränderungsindex von Grawe und Braun nicht zwischen einer zufälligen, meßfehlerbedingten Fluktuation und einer systematischen, tatsächlichen Veränderung unterscheiden.

Wie kann man eine intraindividuelle Effektstärke ohne diesen Nachteil definieren? Nach den bisher dargestellten Überlegungen ist die Antwort nun recht einfach. Anstelle der Post-Prätest-Differenz setze man einfach die Differenz zwischen dem unter der Nullhypothese „keine wahre intraindividuelle Veränderung" *erwarteten Posttestwert* und dem tatsächlichen Posttestwert, d. h. den Zähler der von uns vorgestellten inferentiellen Veränderungskenngröße und teile diesen Zähler dann durch die Standardabweichung des Prätests:

$$V_{\text{deskript}} := \frac{Y - E_0(Y|X)}{Std\ (X)}. \tag{4}$$

Unter der im letzten Abschnitt genannten Präzisierung der Nullhypothese „keine wahre intraindividuelle Veränderung" folgt dann:

$$V_{\text{deskript}} := \frac{[Y - E(X)] - Rel(X)\cdot[X - E(X)]}{Std\ (X)}. \tag{5}$$

Die Abweichung der tatsächlichen Veränderung von der unter der Nullhypothese erwarteten Veränderung wird hier also in der Streuungseinheit des Prätests gemessen. Diese Normierung ist zwar recht willkürlich und andere Normierung wären ebenfalls möglich, aber sie ist an der üblichen Effektstärkennormierung orientiert und „mißt“ Veränderungen eben in den relativ anschaulichen Streuungseinheiten des Vortests.

Als Quintessenz unserer bisherigen Ausführungen schlagen wir folgende Änderungen des Vorgehens von Grawe und Braun (1994) vor: Zur Beschreibung verwende man anstelle von $(Y - X) / Std(X)$ die deskriptive Veränderungskenngröße V_{deskript} und markiere diese mit einem Sternchen, falls $V_{\text{infer}} \geq 1.96$ (5%-Signifikanzniveau) bzw. mit einem Kreuz, falls $V_{\text{infer}} \geq 1.65$ (10%-Signifikanzniveau). (Bei einseitigen Fragestellungen liegen die entsprechenden Signifikanzniveaus bei 2.5% bzw. 5%). Die Voraussetzungen zur Anwendung dieses Verfahrens sind nicht anders als die bei Grawe und Braun: hinreichend große Stichproben für den Prätest *X*, aus denen Erwartungswert, Standardabweichung und Reliabilität geschätzt werden können. Gegebenenfalls können diese Größen auch aus vergleichbaren Stichproben, die z. B. in einem Testmanual publiziert sind, übernommen werden.

5. Ein Anwendungsbeispiel

Dieses Verfahren soll am Beispiel der Skala *Erschöpfung* aus dem *Gießener Beschwerdebogen* von Brähler und Scheer (1983) illustriert werden. Der Gießener Beschwerdebogen wurde den Patienten der Gelderland Klinik vor und nach dem Klinikaufenthalt vorgelegt. Die Skala besteht aus den Items *Schwächegefühl*, *Schlafbedürfnis*, *Erschöpfbarkeit*, *Müdigkeit*, *Benommenheit* und *Mattigkeit*, die mit den Kategorien 0 (nicht) bis 4 (stark) beurteilt werden. Gefragt wird: Wie stark fühlen Sie sich durch die folgenden Beschwerden belästigt? Der Testwert der Skala *Erschöpfung* wird dann aus der Summe der Werte der o. g. sechs Items gebildet. Für eine Patientengruppe von 777 Patienten ergaben sich die in Tabelle 2 angegebenen Kennwerte.

Tabelle 2. *Mittelwerte, Standardabweichungen und Reliabilität der Skala „Erschöpfung“ des Gießener Beschwerdebogens*

	Mittelwert	Standardabweichung	Reliabilität
Prätest	11.04	5.84	0.87
Posttest	7.47	5.36	

Anmerkung: Die Angaben beziehen sich auf $N = 777$ Patienten der Gelderland-Klinik.

