Ambulantes Assessment von Befinden, Stimmungen, Emotionen – Zur Methodik von Selbsteinstufungen (Selbstberichten)

Jochen Fahrenberg (Freiburg i. Br.)

Stand Juli 2006

Gliederung

1 Absichten dieser Darstellung
2 Aufgabenstellung des Assessment und Assessmentstrategien
3 Konstruktion und Evaluation von Methoden zur Beschreibung der Befindlichkeit
Itempool
Auswahlstrategien für Items
Intra-individuelle Varianz statt inter-individueller Varianz
Kovarianzzerlegung
Unipolare oder bipolare Items?
Formulierung von Adjektivskalen und Items
Anzahl der Stufen, Graduierung
Skala mit oder ohne Mittelpunkt?
Gestaltung der Skalen für das Display eines hand-held Computers
Skalenqualität und Skalierung
Verteilungsform (Schiefe und Kurtosis)
Direkte oder indirekte Veränderungsskalierung?
Veränderungsmessung
Ausgangswert-Abhängigkeiten
Änderungssensitivität
Zeitraster
Autokorrelation
Trends
Momentane und rückblickende Einstufungen
Post-Monitoring-Fragebogen und Interview
Retrospektionseffekte
Statistische Definition der intra-individuellen Variabilität
Psychometrische Zuverlässigkeit
Zufallskritische Bewertung von individuellen Zustandsänderungen
Problematischer Gebrauch der konventionellen Itemanalyse und Faktorenanalyse
Fehlbewertungen faktorenanalytischer Resultate
Circumplex-Darstellung
Fortbestehende semantische Probleme und Alternativen?
4 Auswahlgesichtspunkte und Übersicht
5 Einstufung der Befindlichkeit – Einzelne Items oder Skalen wie AD-ACL und PANAS?
6 Statistische Konzepte

Anmerkung 1: Allgemeine Assessment-Modelle
Anmerkung 2: Historische Notiz (Ambulante Selbstbeobachtung durch H. Münsterberg, 1908)

Für die kritische Durchsicht des Manuskripts und für die anregenden Kommentare danke ich Ulrich W. Ebner-Priemer, Mannheim, Thomas Kubiak, Greifswald, Thomas Prill, Mannheim, und Peter Wilhelm, Freiburg (Schweiz)

1 Absichten dieser Darstellung

Befinden, Emotionen und Symptome (Beschwerden) in ihren alltäglichen Verläufen zu erfassen, ist ein zentrales Anliegen des ambulanten Assessment. Das computer-unterstützte Verfahren ist hier aus mehreren Gründen die Methode der Wahl. Außer der ökologischen Validität sind die besonderen, adaptiven Möglichkeiten dieser programmierten Datenerhebung, die Erfassung von Zeitpunkt und Kontext der Selbstberichte und eine hohe kontrollierte Compliance zu nennen (siehe Positionspapier, Fahrenberg, Myrtek, Pawlik & Perrez, 2007).

In testmethodischer Hinsicht entsprechen die Prinzipien und die Methodenprobleme weitgehend denen der schriftlichen Selbsteinstufungen (Skalen), die ja seit mehr als fünfzig Jahren breit angewendet werden. Da die heutigen Untersuchungsansätze des ambulanten Assessment häufig aus anderen Fach-Richtungen als der Psychologischen Diagnostik/Testmethodik stammen, kann es nützlich sein, an Prinzipien der psychologischen Test-Theorie zu erinnern und einige Methodenprobleme hervorzuheben.

Zu den Strategien und Methodenproblemen gehören u.a.

- die Präzisierung der Assessmentstrategie hinsichtlich des angezielten Konstrukts im Varianzraum von Personen, Settings/Situationen, Merkmalen, Zeitpunkten (Wiederholungen) und ggf. Kriterieninformationen;
- Itemselektion mit den wichtigen formalen Anforderungen, u.a. hinsichtlich Verteilungsform, Varianz innerhalb und zwischen Personen;
- die testmethodischen Besonderheiten, wenn statt stabiler Eigenschaften die intra-individuelle Variabilität erfasst werden soll (Veränderungsmessung);
- Gütekriterien und die Abwägung zwischen innerer Konsistenz (Reliabilität), Konstrukt- und Kriterien-Validität, Testökonomie und Standardisierung;
- die Frage nach den Vorteilen und Nachteilen bei der Verwendung von einzelnen Items oder von Skalenwerten, die durch Aggregaten von Itemwerten gebildet werden;
- Überlegungen zum Skalenniveau und zur Auswahl geeigneter statistischer Analysenkonzepte.

Die gegenwärtigen Lehrbücher der Testtheorie und Testkonstruktion behandeln solche Methodenfragen ganz überwiegend im Hinblick auf (1) Fähigkeitstests, wo das Konzept der Parallelmessungen sinnvoll ist oder auf (2) Persönlichkeitsfragebogen verschiedenster Art, wo es ebenfalls um relativ überdauernde (stabile) Eigenschaften geht. Die Diagnostik von Zustandsänderungen mit ihren speziellen Aspekten oder die Prinzipien der allgemeinen Assessmenttheorie werden in der Regel kaum behandelt. Die gegenwärtigen Lehrbücher in Deutschland und in den USA (und wahrscheinlich auch der akademische Unterricht) sind noch weit davon entfernt, die Besonderheiten des ambulanten Assessment zu berücksichtigen.

In der folgenden Übersicht werden einige Aspekte dieser Methodologie geschildert, doch ist keine Übersicht über die einzelnen Verfahren bzw. die Software beabsichtigt. Die Evaluationen und Schlussfolgerungen sind thesenartig formuliert und können an dieser Stelle nicht sehr ausführlich begründet werden. Einige der im Beitrag angesprochenen Methodenprobleme sind an anderer Stelle mit entsprechenden Literaturhinweisen weiter ausgeführt (Fahrenberg, Leonhart & Foerster, 2002b).

Zur Illustration einzelner Aspekte werden statistische Ergebnisse der acht Freiburger MONITOR-Studien herangezogen. Dazu gehört auch die Wochenstudie: Selbstberichte von 33 Studierenden anhand von 8 aktuellen und 3 retrospektiven Items (sowie einem Reaktions-Test) 6 mal täglich an 7 Tagen (Beginn in drei Untergruppen um je zwei Tage versetzt), außerdem mit Tages- und Wochen-Rückblick (siehe Fahrenberg et al., 2002a; Fahrenberg, Leonhard & Foerster, 2002b; und die Dokumentation
http://www.jochen-fahrenberg.de/uploads/media/Freiburger_Tageslaufstudien_mit_MONITOR.pdf

Dieser Beitrag möchte den Erfahrungsaustausch und eine weiterführende Diskussion anregen.

2 Aufgabenstellung des Assessment und Assessmentstrategien

Die Präsenz und die Veränderung von subjektiven Zuständen soll durch Selbsteinstufungen der Untersuchungsteilnehmer erfasst werden. Typische Inhalte sind: Befindlichkeit, Stimmungen, Emotionen – oder spezieller: Schmerzen, erlebte Belastung-Beanspruchung-Überforderung ("Stress"), psychische und körperliche Symptome (Beschwerden). Darüber hinaus kommen beim ambulanten Assessment in der Regel weitere Datenklassen hinzu, u.a. Informationen über Setting/Situation als Kontext, die raumzeitliche Adresse des Selbstberichtes sowie Berichte über Tätigkeiten, soziale Aspekte, Verhaltensweisen u.a. Daten (siehe Fahrenberg et al., 2002b).
Kontext: die Gesamtheit der *relevanten* Bedingungen (Rahmenbedingungen des Geschehens) mit bestimmten kontextuellen Variablen einschließlich der ambienten Parameter der Umwelt, z. B. Lärmpegel, Helligkeit, Temperatur. Begrifflich können Setting, Behavior Setting und Situation unterschieden werden: Setting: ein primär räumlich und durch Gegenstände und Anordnungen objektiv beschreibbarer Kontext (Aufenthaltsort, Tätigkeiten), z. B. das Setting einer Wohnung. Behavior Setting (Barker): ein Setting mit typischem Verhaltensprogramm ("Aufforderungscharakter"), z. B. ein Hörsaal oder ein Restaurant. Situation: ist wesentlich durch die subjektive, erlebnismäßige Beschreibung eines Settings bestimmt und schließt Intentionen und entsprechende Handlungen ein.

Die Selbstberichte (Protokolle) werden gewöhnlich nach einem vorher festgelegten Erhebungsplan (Stichprobentechnik) wiederholt über den Zeitraum eines oder mehrerer Tage bzw. Wochen erhoben. Durch die Wiederholung der Datenerhebung (repeated measurement design) entsteht grundsätzlich eine serielle Abhängigkeit (Autokorrelation) der Daten, denn es werden mehr oder minder große, gemeinsame Einflussquellen existieren: auf der technischen Seite und durch den Erhebungsplan, durch Änderungen der Compliance oder durch Effekte der Übung, der Habituation und Sättigung (Motivationsverlust) u.a. Diese Effekte stellen zugleich typische "Gefährdungen der internen Validität" (Campbell & Stanley) dar.

Assessment und Assessmentstrategien
Der Begriff Assessment betont die pragmatische Seite einer Datenerhebung in Anwendungsfeldern, so dass eine Präzisierung des Untersuchungsziels dazu gehört: zu welchem praktischen Zweck dient die Datenerhebung, welchen "Mehrwert" haben die gewonnenen Daten gegenüber herkömmlichen Methoden, wenn es um die Vorhersage von Kriterien geht und welcher Entscheidungsnutzen in der Praxis wird behauptet?

Assessmentstrategien (siehe Anmerkung 1) sind Pläne, die festlegen, welches Konstrukt mit welchem Untersuchungs- und Auswertungs-Konzept erfasst werden soll. Welche spezielle Auswahl von Personen, Variablen, Situationen und Terminen wird getroffen und auf welche Einheiten sind die empirischen Aussagen deswegen begrenzt? Die klare Unterscheidung des angezielten theoretischen Konstrukts und der Modi des mehrdimensionalen Datenraums (noch vor der inhaltlichen Spezifizierung des gemeinten Konstrukts) ist wichtig, um bestimmten Schwierigkeiten und möglichen Missverständnissen vorzubeugen. So gab es in der psychologischen und psycho-physiologischen Forschung über Aktivierungsprozesse begriffliche Unklarheiten und konträre Ergebnisse bzw. Interpretationen bis erkannt wurde, dass eine prägnantere Unterscheidung unerlässlich ist – zwischen (1) der allgemeinen biobehavioralen Aktivierungsreaktion (gemittelt über Personen, querschnittlich) und (2) den Aktivierungsprozessen als individuell durchaus differierende Verläufe ohne konsistente Zusammenhänge der einzelnen Aktivierungsparameter (längsschnittlich).

Typisch für das ambulante Assessment sind die Fragen nach inter-individuellen Unterschieden der intra-individuellen Variabilität, z.B. Belastung-Beanspruchung-Überforderung am Arbeitsplatz, klinische Symptomverläufe, Schmerz-Tagebücher. Inwieweit unterscheiden sich bestimmte Personen oder Personengruppen in diesen Zustandsänderungen, und gibt es Zusammenhänge mit Kriterien? Hier sind strategische Unterscheidungen nützlich. Stemmler (1992, 1996, S. 261f) schlug in Anlehnung an Cattells Datenbox eine Klassifikation von neun Assessment-Modellen vor.

Die intra-individuelle Variabilität wurde bereits von W. Stern, später von R.B. Cattell, R. Heiß, D. Fiske u.a. Autoren als ein zentrales Thema der Differentiellen Psychologie/ Persönlichkeitsforschung betont (vgl. die Darstellung von Asendorpf, 1991; und die Übersicht, nur über die amerikani-

sche Literatur, Eid & Diener, 1999). Die empirischen Untersuchungen stützen sich jedoch bis in neuere Zeit (auch in der Forschung bei Eid & Diener, 1999; Scherer et al., 2004; Schmidt-Atzert, 1996; Watson, 1997) ganz überwiegend auf die Tagebuch-Methode bzw. Fragebogen, obwohl seit Jahrzehnten (Pawlik & Buse, 1982) computer-unterstützte Methoden verfügbar sind, dann auch die Experience Sampling Methode ESM mit Uhr und Booklets (Larson & Csikszentmihalyi, 1983; Hektner & Csikszentmihalyi, 2002; Brandstätter, 1983; de Vries, 1992; verspätet ESM auch mit zuverlässigerem handheld Computer, siehe Barrett & Barrett, 2001) sowie das Ecological Momentary Assessment EMA (Stone & Shiffman, 1994).

Methodenstudien haben gezeigt, dass bei Fragebogen, insbesondere bei tagebuchähnlich wiederholter Anwendung, substantielle Verzerrungen auftreten (vgl. Baumann, Thiele & Laireiter, 2003; Lucas & Baird, 2005). Die Compliance ist sehr eingeschränkt, d.h. ein hoher Prozentsatz der Fragebogen bzw. Skalen in Tagebuchform wird erst nachträglich ausgefüllt. Außerdem wurde zunehmend erkannt, dass es ausgeprägte psychologische Retrospektionseffekte geben kann (Käppler et al., 2001; Gorin & Stone, 2001; Stone & Litcher-Kelly, 2005). Da weder der Anteil verspäteter Einträge noch das Ausmaß der retrospektiven Verzerrungen kontrolliert werden können, bestehen grundsätzliche Zweifel hinsichtlich aller Tageslauf- und Längsschnitt-Untersuchungen mit konventioneller Papier-und-Bleistift-Methode.

3 Konstruktion und Evaluation von Methoden zur Beschreibung der Befindlichkeit

Itempool

Die publizierten deutschen "Stimmungsskalen" enthalten bereits eine umfangreiche Zusammenstellung von geeignet erscheinenden Deskriptoren, die den simplen Übersetzungen aus der englischen Sprache in der Regel vorzuziehen sind: siehe u.a. die EWL (Janke & Debus, 1978), die SKAS (Hampel, 1972, 1977) und die BfS (von Zerrsen, 1976).

Zur Übersicht über wichtige Konstruktbereiche von Befindlichkeit ("Stimmung") werden hier die Skalenbezeichnungen genannt (siehe auch Tabelle 1):

EWL: Aktiviertheit, Konzentriertheit, Desaktiviertheit, Müdigkeit, Benommenheit, Extravertiertheit, Introvertiertheit, Selbstsicherheit, Gehobene Stimmung, Erregtheit, Empfindlichkeit, Ärger, Ängstlichkeit, Deprimiertheit, Verträumtheit.

SKAS (SES): Gehobene Stimmungslage, gedrückte Stimmungslage, Mißstimmung, Ausgeglichene Stimmungslage, Trägheit, Müdigkeit.

Die zugehörigen Items haben sich in z.T. sehr umfangreichen Untersuchungen bereits bewährt. Die Daten dieser Skalenkonstruktionen wurden allerdings primär in querschnittlichen Erhebungen gewonnen, wobei meist darauf geachtet wurde, dass die Untersuchten sich in möglichst unterschiedlichen Situationen befanden. Selbst wenn wiederholte Erhebungen stattfanden, erfolgte die Skalenkonstruktion vorwiegend aufgrund der interindividuellen Varianz. Die für Veränderungsmessung grundsätzlich – mindestens zu Kontrollzwecken – zu fordernde Konstruktion auf der Basis der intraindividuellen Varianz fehlte und ist bis in die Gegenwart unüblich.

Der Pool der Items von EWL, SKAS und BfS wird vermutlich die gängigsten deutschen Deskriptoren subjektiver Zustände enthalten. Oft gingen eingehende Diskussionen über deren semantische Aspekte und über statistische Aspekte der Verteilung und Interkorrelation voraus. In diesem Pool fehlen allerdings Körperwahrnehmungen, speziellere Beschwerden, Symptome, psychopathologische Phänomene.

Bereits für Laboruntersuchungen, z.B. in der die psychophysiologischen Aktivierungsforschung mit wiederholten Selbsteinstufungen, waren die aus vielen Items bestehenden Listen zu lang; für das ambulante Assessment gilt dies um so mehr. Statt sich praktisch auf eine oder zwei der typischen Skalen der Papier-und-Bleistift-Tests zu beschränken und wichtige andere Bereiche zu vernachlässigen, entschieden sich hier viele Untersucher dafür, einzelne Items zu verwenden (vgl. die publizierten Listen z.B. Heger, 1990; Käppler, 1994; Pawlik & Buse, 1982; Perrez & Reicherts, 1989; Perrez, Schoebi & Wilhelm, 2000).

Auswahlstrategien für Items

Einige Untersucher werden eher eine relativ breite Beschreibung von Befindensänderungen (auch als Hintergrund bestimmter Ziel-Items) anstreben, andere Untersucher werden sich von vornherein hypothesengeleitet für wenige spezielle Items entscheiden. Wenn über spezielle Ziel-Items (z.B. "Schmerz", "Job-Stress") hinaus Deskriptoren allgemeinerer Befindlichkeits-Änderungen ausgewählt werden sollen, ergeben sich mehrere Methodenprobleme. Diese testmethodischen Fragen werden im Folgenden skizziert, einschließlich der Frage "Einzel-Items oder Skalen wie AD-ACL oder PANAS?".