Für eine Person mit einem Prätestwert von: $X = 13$ und einem Posttestwert von: $Y = 7$ ergäbe sich also:

$$V_{\text{infer}} = \frac{[7 - 11.04)] - 0.87 \cdot [13 - 11.04]}{5.84\sqrt{1 - 0.87^2}} = \frac{-4.04 - 1.71}{5.84\sqrt{0.24}} = -2.0,$$

$$V_{\text{deskript}} = \frac{[7 - 11.04)] - 0.87 \cdot [13 - 11.04]}{5.84} = \frac{-4.04 - 1.71}{5.84} = -0.98.$$

Der *inferentielle* intraindividuelle Veränderungskennwert dieser Person beträgt in diesem Beispiel also −2.0 und ist damit, bei zweiseitiger Testung auf dem 5% Niveau statistisch signifikant. Damit können wir die Nullhypothese, daß es sich hier um eine zufällige, meßfehlerbedingte Veränderung handelt, verwerfen. Der deskriptive intraindividuelle Veränderungskennwert dieser Person dagegen beträgt in diesem Beispiel −0.98. Die Verringerung der Erschöpfung dieser Person liegt demnach in der Größenordnung bei ca. einer Streuungseinheit der Skala im Vortest.

Würde man die Effektsärke nach Grawe und Braun (1994) berechnen, ergäbe sich (7 - 13)/5.84 = 1.02. Die deskriptiven Veränderungsindices unterscheiden sich in diesem Beispiel eines hoch reliablen Instruments also kaum. Lediglich der inferentielle Veränderungsindex gibt noch die zusätzliche Information über die statistische Signifikanz.

Ganz anders sieht dies jedoch aus, wenn man ein sehr unreliables Instrument verwendet, das ansonsten die gleichen Eigenschaften, insbesondere gleichen Mittelwert und gleiche Varianz im Prätest hat. Der Index von Grawe und Braun wäre bei einem Prästestwert von 13 und einem Posttestwert von 7 unverändert gleich (7 − 13) / 5.87 = −1.02, selbst dann die Reliabilität des Instruments gleich 0 wäre. Ohne die Berücksichtigung der Information über die Reliabilität würde man auch hier von einer relativ starken Veränderung sprechen, obwohl sie ausschließlich durch Meßfehlerfluktuationen zu erklären wäre. Sowohl der von uns vorgeschlagene inferentielle als auch der deskriptive Veränderungskennwert wären in diesem Extremfall dagegen gleich (7 - 11.04) / 5.87 = −0.69 und damit *nicht* signifikant.

Prüft man auf dem 5%-Niveau würde man bei einer Reliabilität von 0 auch nur 5% „signifikante“ Werte beim inferentiellen Veränderungskennwert erhalten. Auch hier wäre man nicht völlig gegen Fehlurteile geschützt, aber diese würden sich eben in dem üblichen, bekannten Rahmen halten, den man selbst durch das Signifikanzniveau festsetzt.

6. Die Berücksichtigung situativ bedingter Veränderungen

Unsere bisherigen Überlegungen bewegen sich ganz im Rahmen der Klassischen Testtheorie, in der die einzigen Varianzquellen „wahre Werte“ der Personen und „Meß-