Intra-individuelle Varianz statt inter-individueller Varianz

Die grundsätzliche Forderung wurde bereits erwähnt: Bei der Veränderungsmessung geht es um die inter-individuellen Unterschiede der intra-individuellen Variabilität und nicht etwa um stabile, reproduzierbare Eigenschaftsdimensionen wie bei Persönlichkeits-Fragebogen. Diese Adäquatheitsfrage wird selten diskutiert oder strategisch umgesetzt. Deshalb muss hervorgehoben werden:

Testmethodisch ist es wichtig auf den relativen Varianzanteil der Komponenten (Personen x Situationen bzw. Wiederholungen versus Personen) zu achten und geeignete Datensätze in dieser Hinsicht zu evaluieren.

Als rationale Selektion wurde ein Auswahlverfahren bezeichnet, Items aufgrund formaler Eigenschaften auszuwählen – neben den inhaltlich begründeten Präferenzen, den Prädiktoren-Kriterien-Beziehungen, der Testökonomie und dem Entscheidungsnutzen.

Differenzierungsleistung von Items u. a. Variablen:

zwischen Personen (als Ausdruck relativ überdauernder Personenunterschiede);

zwischen Untersuchungstagen (als Ausdruck der allen Personen gemeinsamen Tageseffekte und Trends, z. B. durch Gewöhnung u. a.);

zwischen den Tageszeiten (als Ausdruck genereller Verläufe im Tagesprofil, z.B. als zirkadiane Periodik u. a.);

zwischen verschiedenen Settings/Situationen (als Ausdruck situationstypischer Merkmale oder Merkmalsmuster des Befindens und Verhaltens bei allen Personen);

hinsichtlich des relativen Anteils der Wechselwirkung von Person x Tag (Individualität der Tagesprotokolle), der Wechselwirkung Person x Tageszeit (Individualität der Tagesläufe) und der Wechselwirkung Tageszeit x Tag (Unterschiede des Tageslaufs an Wochen-Tagen) sowie des Residuums.

In der psychophysiologischen Aktivierungsforschung (vgl. Fahrenberg & Myrtek, 2005) wurden laborinterne Regeln der Parameterselektion erarbeitet, die u.a. die signifikante Diskrimination zwischen typischen Aufgaben-Situationen und zwischen Personen, das Fehlen extremer Verteilungsanomalien (Schiefe, Kurtosis) und das Fehlen hoher Redundanzen umfassen. Eine Innerhalb - Personen Korrelation $r > 0.70$ wurde als problematisch, noch höhere Koeffizienten $r > .80$ durchweg als Anlass zur Eliminierung eines Parameters aus testökonomischen Gründen genommen.

Diese Prozedur, so muss noch einmal betont werden, ist nicht für eine schematische Anwendung gedacht, sondern als systematisches Raster, um Auswahlentscheidungen zu erleichtern. Auch ungünstige Verteilungen werden einen Untersucher kaum dazu veranlassen, einen für die Fragestellung besonders wichtigen Parameter von der weiteren Analyse auszuschließen. Als Beispiel kann das Item "ärgerlich, gereizt" dienen. Trotz des unübersehbaren "Bodeneffektes" in solchen Studien (s. Fahrenberg et al., 2002b) wurde es beibehalten, denn wenn diese seltenen Ärgererlebnisse angegeben werden, hat dies eine besondere psychologische Bedeutung.

Nur deskriptiv, d.h. ohne Anspruch auf statistische Hypothesenprüfung, können zwei-faktorielle oder drei-faktorielle Varianzanalysen als Messwiederholungsmodelle gerechnet werden, um auf einfache Weise die relative Größe der verschiedenen Varianzquellen (Haupteffekte und Wechselwirkungen) zu bestimmten und im Hinblick auf die zur Auswahl stehenden Items/Skalen zu vergleichen. Für

den Datensatz der Freiburger Wochenstudie wurden solche Tabellen mitgeteilt (Fahrenberg et al., 2002a, 2002b).

Kovarianzzerlegung

Die Kovarianzzerlegung ist ein aufschlussreiches Verfahren, um die Quellen der Variation und Kovariation in einer mehrdimensionalen Datenbox zu unterscheiden (siehe u. a. Stemmler, 2001; Stemmler & Fahrenberg, 1989). Die Ergebnisse sind für die Interpretation einer Untersuchung nützlich und erleichtern die weitere Versuchsplanung. Die Kovarianzzerlegung folgt denselben Regeln wie die Zerlegung der Quadratsummen SS in einem experimentellen Versuchsplan, abgesehen davon, dass es hier um Varianz-Kovarianzmatrizen und Korrelationsmatrizen geht. Innerhalb-Personen werden die Kovarianzen über alle Personen gepoolt (SAS-Makro CVZ von F. Foerster). Ein Untersuchungsplan mit Personen und Messwiederholungen (hier die wiederholten Eingaben bzw. Messungen an einem Tag) erlaubt die Unterscheidung der Quellen "Zwischen Personen", "Innerhalb Personen" und "Residuum". Falls diese Datenerhebung noch an weiteren Tagen erfolgt, kann mittels drei-faktorieller, deskriptiv angewendeter Varianzanalyse weiter differenziert werden: Personen, Abfragen, Wochentag, Personen x Abfrage, Personen x Wochentag, Abfrage x Wochentag und Residuum (Beispiele vgl. Fahrenberg et al., 2002b). Auch im Verlauf der schwieriger zu handhabenden Multi-Level-Analysen können diese Varianzquellen separiert werden.

Unipolare oder bipolare Items?

In älteren Stimmungsskalen wurden oft bipolare Items verwendet, z.B. fröhlich–traurig, konzentriert – unkonzentriert. Bei anderen Deskriptoren ist es jedoch sprachlich schwierig, passende Gegenpole zu finden, ohne deutlich andere Konnotation einzuführen: gereizt – friedlich, gestresst – entspannt, nervös – ruhig, verärgert – gut gelaunt, usw. Die meisten bipolar formulierten Items könnten größere semantische Probleme aufgeben als ihre unipolaren Abschnitte. Deswegen wurden zunehmend unipolare Items verwendet. Als Konsequenz scheinen sich dann in den Faktorenanalysen bestimmte Bereiche in zwei relativ unabhängige Subskalen aufzuspalten (vgl. EWL-Skalen und "Gehobene Stimmung" und "Gedrückte Stimmung" SKAS oder neuerdings Positive Affect und Negative Affect PANAS).

Unipolare Skalen, z.B. von (1) überhaupt nicht zutreffend bis (7) völlig zutreffend, bringen oft das schwierige Problem der Bodeneffekte bzw. Deckeneffekte mit sich, d.h. schiefe Verteilungen mit einer eingeschränkten Diskriminationsleistung an den betreffenden Skalenende.

Formulierung von Adjektivskalen und Items

Bei der sprachlichen Formulierung sind verschiedene Gesichtspunkte zu bedenken: umgangssprachlich verständlich, semantisch möglichst eindeutig, d. h. unter Vermeidung von Fremdwörtern, von komplizierter Grammatik, doppelter Verneinung oder regional unterschiedlichem Sprachgebrauch. Hier helfen u. U. gründliche Formulierungsversuche und Diskussion in einer Gruppe. Wie kann das Gemeinte am besten ausgedrückt werden? Häufig gibt es Diskussionen um die Graduierungen, d. h. die Anzahl der Stufen und die Benennung bzw. Verankerung der Stufen.

Anzahl der Stufen, Graduierung

Die Meinungen sind geteilt, wie viele Stufen sinnvoll sind. Die Antwort hängt von dem Iteminhalt, von der Formulierung (unipolar, gutgelaunt ... nicht gut gelaunt, bipolar gut gelaunt ... schlecht gelaunt), von der Population u. a. Aspekten ab. Bei Studierenden kann durchaus auch an mehr als 7 Stufen, d.h. auch 11 Stufen oder, je nach Merkmal, sogar bis zu 21 Stufen, gedacht werden. Eine "Visual Analog Scale" VAS ist praktisch feiner abzustufen als eine typische Likert-Skala.

In der Freiburger MONITOR-Software werden verwendet:

(1) unipolare, 7 stufige Adjektivskalen in Gestalt von miteinander verbundenen Kästchen, wobei nur die Extrema verbal benannt sind:
"Ist die momentane Situation ..." oder "Fühlen Sie sich momentan ..." mit den Skalenstufen 1 = gar nicht ... bis ... Skalenstufe 7 = völlig;

(2) eine visuelle Analog-Skala mit einer verbalen Verankerung der Skalenenden (gar nicht – völlig), wobei die Cursor-Position in 21 (unsichtbare) Stufen aufgelöst wird. Diese VAS könnte Tendenzen zur Mitte etwas verringern und insgesamt eine höhere Varianz begünstigen.

Zu diesem Aspekt ist ein indirekter Vergleich zwischen zwei Untersuchungen möglich, in denen dieselben Items, jedoch verschiedene Skalenformate verwendet wurden. Die VAS-Mittelwerte und SD der 6 Selbsteinstufungen eines Tages liegen von 64 Studierenden vor (n = 367 Datenpunkte), die entsprechenden Statistiken bei Verwendung von 7-stufigen Skalen stammen aus der Wochenstudie 33 Studierenden (n = 1323 Datenpunkte) und lauten z.B. für die folgenden vier Items:

Tabelle 1: Anzahl der Skalenstufen und Verteilung der Selbsteinstufungen

	VAS mit 21 Stufen				Likert-Item mit 7 Stufen			
	M	SD	Schiefe	Kurtosis	M	SD	Schiefe	Kurtosis
aktiv/leistungsfähig	11.50	5.01	-.31	-.85	3.86	1.43	-.17	-.65
ärgerlich/gereizt	2.85	4.12	**1.79**	**2.78**	1.96	1.36	**1.50**	**1.57**
bedrückt	3.75	4.66	1.30	0.70	2.38	1.48	.90	-.17
körperlich wohl	13.0	4.63	-.66	-.09	4.64	1.44	-.42	-.17

Wesentlich ist, dass die VAS deutlich größere Varianzen, d.h. das Doppelte bis Dreifache, ergibt, also eine sehr erwünschte höhere Diskriminationsleistung zwischen einzelnen Bedingungen. Beim Item "ärgerlich, gereizt" sind jedoch die Schiefe und Kurtosis der Verteilung tendenziell noch stärker ausgeprägt als bei der kurzen Skala. Eine Methodenstudie mit beiden Skalentypen an derselben Stichprobe würde einen zuverlässigeren Vergleich ermöglichen.

Hinsichtlich der Quantoren gibt es eine Anzahl von methodischen und empirischen Untersuchungen. Rohrmann (1978) untersuchte das populäre Verständnis solcher Graduierungen und plädierte für die folgenden Abstufungen, gab jedoch auch Hinweise für feinere Graduierungen:
Häufigkeitsskala: nie – selten – gelegentlich – oft – immer
Intensitätsskala: (gar) nicht – wenig – mittelmäßig – überwiegend (oder ziemlich, annähernd) – völlig
Wahrscheinlichkeitsskala: keinesfalls – wahrscheinlich nicht – vielleicht – ziemlich wahrscheinlich – sicher
Bewertung von Aussagen: stimmt nicht – wenig – mittelmäßig – ziemlich – sehr zutreffend
oder: gar nicht – wenig – teils-teils – ziemlich – völlig zutreffend

Skala mit oder ohne Mittelpunkt?
Sowohl die Skala mit ungerader Anzahl von Stufen als auch die Skala mit gerader Anzahl haben Argumente für sich. Viele Personen möchten eine mittlere Position zur Verfügung haben, doch bleibt unklar, ob die betreffende Antwort "weder – noch" oder nur "unentschieden / kann es nicht sagen" bedeutet. Fehlt eine mittlere Stufe, werden die Befragten zur Entscheidung gezwungen und könnten damit unzufrieden sein. Die Formulierung der Skala legt bereits Urteilsheuristiken nahe (Schwarz, 1990; Schwarz & Scheuring, 1992), z.B. könnte die mittlere Skalenstufe als "normaler" Wert der Antwort interpretiert werden.

Diese Überlegungen beziehen sich nicht allein auf bipolare Skalen. Eine mittlere Position muss vorhanden sein, wenn in direkten Veränderungsskalierungen nach der Zunahme oder Abnahme der Merkmalsausprägung gefragt wird und deswegen auch eine Position "unverändert" existieren muss.

Gestaltung der Skalen für das Display eines hand-held Computers
Für die kleinen Displays von hand-held Computern ergeben sich typische Layout-Probleme, aber auch einige originelle Lösungsversuche (siehe die speziellen Anwender-Publikationen, z.B. die ausführlich dokumentierte MONITOR-Methodik (Hüttner & Leonhart, 2002; vgl. auch http://www4.psychologie.uni-freiburg.de/einrichtungen/Psychophysiologie/ (Ambulatory Assessment – MONITOR 9) und Beiträge zum Design von User Interfaces, Palmblad & Tiplady, 2004).

Skalenqualität und Skalierung
Aussagen über subjektive Zustände werden erhoben u.a.:
(1) kategorial (vorhanden/nicht vorhanden, oder im Stil einer Multiple-Choice-Liste), d.h. auf einer Nominal-Skala;

(2) mit abgestuften Antwort-Möglichkeiten: numerisch ("Thermometer-Skala"), graphisch (visuelle Analogskala VAS) oder mit verbal benannten Stufen, also eine Größer-Kleiner-Beziehung (Rangordnung, auf einer Ordinal-Skala) wiedergebend;
(3) durch Kombination einzelner Items aufgrund einer statistischen Itemanalyse zu einer Skala, oft als sog. Likert-Skala organisiert, d.h. aus einer Anzahl von Items, mit jeweils quantitativ abgestuften Antwort-Möglichkeiten, aufaddiert und insofern eine Intervall-Skala postulierend;
(4) als direkte Einschätzung der Intensität (numerisch oder graphisch), z.B. als x Prozent des erlebten Maximums oder als Stufe x eines als 100 gesetzten Maximums, z.B. auch des vom Befragten bisher erlebten Maximums (eine ipsative, d.h. personeneigene Skalierung, die über die ipsative Normierung hinausgeht);
(5) viel seltener mittels der Thurstone-Skala, für die durch bestimmte psychometrische Techniken, z.B. durch den systematischen Vergleich aller Deskriptoren (Methode des Paarvergleichs), Distanzen zwischen den benannten Stufen berechnet werden, denen dann für die betreffende Untersuchungsgruppe Intervallskalen-Charakter zugesprochen werden kann, z.B. die Skala Allgemeiner zentraler Aktiviertheit (Bartenwerfer, 1963);
(5) gelegentlich nach der von Guttman vorgeschlagenen Circumplex-Skalierung, d.h. mit zyklisch angeordneten Deskriptoren, wobei die Richtung des einzutragenden Vektors eine zweidimensionale Konzeption, z.B. Valenz und Aktivierung, und die Länge des Vektors die Intensität des Zustandes repräsentieren soll (weitere Literatur zu Messung und Skalierung, siehe Borg & Staufenbiel, 1997; Orth, 1983).

Typische Ordinaldaten werden gewonnen, wenn ein einzelner kompetenter Beobachter oder eine Gruppe trainierter Beobachter die Ausprägung von Merkmalen beurteilen und ihre Einschätzungen in Rangordnungen von Größer-Kleiner-Beziehungen ausdrücken. Demgegenüber bestehen Erlebnisberichte aus Selbstbeurteilungen. Weder ist ein direkter Vergleich mit der Befindlichkeit anderer Menschen möglich noch besteht in der Regel ein methodisches Training. Ob die Einstufungen faktisch wiederholbar sind oder ob eine Beurteiler-Übereinstimmung besteht, kann grundsätzlich nicht geprüft werden. Selbsteinstufungen des Befindens liefern also nominale oder ordinale Daten besonderer Art. Es sind subjektive Schätzverfahren hinsichtlich nicht direkt messbarer, subjektiver Befindlichkeiten, d.h. von Merkmalen mit unbekanntem numerischen Relativ, vermutlich von Individuum zu Individuum unterschiedlich, und eventuell auch von Deskriptor zu Deskriptor und von Situation zu Situation mit wechselnden, pseudo-numerischen Bezugssystemen.

Wer die Definitionen einer Intervallskala kennt, wird grundsätzlich zweifeln, dass es sich bei den heute gängigen Likert-Skalen zur Selbsteinstufung subjektiver Zustände um Intervallskalen handeln kann: Die Gleichheit der Skalenintervalle ist nicht gegeben und folglich sind die Verhältnisse der Intervalle nicht definiert. Deshalb sind – streng genommen – lineare Transformationen und die typischen "metrischen" Rechenoperationen definitionsgemäß unzulässig; auch die simple Addition einzelner, heterogener Itemwerte zu einem Skalenwert verletzt die Grundannahme. Es kann nur, Item für Item, eine Rangordnung der Einstufung auf einem individuellen Kontinuum mit Größer-Kleiner-Beziehung behauptet werden. Über die Konsequenzen dieses Sachverhalts existieren allerdings in der Fachliteratur große Meinungsunterschiede.
In der psychologischen Testmethodik und Forschung ist es eine weit verbreitete Gewohnheit, auch diesen konventionellen, als numerisch erscheinenden Selbsteinstufungs-Daten den Charakter von Intervall-Skalen zuzubilligen. Die großen Vorteile liegen in der Anwendung der parametrischen statistischen Tests, deren Rechenoperationen Intervallskalen voraussetzen, und in der anscheinend besseren "Informationsausschöpfung". Gelegentlich wird auch betont, dass statistische Auswertungen wie Varianzanalysen relativ robust gegenüber der Verletzung ihrer Voraussetzungen wären oder dass es durchaus zulässig sei, auch parametrische statistische Tests sozusagen als deskriptiv-explorative Strategien aufzufassen und erst bei fortgeschrittenen Entscheidungsexperimenten strenger zu werden. Früher waren auch weitaus weniger statistische Konzepte zur Auswertung von Ordinal- und Kategorial-Daten ausgearbeitet (siehe Bortz, Lienert & Boehnke, 2000).