fehler“ sind. Nun ist aber nach vielen Jahren der Situationismus und Interaktionismus-Debatte nicht mehr zu übersehen, daß auch bei intendierter Messung von Person- bzw. Persönlichkeitseigenschaften mit situativen und/oder interaktiven (d. h. Interaktion zwischen Person und Situation) Effekten zu rechnen ist: *Psychologische Messung finden niemals in einem situationalen Vakuum statt.* Bei einem Test, der lediglich zu einer einzigen Meßgelegenheit erhoben wird, mißt man, genau genommen, gar nicht die Person, sondern die Person-in-der-Situation (s. z. B. Anastasi, 1983). In welchem Ausmaß auch situative und interaktive Effekte eine beobachtbare Testwertvariable bestimmen, kann man unter bestimmten Annahmen aus der Differenz zwischen der Paralleltest-Korrelation und der Retest-Korrelation ersehen. Parallele Tests, die innerhalb einer Meßgelegenheit erhoben werden, werden i. d. R. auch in derselben psychischen und objektiven Situation erhoben, in der sich die betreffende Person gerade befindet. Diese Situation ist eine potentielle systematische Varianzquelle, die sich also *in beiden* parallelen Messungen niederschlagen könnte. Situative und interaktive Effekte sind Bestandteil der Varianz *zwischen* verschiedenen Personen, da sich jede Person bei der Messung in einer für sie spezifischen psychischen und/oder objektiven Situation befindet. Man denke hier bspw. an die unterschiedliche Dauer des Schlafs in der Nacht vor einer Leistungsmessung oder an unterschiedliche Stimmungen, wenn Persönlichkeitseigenschaften erhoben werden sollen. Wenn die beiden Paralleltests zur *selben* Meßgelegenheit erhoben werden, gehen diese und viele andere situativ bedingten Unterschiede *systematisch* in beide Paralleltests ein und erhöhen deren Korrelation. Werden dagegen zwei Tests, die dieselbe Personeigenschaft messen, zu unterschiedlichen, hinreichend weit auseinanderliegenden Meßgelegenheiten erhoben, ist es plausibel anzunehmen, daß die Retest-Korrelation *nicht* durch situativ und interaktiv bedingte Effekte erhöht sein wird. Demnach kann man unter einigen relativ plausiblen Annahmen aus der Differenz zwischen Parallel- und Retestkorrelation das Ausmaß situativer und/oder interaktiver Effekte abschätzen.

In der *Latent-State-Trait-Theorie* (LST-Theorie; s. z. B. Steyer, Ferring & Schmitt, 1992; Steyer & Schmitt, 1990) wird die Klassische Testtheorie (KTT) insofern erweitert, als daß nicht mehr nur zwischen wahrem Wert und Meßfehler, sondern zwischen latenten Traitvariablen, latenten Statevariablen und Meßfehlern unterschieden wird. Der Wert einer latenten Traitvariablen charakterisiert die Person, wohingegen der Wert einer latenten Statevariablen ein Merkmal der Person-in-der-Situation ist. Entsprechend wird auch der Reliabilitätskoeffizient der Klassischen Testtheorie in der LST-Theorie durch den *Konsistenzkoeffizienten* ergänzt, den durch den Personfaktor erklärten Varianzanteil der Testwertvariablen, wohingegen der *Reliabilitätskoeffizient* der LST-Theorie der durch die Faktoren *Person* und *Situation* (incl. ihrer Interaktion) erklärte Varianzanteil ist. Unter bestimmten Annahmen, die im wesentlichen analog zu den Paralleltestannahmen der KTT sind, kann der Konsistenzkoeffizient durch die Retestkorrelation geschätzt werden.

Deinzer et al. (1995) haben für verschiedene Persönlichkeitstests gezeigt, daß der Konsistenzkoeffizient bei den meisten der involvierten Skalen um 5 bis 10 Prozentpunkte niedriger liegt als der zugehörige Reliabilitätskoeffizient. Das heißt zugleich, daß 5 bis 10 Prozent der Varianz von Persönlichkeitsskalen auf situative und interaktive Effekte zurückzuführen sind.