Die Interpretation von Ordinalskalen-Daten als Intervall-Messungen ist besonders erstaunlich, wenn sehr anspruchsvolle statistische Strukturanalysen und Modellierungen gerade anhand metrisch

sehr zweifelhafter Selbstbeurteilungen in Fragebogen unternommen werden – unter Einschluss aller zusätzlichen retrospektiven Verzerrungen u.a. Mängel.

Die "liberale" Einstellung hat wohl den Hintergrund, dass in solche messmethodischen und theoretischen Vorentscheidungen zur Operationalisierung/Abbildung/Repräsentation latenter Eigenschaften so viele Voraussetzungen eingehen, dass die Argumentation unübersichtlich wird. Dass diese Entscheidungen beliebig wären, ist auch dann nicht anzunehmen, wenn vorsichtig formuliert wird: "Die Skalenqualität einer Messung ist also letztlich von theoretischen Entscheidungen, d.h. von Interpretationen abhängig." (Bortz et al., 2000, S. 66)

Es gibt jedoch weiterhin Gutachter internationaler Journals, die darauf bestehen, dass die statistische Prüfung von Hypothesen Rücksicht auf das Skalenniveau nimmt und Ordinaldaten auch mit statistischen Tests für Ordinaldaten ausgewertet werden müssen. Ein pragmatischer Weg kann es u.U. sein, zentrale Untersuchungshypothesen sowohl parametrisch als auch nicht-parametrisch zu prüfen und eventuelle Widersprüche vertiefend zu diskutieren.

Verteilungsform (Schiefe und Kurtosis)

Auffällige Verteilungen, d.h. extreme Schiefe oder Kurtosis, sind unerwünschte Sachverhalte, denn sie zeigen an, dass in bestimmten Bereichen der Skala zwischen verschiedenen Personen/ Zuständen/ Situationen kaum differenziert werden kann. Gelegentlich wird übersehen, dass es nicht auf die sog. "Normalität" (Gauß' Fehlerverteilung) der Verteilung an sich, sondern (1) auf die maximale Differenzierung zwischen Personen/Zuständen/Situationen und (2) im bivariaten Fall auf die Homogenität der Zeilen- und Spaltenvarianzen (bivariate Normalverteilung) ankommt.

Durch eine rechnerische Transformation (z-Werte, logarithmisch u.a.) können Verteilungen solcher Skalenwerte "normalisiert" werden, doch ist der Nutzen dieses Verfahrens umstritten. Die empirische Diskriminationsleistung ist nachträglich auf diese Weise nicht zu verbessern, sondern höchstens durch eine konstruktive Revision der Skala. Zumindest im Bereich der Selbsteinstufungen beziehen sich Datentransformationen nur auf die Oberfläche – im Vergleich zu den viel ernsteren Problemen der Semantik solcher Urteile und der Ordinal-Intervall-Skalierung.

Direkte oder indirekte Veränderungsskalierung?

Veränderungen der Befindlichkeit können auf zweierlei Weise bestimmt werden: indem bei einer zweiten Einstufung die numerische Differenz von der ersten zur zweiten Einstufung gebildet wird oder indem eine direkte Einstufung von Richtung und Betrag der Zustandsänderung erbeten wird. Darüber hinaus könnte der Verlauf, z.B. als Wachstumsfunktion oder als Polynom modelliert werden, um die Änderungen darauf zu beziehen.

Methodenstudien mit Fragebogendaten sprechen dafür, dass die indirekte und die direkte Veränderungsmessung nicht äquivalent sind (Baumann, Sodemann & Tobien, 1980; Rösler, Baumann & Marake, 1980; Schneider, 1982).

In einer Methodenstudie anhand der EWL wurde von Baltisssen und Boucsein (1987) zweierlei verglichen: (1) Zustandsskalierungen mittels indirekter Einstufungen auf 7stufigen Skalen (0 bis 6) mit anschließender Berechnung der Differenzen (13 Werte möglich, nachträglich reduziert auf 7 Stufen, von – 3 bis + 3) mit (2) direkte Veränderungsskalierungen auf einer 7stufigen Skala von – 3 bis + 3. In einem experimentellem Versuchsplan wurde bei der Experimentalgruppe durch Sprechangst bzw. Lärm eine psychophysische Aktivierung induziert. Evaluiert wurden die Korrelation beider Veränderungswerte sowie die Anzahl hypothesenkonformer Effekte in den psychophysiologisch induzierten Reaktionen. Die Schlussfolgerungen lauteten: "Die Interkorrelationen der Skalierungsverfahren machten bei einem gemeinsamen Varianzanteil von ca. 40 % deutlich, dass die Verfahren eigenständige Methoden im Sinne verschiedener Beobachtungsebenen darstellen." ... "Es ergaben sich Hinweise auf eine differentielle Validität derart, dass Stresseffekte im Bereich körperlicher Symptome besser durch die Veränderungsskalierung, im Bereich der Befindlichkeit eindeutiger durch die Zustandsskalierung abgebildet werden." (S. 1). Die direkte Veränderungsskalierung zeigte tendenziell eher hypothesenkonforme Beziehungen zur gemessenen Herzfrequenz und EDA. Insgesamt gaben die Autoren eine vorsichtige Empfehlung für die direkte Veränderungsskalierung.

In der psychophysiologischen Forschung wurden nicht selten sowohl Zustands- als Veränderungs-Skalierungen von subjektivem Befinden und Körperwahrnehmungen eingesetzt (vgl. Fahrenberg & Myrtek, 2005; Schneider, 1982) und weitere psychometrische Aspekte berücksichtigt. Eine

deutliche Überlegenheit einer der beiden Verfahren war nicht festzustellen – die psychophysiologischen Interkorrelationen blieben in allen Fällen geringfügig.

Die Abbildung von Zustandsänderungen durch indirekte oder direkte Einstufungen sind nicht äquivalent. Dieses Methodenproblem scheint auch in neueren Arbeiten keine zu verallgemeinernde Antwort gefunden zu haben. Statt auf überlegene Modellierungsansätze zu hoffen, muss hier eher ein grundsätzliches Defizit in der Numerik der subjektiven Einstufungen angenommen werden. Bei unreflektierter Verwendung nur einer der Methoden besteht ein u.U. nicht unwichtiger Bias.

Eine Entscheidung, welche der beiden Methoden einen höheren deskriptiven Nutzen hat (in Begriffen von empirischer Validität, Generalisierbarkeit, Entscheidungsnutzen), ist kaum möglich, sofern keine geeigneten Untersuchungen mit überzeugenden Kriterien vorliegen.

Veränderungsmessung

Die Messung einer Veränderung, d. h. der Zunahme oder der Abnahme eines quantitativ erfassten Merkmals, scheint zunächst eine einfache Aufgabe zu sein. Biometrisch ist jedoch dieses "measurement of change" eine komplizierte Angelegenheit. Bei der Definition von geeigneten Veränderungsmaßen (Reaktionswerten) müssen funktionelle, statistische (sog. Fehlermodelle) und rechnerische Abhängigkeiten bedacht werden. In einigen Funktionsbereichen bestehen Ausgangswert-Beziehungen. Diese Abhängigkeiten erschweren die Effektbeurteilung. Zum Beispiel: Sind Hypertoniker blutdruckreaktiver oder reagieren sie nur auf einem höheren Niveau? (Fahrenberg, Foerster & Wilmers, 1995). Veränderungsmessungen sind jedoch für viele psychologische Fragestellungen und die Prozessforschung notwendig (weitere Literatur siehe u.a. Baumann, Fähndrich, Stieglitz & Woggon, 1990; Fahrenberg et al., 2002b). Für die konventionelle psychologische Testdiagnostik von Eigenschaftsausprägungen stellt sich diese Aufgabe in der Regel einfacher dar, denn für die Schätzungen werden die konventionellen Reliabilitäts-Bestimmungen (Retest-Reliabilität, Konsistenz-Reliabilität) verwendet, die in der Prozessforschung inadäquat sein können.

Ausgangswert-Abhängigkeiten

Ein ebenfalls nicht befriedigend gelöstes Methodenproblem ist die Ausgangswert-Abhängigkeit von Veränderungswerten, die bei einer bias-freien Evaluation zu berücksichtigen ist. Veränderungswerte, die als einfache Differenzen gebildet werden, sind deswegen von den Ausgangswerten rechnerisch abhängig, und mit diesen in der Regel negativ korreliert. Ob darüber hinaus eine homöostatisch begründete, intra-individuelle Ausgangswert-Beziehung existiert (je höher der Ausgangswert, desto geringer der biologisch mögliche Anstieg) ist umstritten, denn außerdem gibt es den Effekt der Regression zur Mitte (geteilte Messfehler). Eine Differenzierung verlangt explizite Fehlermodelle (siehe Kendall and Stuart) aufgrund adäquater Schätzung der Reliabilitäten. Es werden sog. "wahre Werte" unter Berücksichtigung der Reliabilität aufgrund einer wiederholten Messung des Ausgangswertes bzw. des Reaktionswertes bestimmt (Stemmler, 2001). Umfangreiche Methodenstudien lieferten viele Belege für AW-Abhängigkeiten (Foerster, 1995), doch blieben die Resultate für einzelne Funktionen relativ inkonsistent, ohne dass sich die Gründe genau angeben lassen (Fahrenberg & Myrtek, 2005). Das Interesse an diesem ungelösten Problem hat stark abgenommen – zugunsten von Ausgangswert-adjustierten (residualisierten) Veränderungswerten, bei denen unterschiedslos alle Komponenten solcher AW-Beziehungen pauschal eliminiert werden.

Im Vergleich zu physiologischen Messwerten, sind die zugrundeliegenden AW-Beziehungen bei Einstufungen der Befindlichkeit praktisch vielleicht weniger von Belang: nicht nur wegen der mangelnden Intervall-Metrik, sondern auch wegen der massiven Restriktion des Skalen-Bereichs und der ohnehin unsicheren Beurteilung der Reliabilität. Während z.B. die Herzfrequenz praktisch mit perfekter Zuverlässigkeit gemessen werden kann, fällt es schwer, den "Messfehler" von Selbstbeurteilungen zu bewerten.

Änderungssensitivität

Im Zusammenhang der Veränderungsmessung ist das Konzept der Änderungssensitiviät entstanden (vgl. Krauth, 1983). Ist eine bestimmte Aufgabe oder ist ein Item überhaupt geeignet, Zustandsänderungen zu erfassen, d.h. intra-individuelle Varianz (und nicht nur interindividuelle Unterschiede oder Fehlerkomponenten) anzuzeigen. Ein idealer Parameter würde sowohl intra- als inter-individuelle

Varianz bei minimalem Messfehler abbilden wie es für die Herzfrequenz in psychophysiologischen Untersuchungen bekannt ist.

Das relativ einfache Verfahren der Kovarianzzerlegung oder die Multi-Level-Analysen für metrische Daten können wichtige Anhaltspunkte für die Evaluation der Änderungssensitivität liefern.

Zeitraster

Die Änderungssensitivität der Methodik ist wesentlich begrenzt durch das Zeitraster der Datenerhebung. Wenn die Abstände zwischen den Erlebnisberichten eine oder mehrere Stunden betragen, werden viele der rascher wechselnden Zustandsänderungen bzw. Ereignisse nicht erfasst werden. Selbst bei einem 30-Minuten-Raster im psychophysiologischen Blutdruck-Monitoring (Käppler, 1994) können kurzfristige emotionale Reaktionen und auch plötzliche Blutdruckanstiege "übersehen" werden. Die programmiertechnisch vorgesehene Möglichkeit, bei besonderen Ereignissen eine computerunterstützte Protokollierung selbständig auszulösen, wird erfahrungsgemäss nur selten genutzt. Ein methodisch überlegenes Verfahren ist das interaktive Monitoring, indem der Selbstbericht durch die nicht-metabolisch bedingte Zunahme der Herzfrequenz getriggert wird (Myrtek, 2004). Allgemeine Regeln, wann ein Zeitraster adäquat ist, kann es nur geben, wenn die zeitliche Struktur des zugrundeliegenden Prozesses bekannt ist. In Methodenstudien könnte das Raster systematisch variiert und der Informationsgewinn bzw. der Informationsverlust analysiert werden (vgl. die Diskussion in Fahrenberg et al., 2002; vgl. Beispiele von Ebner-Priemer, 2005). In der konkreten Versuchsplanung werden neben den inhaltlichen Hypothesen und Erwartungen auch triviale Gründe maßgeblich sein, die Datenerhebung zumutbar zu gestalten und an den Tageslauf anzupassen.

Autokorrelation

Die grundsätzlich bestehende serielle Abhängigkeit bei wiederholter Einstufungen der Befindlichkeit bzw. bei anderen Messwiederholungen ist oft betont worden. Die Koeffizienten der Autokorrelation – bzw. im Hinblick auf die Vorhersagbarkeit – die Beta-Koeffizienten der Autoregression sind in den Zeitreihen typischer Verhaltensexperimente bzw. Einzelfall-Studien im allgemeinen gering, und es gibt eine Kontroverse, ob sie nicht vernachlässigt werden können (Huitema, 1988; Matyas & Greenwood, 1991). Aus ihren Monitoringstudien berichteten Pawlik und Buse (1992) Autokorrelationskoeffizienten bei einigen ausgewählte Items des Befindens und Verhaltens. Die Koeffizienten lagen nach einem Intervall von 15 Minuten in der Größenordnung von 0.18 und blieben dann über die folgenden Stunden relativ konstant zwischen 0.20 und 0.30. Für eine Serie von Aufmerksamkeitstests, die im Abstand von 90 Minuten durchgeführt wurden, ergeben sich, sobald eine Bereinigung für den (Übungs-) Trend vorgenommen wird, deutlich geringere Koeffizienten (0.23, 0.08 und 0.04 für lag 1 bis 3) und für Selbsteinstufungen etwas höhere Koeffizienten (0.38, 0.24, 0.17). Bei einer Zeitverschiebung von lag 8, was einem ungefähren Abstand von 24 Stunden entsprach, traten noch einmal höhere Koeffizienten im Sinne einer zirkadianen Periodik auf (Pawlik & Buse, 1999)

Die Höhe der Autokorrelation hängt maßgeblich vom Zeitraster und vom untersuchten Prozess ab. Für das langsamer veränderliche Allgemeinbefinden sind höhere Autokorrelations-Koeffizienten zu erwarten als in Zeitreihen mit dynamischen emotionalen Episoden. Bei einem Zeitraster von ca. 2 Stunden scheinen die Abhängigkeiten nur gering, fast vernachlässigenswert zu sein. In der Freiburger Wochenstudie waren die Koeffizienten der Autokorrelation sowohl bei Befindens-Items als auch für den Reaktions-Test niedrig. Koeffizienten $\geq .20$ zeigen sich bei Lag 1 für aktiv (.21), bedrückt (.20), gute Stimmung (.26) und – am höchsten – körperlich wohl (.31). Bei Lag 2, d.h. nach einem halben Tag bleibt allein der Koeffizient für körperlich wohl (.21) übrig, alle anderen gehen gegen Null. Dies gilt auch bei Lag 6, d.h. nach 24 Stunden: hier sind es noch die Koeffizienten für aktiv (.13) und Mittelwert der Reaktionszeit (.14). Da es in dieser Hinsicht auffällige Unterschiede zwischen den Personen gibt, wurden die Koeffizienten der individuellen Zeitreihen über die z'-Funktion aggregiert. Die durchschnittliche serielle Abhängigkeit der Selbsteinstufungen ist hier also schon beim zweiten Intervall nur noch gering ausgeprägt und ohne praktische Bedeutung.

Die Höhe der Autokorrelation kann durch das Vorkommen markanter Effekte beeinflusst, u.U. inflationiert werden, oder durch Datenlücken reduziert werden. Solche Autokorrelations-Analysen stehen unter dem Vorbehalt, dass die Daten zu (nahezu) äquidistanten Zeitpunkten vorliegen sollen. Da diese Voraussetzung bei psychologischen Zeitreihen, die länger als ein Wach-Intervall dauern,

grundsätzlich nicht gegeben ist, sind die über einen Tag hinausgehenden Berechnungen nur vorsichtig zu vergleichen und zu interpretieren, jedoch z.B. für ein 24-Stunden-Lag nicht uninteressant.