Was hat dies für Konsequenzen für die Diagnostik intraindividueller *Traitveränderungen*? Will man nicht Veränderungen diagnostizieren, die möglicherweise nur situativ bedingt sind, muß man in den von uns oben vorgestellten Kenngrößen die Reliabilität des Prätests durch die *Konsistenz* des Vortests (s. dazu Steyer, Ferring & Schmitt, 1992) ersetzen. Die beiden Kenngrößen können dann wie folgt berechnet werden:

$$V_{\text{infer}} = \frac{[Y - E(X)] - Con(X) \cdot [X - E(X)]}{Std(X)\sqrt{1 - Con(X)^2}}, \tag{6}$$

$$V_{\text{deskript}} := \frac{[Y - E(X)] - Con(X) \cdot [X - E(X)]}{Std(X)}. \tag{7}$$

7. Diskussion

Mit der in Abschnitt 3 dargestellten inferentiellen Veränderungskenngröße kann man relativ einfach entscheiden, ob eine statistisch signifikante intraindividuelle Veränderung vorliegt. Ist man eher, oder, in Ergänzung dazu, auch an einer Kenngröße interessiert, mit der man das Ausmaß der Veränderung in Einheiten der Standardabweichung des Vortests abschätzen und verschiedene Skalen miteinander vergleichen kann, bietet sich die entsprechende deskriptive Veränderungskenngröße (s. Abschnitt 4) an. Im Gegensatz zu der von Grawe und Braun (1994) vorgestellten Kenngröße wird dabei in Rechnung gestellt, daß die verwendeten Maße nicht fehlerfrei erhoben werden. Während bei Grawe und Braun eine Post-Prä-Differenz, die ausschließlich auf Meßfehlern beruht, genauso groß aussehen kann, wie eine Post-Prä-Differenz, die auf perfekt reliablen Messungen basiert, kann dies bei der hier vorgestellten Kenngröße nicht passieren. Darüber hinaus kann man sich mit der inferentiellen Veränderungskenngröße dagegen absichern, daß eine Veränderung nur durch zufällige Fluktuationen zustande kommt. Schließlich kann man, die neueren Erkenntnisse der Situationismus-Interaktionismus-Debatte in Rechnung stellend, den Konsistenzkoeffizienten anstelle des Reliabilitätskoeffizienten verwenden, wenn es explizit um die Diagnose einer *Traitveränderung* geht.

Auch wenn wir hier nur die Veränderung einer einzigen Variablen betrachtet haben, darf dies nicht so mißverstanden werden, als daß wir anregen wollten, bei der Evaluation einer Therapie ebenfalls nur eine einzige Variable zu betrachten. Daher sei hier noch einmal hervorgehoben, daß wir den Vorschlag von Grawe und Braun (1994), ein Veränderungsprofil bzgl. mehrerer Variablen gleichzeitig zu betrachten, sehr befürworten. Unser Vorschlag bezieht sich nur auf die Berücksichtigung der Meßfehlerbehaftetheit schon bei der *Berechnung* des Veränderungskennwerts, und nicht erst in seiner Interpretation, bei der Grawe und Braun sicherlich eine mangelhafte Reliabilität

ebenfalls berücksichtigen würden.

Schließlich sei angemerkt, daß keine der beiden von uns vorgeschlagenen Veränderungskenngrößen als Variable bei Untersuchungen zu interindividuellen Unterschieden intraindividueller Veränderungen herangezogen werden sollte. Warum diese Einschränkung? Die Antwort lautet: Dazu gibt es bessere Verfahren, bei denen eine (notwendigerweise wieder fehlerbehaftete) Schätzung intra-individueller Veränderungskennwerte unnötig ist. Bei diesen alternativen Verfahren ist es *nicht* nötig, die intraindividuellen Veränderungskennwerte zu berechnen (genauer: zu schätzen), um die dabei entstehende Variable mit anderen, möglicherweise erklärenden Variablen zu korrelieren; stattdessen werden die betreffenden Korrelationen und anderen Zusammenhangsmaße direkt, basierend auf einem Modell, geschätzt (s. dazu Steyer, Eid und Schwenkmezger, in Druck). Wir beziehen uns hier auf dasselbe Prinzip, das auch schon für einfache faktorenanalytische Modelle gilt: Auch hier sollte man *nicht* die Faktorscores schätzen und miteinander korrelieren, wenn man an den Korrelationen der betreffenden Faktoren interessiert ist. Stattdessen kann man im Rahmen von Strukturgleichungsmodellen die betreffenden Korrelationen zwischen den Faktoren direkt und genauer schätzen, ohne dabei die (schätzfehlerbehafteten) Faktorscores verwenden zu müssen, was zu einer verfälschten Schätzung der Korrelation zwischen den Faktoren führen würde.

Fassen wir zusammen! Wir meinen, daß das von Grawe und Braun (1994) vorgeschlagene Verfahren sehr nützlich ist, aber in mehrerlei Hinsicht verbessert werden kann. Zum einen ist unsere deskriptive Veränderungskenngröße meßfehlerkorrigiert, ohne daß die Interpretation als Effektstärkenmaß verloren geht. (Allerdings halten wir die Bezeichnung „Effektstärke“ in diesem Kontext für etwas gefährlich, weil hier zu schnell eine kausale Interpretation vorgenommen könnte, die aber i. d. R. in keiner Weise abgesichert ist.) Zum anderen wird diese deskriptive Veränderungskenngröße durch eine entsprechende inferentielle Veränderungskenngröße ergänzt, mit der sehr schnell die statistische Signifikanz der intraindividuellen Veränderung beurteilt werden kann. Schließlich kann man mit den im letzten Abschnitt modifizierten Veränderungskenngrößen nicht nur dem Meßfehlerproblem begegnen, sondern auch situative und/oder interaktive Effekte angemessen berücksichtigen, wenn es um die Diagnostik einer Traitveränderung geht.

Literatur

Anastasi, A. (1983). Traits, states and situations: A comprehensive view. In H. Wainer & S. Messick (Eds.). *Principals of modern psychological measurement* (pp. 345-356). Hillsdale, NJ: Erlbaum.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In: C. W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison: University of Wisconsin Press.

Brähler. E. & Scheer, J. W. (1983). *Der Gießener Beschwerdebogen (GBB).* Bern: Huber.

Collins, L. & Horn, J. (Eds.) (1992). *Best methods for the analysis of change.* Washington, D.C.: APA.

Cronbach, L. J., & Furby, L. (1970). How we should measure “change“ − Or should we? *Psychological Bulletin, 74,* 68-80.

Deinzer, R., Steyer, R., Ostendorf, F., Neubauer, A., Eid, M., Notz, P. & Schwenkmezger, P. (1995). Situational effects in trait assessment: The FPI, NEOFFI, and EPI questionnaires. *European Journal of Personality, 9,* 1-23.

Dubois, P. H. (1957). *Multivariate correlational analysis.* New York.

Foerster, F. (1995). On the problems of initial-value-dependencies and measurement of change. *Journal of Psychophysiology, 9,* 324-341.

Grawe, K. & Braun, U. (1994). Qualitätskontrolle in der Psychotherapiepraxis. *Zeitschrift für Klinische Psychologie, 23,* 242-267.

Grawe, K., Donati, R. & Bernauer, F. (1994). *Psychotherapie im Wandel. Von der Konfession zur Profession.* Göttingen: Hogrefe.

Harris, C. W. (Ed.) (1963). *Problems in measuring change.* Madison: University of Wisconsin Press.

Huber, H. P. (1973). *Psychometrische Einzelfalldiagnostik.* Weinheim: Beltz.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Jöreskog, K. & Sörbom, D. (1993). *LISREL 8. A guide to the program and its applications.* Chicago, IL: SPSS Inc.

Krauth, J. (1983). Bewertung der Änderungssensitivität von Items. *Zeitschrift für Differentielle und Diagnostische Psychologie, 4,* 7-28.

Lord, F. M. (1963). Elementary models for measuring change. In: C. W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison: University of Wisconsin Press.

Manning, W. H. & Dubois, P. H. (1962). Correlational methods in research on human learning. *Perceptual and Motor Skills, 15,* 287-321.