Streudiagramme (Scatter Plots) können auch verwendet werden, um serielle Abhängigkeiten zu veranschaulichen, z.B. wenn die Abweichung des Itemwertes vom individuellen Mittelwert gegen den jeweils vorausgegangenen Wert eingetragen wird (Totterdell, Briner, Parkinson & Reynolds, 1996). Scatter Plots dieser Art können die Autokorrelation mit dem lag 1 repräsentieren. In einer zweiten Graphik können aufeinanderfolgende Veränderungen gegeneinander aufgetragen werden. Dadurch wird das Ausmaß der Veränderung in sukzessiven Intervallen deutlich (siehe die MQSD-Statistik). Beide Darstellungen können Unterschiede zwischen Items aufzeigen: "Fingerabdrücke in Zeitreihen" (Totterdell et al., 1994, 1996). Im Unterschied zu den üblichen Zeitreihenanalysen wird hier nicht vorausgesetzt, dass der Prozess stationär ist, d. h. frei von Trends in Mittelwert und Standardabweichung. Die Darstellung soll an ein besseres Verständnis von dynamischen Verläufen (nicht-stationären, nicht-linearen Entwicklungen) in den Variablen heranführen. So könnten z. B. Änderungen der Stimmung vom aktuellen Niveau der Stimmung oder von einer vorausgegangenen Schlafphase abhängen. Einfach die nicht-stationären Anteile (Trends) zu entfernen, wäre keine überzeugende Lösung.

Anschauliche Beispiele für solche graphischen Darstellungen von Lag1- und Lag2-Darstellungen gibt u.a. Ebner-Priemer (2005) für Patientendaten zur aversiven Anspannung, außerdem wurden MQSD-Statistiken berechnet.

Bei einem höher auflösenden Zeitraster können sich natürlich stärkere Abhängigkeiten zeigen. Dieser Aspekt sollte also geprüft, d.h. weder vernachlässigt noch überschätzt werden. Die Autokorrelationsfunktion (Autoregression) und die Beschreibung der intra-individuellen Variation durch das MQSD-Maß könnten dazu beitragen, unter statistisch-operationaler Perspektive zwischen eher stationären Zuständen (Stimmungen) und deutlich abgehobenen dynamischen Zustandsänderungen (Emotionen) zu differenzieren.

Trends

Montone Trends in der Befindlichkeit gesunder Untersuchungsteilnehmer scheinen selten zu sein. In Zeitreihen psychologischer Testdaten sind jedoch in der Regel deutliche Trends zu erkennen, die durch Übung, Lernen bestimmter Strategien oder Ermüdung/Sättigung verursacht sind (Buse & Pawlik, 1996; 2001; Pawlik & Buse, 1994; siehe auch Fahrenberg et al., 1977). Gemeinsame Trends in Zeitreihen können zu einer Fehleinschätzung der Kovariation führen. Doch die simple Eliminierung von Trendkomponenten kann, weil zugleich andere Varianzkomponenten verloren gehen, Nachteile haben (siehe die Diskussion, Fahrenberg et al., 2002b).

Momentane und rückblickende Einstufungen

Der Vorzug der computer-gestützten Protokollierung von Selbstberichten im ambulanten Assessment liegt gerade in dem Zugang zum momentanen Befinden. Aktuelle Aussagen werden weitaus eher die individuelle Befindlichkeit repräsentieren als die retrospektiven, subjektiv auf unbekannte Weise aggregierten, summarischen Urteile im Fragebogen bzw. Tagebuch. Die computer-unterstützten Selbsteinstufungen sind in viel größerer Dichte der Informationen verfügbar, sie werden mit wesentlich höherer Compliance und technischer Zuverlässigkeit gewonnen als in einem Fragebogen, und diese Auskünfte sind kontextbezogene Daten mit "Adressen", denn sie sind in dem angegebenen Setting mit einer genauen Zeitangabe verankert.

Als Absicherung gegen das Übersehen wichtiger Veränderungen aufgrund des ausgewählten Zeitrasters ist es zweckmäßig, den Items zum momentanen Befinden einige rückblickende Fragen anzufügen, z.B.:

Gab es seit der letzten Eingabe besondere Ereignisse? (bei "Eingabe "Ja" folgt die Aufforderung: Bitte Ereignisse in Stichworten eingeben!)

Wie stark erlebten Sie seit der letzten Eingabe Stress?

Gab es seit der letzten Eingabe positive Veränderungen?

Diese Art der Protokollierung wäre – wie alle anderen Schritte des Programmablaufs – zu Beginn der Untersuchung zu besprechen und zu trainieren.

Post-Monitoring-Fragebogen und Interview

Es bietet sich an, zum Abschluss eines ambulanten Assessment einen Fragebogen einzusetzen, um die Teilnehmer in standardisierter Weise nach Kommentaren und eventuellen Verbesserungsmöglichkeiten zu fragen. Weitere Themen wären Akzeptanz und Compliance, die rückblickende Beurteilung des durchgeführten "Skalierungs-Trainings" zu Untersuchungsbeginn, die Exploration, ob es besondere Ereignisse, eventuell mit Datenschutz-Problemen gab u.a.

Beim psychophysiologischen Monitoring haben sich strukturierte Interviews anhand einer unmittelbar angeschlossenen, ersten Auswertung der Selbstberichte und der Messwerte von Blutdruck und Herzfrequenz bewährt. In komplementärer Weise wird nach Ereignissen gefragt, die aus dem übrigen Tagesablauf herausragten. Dies konnten Erlebnisse mit einer besonderen emotionalen Qualität oder/ und einer besonderen körperlichen Belastung sein. Hatten Patienten solche Episoden beschrieben, wurde geprüft, ob zeitgleiche Blutdruckdaten vorlagen. In einem weiteren Durchgang wurde mit dem Patienten in Stichworten der gesamte Tagesablauf nachvollzogen, um weitere Hinweise für Segmentierungen und Aggregationen zu erhalten. Im dritten Durchgang konnten die genauen Blutdruckwerte als Ausgangsbasis verwendet werden, um nach psychologischen Entsprechungen der physiologischen Episoden (relative Maxima, Minima) zu suchen (siehe Fahrenberg & Myrtek, 2005).

Retrospektionseffekte

Retrospektive Einstufungen von Befinden, Beanspruchung (Stress), Beschwerden, Schmerzen usw. stimmen mit den aktuellen Einstufungen nicht gut überein. Erinnerungstäuschungen "recall-error" können in differentiell-psychologischen und klinisch-psychologischen Untersuchungen, die sich – wie mehrheitlich der Fall – auf Interviews, Fragebogen oder Tagebücher stützen, eine systematische Verzerrung mit problematischen Konsequenzen bedingen. Solche Diskrepanzen zwischen momentanen und rückblickenden Selbsteinstufungen liefern ein starkes Argument für die computer-unterstützte Methode der Selbstberichte.

Inzwischen existieren zahlreiche Untersuchungen über solcher Effekte (Gorin & Stone, 2001; Pohl, 2004; siehe Übersicht Fahrenberg, 2006), insbesondere zum negativen Retrospektionseffekt (Käppler et al., 1993; Fahrenberg et al., 2002a). Beim Vergleich der gemittelten Tageswerte mit den am Abend bzw. am nächsten Morgen erhobenen Rückblicken zeigt sich häufig eine einheitlich negative Verzerrungstendenz: der Tageslauf wird als belastender, körperlich und geistig anspannender, und die Stimmung als stärker aufgeregt/nervös und ärgerlich/gereizt beschrieben als aufgrund der Tages-Mittelwerte der aktuellen Einstufungen zu erwarten war. Der Befund ist gut reproduziert, doch konnte eine Serie der Methodenstudien die maßgeblichen Bedingungen noch nicht aufklären. Es handelte sich zwar statistisch um einen mittleren bis großen Effekt, aber er war über mehrere Wochentage intra-individuell nicht überzeugend reproduzierbar, und er trat vorwiegend bei bestimmten Adjektivskalen auf: es waren Items mit *negativem* Inhalt und unter diesen vor allem die Items "anstrengend, belastend", "bedrückt" und "Erlebten Sie seit der letzten Eingabe Stress?" Beteiligt sein Könnten u.a. die suggestive Wirkung "negativ" formulierter Items, die statistische Tendenz zur Mitte, situative Tageseinflüsse und eine persönlichkeitspsychologische Tendenz zu negativen Bewertungen (Klagsamkeit), vielleicht auch urteilsheuristische Tendenzen (Fahrenberg et al., 2002a).