Petermann, F. (1978). *Veränderungsmessung.* Stuttgart: Kohlhammer.

Smith, M. L. & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology, 60,* 402-408.

Steyer, R. & Eid, M. (1993). *Messen und Testen.* Berlin: Springer.

Steyer, R., Ferring, D. & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 2,* 79 - 98.

Steyer, R. & Schmitt, M. (1990). The effects of aggregation across and within occasions on consistency, specificity, and reliability. *Methodika, 4,* 58-94.

Steyer, R., Eid, M. & Schwenkmezger, P. (in Druck). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research-online.*

von Eye, A. (Ed.) (1990a). *Statistical methods in longitudinal research. Volume I: Principles and structuring change.* San Diego: Academic Press.

von Eye, A. (Ed.) (1990b). *Statistical methods in longitudinal research. Volume II: Time series and categorical data.* San Diego: Academic Press.

Webster, H. & Bereiter, C. (1963). The reliability of changes measured by mental test scores. In C. W. Harris (Ed.) *Problems in measuring change* (pp. 39-59). Madison: University of Wisconsin Press.

Westmeyer, H. (1983). Änderungssensitive Meßinstrumente in der Gesprächspsychotherapie: theorienah und konstruktvalide? − Eine Analyse am Beispiel der Kieler Änderungssensitiven Symptomliste −. *Diagnostica, 28,* 367-376.

Willet, J. B. (1989). Questions and answers in the measurement of change. *Review of Research in Education, 50,* 345-422

Zielke, M. (1979). Entwicklung der Kieler Änderungssensitiven Symptomliste (KASSL). *Diagnostica, 25,* 78-97.

Zielke, M. & Kopf−Mehnert, C. (1978). *Veränderungsfragebogen des Erlebens und Verhaltens, VEV.* Weinheim: Beltz.

Anhang: Ableitung der Gleichung 3

Daß der Zähler der Gleichung 3 richtig ist, erkennt man wie folgt: Allgemein gelten für die Koeffizienten einer linearen Regression: $\alpha_0 = E(Y) - \alpha_1 E(X)$ bzw. $\alpha_1 = Cov(X,Y) / Var(X)$ und aus der Nullhypothese, daß X und Y parallel sind, folgen: $E_0(Y) = E(X)$, $Var_0(Y) = Var(X)$, sowie $\alpha_1 = Cov_0(X,Y) / [Std(X)\, Std_0(Y)] = Kor_0(X,Y) = Rel(X)$. Mit dem tiefgestellten Index 0 soll hervorgehoben werden, daß es sich hier nicht um den tatsächlichen Erwartungswert bzw. die tatsächliche Varianz des Posttests bzw. die tatsächliche Kovarianz oder Korrelation zwischen Prä- und Posttest handelt, sondern um die entsprechenden, *unter der Nullhypothese geltenden* Parameter.

Daß auch der Nenner richtig ist, folgt aus:

$$\begin{aligned} Var_0[Y - E_0(Y \mid X)] &= Var_0(Y) + Var_0[E_0(Y \mid X)] - 2\, Cov_0[Y, E_0(Y \mid X)] \\ &= Var(X) + Rel(X)^2\, Var(X) - 2\, Cov_0[Y, (\alpha_0 + \alpha_1 X)] \\ &= Var(X) + Rel(X)^2\, Var(X) - 2\, Rel(X)\, Cov_0(Y, X) \\ &= Var(X) + Rel(X)^2\, Var(X) - 2\, Rel(X)^2\, Var(X) \\ &= Var(X) - Rel(X)^2\, Var(X) \\ &= Var(X)\, [1 - Rel(X)^2]. \end{aligned}$$

Autorenhinweis

Korrespondenz bzgl. dieses Artikels sind zu richten an: Prof. Dr. Rolf Steyer, Friedrich-Schiller-Universität Jena, Institut für Psychologie, Am Steiger 3, Haus 1, D-07743 Jena
E-mail: s6stro@rz.uni-jena.de