Statistische Definition der intra-individuellen Variabilität

Wegen der besonderen Bedeutung der intra-individuellen Variabilität im Ambulanten Assessment stellt sich die Frage nach der geeigneten statistischen Beschreibung. Mehrere Autoren, wie z.B. Eid und Diener (1999) diskutierten verschiedene Aspekte der intra-individuellen Varianz, verwendeten letztlich aber nur die Standardabweichung der Itemwerte. Die Varianz bzw. Standardabweichung sind nur bedingt geeignet, denn sie erfassen nicht die serielle Abhängigkeit. Für Zeitreihen wurden deshalb spezielle Statistiken vorgeschlagen: (1) die Häufigkeit des Wechsels von Zunahme und Abnahme der Itemwerte (Vorzeichenwechsel) als nicht-parametrischer Index und (2) das mittlere Quadrat sukzessiver Differenzen (MQSD, englisch MSSD). Die statistische Definition von Variabilität ist komplizierter als es zunächst den Anschein hat. So wird MQSD, im Unterschied zum Vorzeichenwechsel, von Ausreißer-Werten und von allgemeinen Trends beeinflusst. So gibt es in der Psychophysiologie, insbesondere hinsichtlich der Variabilität der Herzfrequenz viele Bemühungen, einen bias-freien Variabilitätsindex zu entwickeln (Berntson et al., 1997); die meisten dieser Indizes setzen jedoch Intervall- oder Verhältnisskalen voraus.

Der Quotient aus MQSD und Varianz kann als einfacher Index der Autokorreliertheit einer Zeitreihe dienen (von Neumann) mit Signifikanzprüfung nach Yamane (1969). In der Freiburger Wochenstudie ergab sich eine weitgehende Übereinstimmung: hohe Werte dieses Quotienten zeigen eine hohe negative Autokorrelation, niedrige Werte (< 1.0) eine positive serielle Korrelation an.

Psychometrische Zuverlässigkeit

Wenn es um Selbstbeurteilungen der Befindlichkeit geht, stellt sich wie in anderen Bereichen der Diagnostik die Frage nach der Zuverlässigkeit solcher Einstufungen. Die Antworten fallen aus drei Gründen besonders schwer: Es sind subjektive Auskünfte ohne weitere Prüfmöglichkeiten, die Einstufungen beziehen sich auf u.U. sehr kurzfristig veränderliche Zustände, und die erhaltenen numerischen Werte sind im Grunde nur Ordinaldaten. Dementsprechend kann es nicht verwundern, dass bisher keine fachlich breit akzeptierte Taxonomie der Kategorien bzw. der Dimensionen der Befindlichkeit erarbeitet werden konnte.

Die klassische Testtheorie wurde im Hinblick auf Intelligenztests entwickelt und später auf Persönlichkeitstests übertragen wurde; sie bezieht sie sich primär auf zeitlich überdauernde, stabile Eigenschaften. Beim Assessment von Befindlichkeit geht es um einen grundsätzlich anderen Sachverhalt, denn hier interessieren gerade die Zustandsänderungen, d.h. eine Variabilität, die sonst in der Regel dem Messfehler zugerechnet wird. Deswegen werden im Folgenden die wichtigsten Prinzipien der klassischen Testtheorie referiert (ohne auf Messmodelle mit korrelierten Fehlern, Latent-Structure-Modelle u.a. einzugehen). Anschließend ist zu überlegen, wie die traditionell berechneten Reliabilitätskoeffizienten im Fall der Befindlichkeitsdaten überhaupt noch zu interpretieren sind. Die teststatistischen Verfahren und Bewertungen stehen in einem inneren Zusammenhang mit theoretischen Vorannahmen zur Operationalisierung der Konstrukte.

Reliabilität

Reliabilität (Zuverlässigkeit) meint die Genauigkeit, mit der ein Test ein reproduzierbares Ergebnis liefert (instrumentelle Präzision), unabhängig davon, ob es dem entspricht, was der Test zu messen vorgibt. Eine adäquate Reliabilität der Messung ist eine notwendige, aber noch nicht hinreichende Bedingung für die Validität des Testverfahrens.

Jedes Messmodell muss bestimmte Annahmen machen, auch wenn diese nur näherungsweise oder nur unter bestimmten Rahmenbedingungen zutreffen. Grundlage der klassischen Testtheorie ist die Annahme, dass sich jeder gemessene Wert aus einem "wahren" Wert und einem Fehleranteil (bzw. mehreren Fehlerkomponenten) zusammensetzt. Weiterhin soll gelten: Der Erwartungswert (Mittelwert) für alle Fehlerkomponenten ist null. Die verschiedenen Fehlerkomponenten sind voneinander unabhängig, und sie sind auch nicht mit dem "wahren" Wert korreliert. Als fundamentale Voraussetzung des Messmodells gilt auch die Objektivität der Daten, d.h. die prinzipiell mögliche Kontrolle durch andere Beobachter – eine Annahme, die offensichtlich bei den allermeisten Selbstbeurteilungen (Skalen, Fragebogen) nicht zutrifft.

Aus den genannten Axiomen ergibt sich: die Varianz des Messwertes setzt sich additiv aus den Varianzen von "wahrem" Wert und Fehler zusammen. Die Reliabilität eines Tests bzw. einer Messung wird statistisch definiert als das Verhältnis von wahrer Varianz zu beobachteter Testvarianz. Wenn die Reliabilität eines Tests bekannt ist, lässt sich schätzen, in welchem Konfidenzintervall der wahre Wert liegen wird. Der fehlerbedingte Anteil an der Streuung eines Messwertes wird mit dem Begriff des Standardmessfehlers beschrieben. Anhand des Standardmessfehlers aus der Normierungsstichprobe des Tests kann statistisch geprüft werden, ob zwischen den Testwerten verschiedener Personen oder – mit noch mehr Vorbehalten – zwischen zwei wiederholten Messungen (Differenzwerte) nur zufällige oder bedeutsame Unterschiede bestehen.

Ein Seitenblick auf die Physiologie ergibt, dass dort das Thema "Reliabilität" einen geringeren Raum einnimmt als in der Testpsychologie. Selbstverständlich liegt dies hauptsächlich an der weithin naturwissenschaftlichen Methodik. Dennoch sind die anstelle von "Re-Test-Reliabilität", "Konsistenz" und "Stabilität" verwendeten Begriffe erwähnenswert. In der Biomedizin ist häufig die Unterscheidung von Auflösung, Genauigkeit und Reproduzierbarkeit einer Messung zu finden:

– Auflösung in zeitlicher Hinsicht;
– Amplitudenauflösung;

– Genauigkeit (Accuracy) bezogen auf eine Referenzmethode (einen Standard mit "wahren" bzw. bestmöglichen Messwerten, dem sog. Goldstandard), statistisch zu definieren als Mittelwert der Differenz zweier Methoden (mittlere Abweichung);
– Reproduzierbarkeit (Precison), statistisch zu definieren als Standardabweichung der Differenzen der Messwerte zweier Methoden.

Häufig wird der von Bland und Altman (1986) vorgeschlagene Koeffizient der "Repeatability" verwendet; er gibt den Bereich an, in den 95 % (± 2 SD) der wiederholten Messungen fallen werden.

Grundsätzlich wird die äquivalente Wiederholbarkeit der Messoperation postuliert. Im Bereich der psychologischen Methodik kann dies höchstens in grober Näherung behauptet werden. Es ist eine sehr fragwürdige Annahme, denn Phänomene wie die methodenbedingte Reaktivität, die zunehmende Vertrautheit mit der Aufgabe, die Ausbildung von Strategien und subjektiven Theorien u.a., müssen negiert werden.

Reliabilitätsschätzungen

Die Reliabilität eines Tests kann auf verschiedene Weise bestimmt werden:

(1) Die Reproduzierbarkeit des Ergebnisses zu verschiedenen Zeitpunkten ("Retest-Reliabilität"),
(2) Die Vergleichbarkeit der Ergebnisse von Parallelformen ("Paralleltest-Reliabilität"),
(3) Die Übereinstimmung von Teilen des Tests Präzision des Test als solchem ("Halbierungs-Reliabilität", in verallgemeinerter Form die "innere Konsistenz" bzw. Homogenität der Items).

Die verschiedenen Formen der Reliabilitätsschätzung werden unterschiedliche Resultate liefern, da jeweils typische Durchführungsbedingungen und Fehlereinflüsse bestehen.

Die Retest-Reliabilität bestimmt sich aus der Korrelation der Testwerte, die zu zwei Zeitpunkten erhoben wurden. Diese auch als Stabilitätskoeffizienten bezeichneten Retest-Koeffizienten beschreiben also die Reproduzierbarkeit der Werte über ein Zeitintervall, jedoch nur die Enge des Zusammenhanges und noch nicht den möglichen Unterschied der Mittelwerte wie es ein Intra-Klassen-Koeffizient leistet.

Die differenzierte Beschreibung von Befindlichkeiten wird insbesondere die Zustandsänderungen zum Ziel haben. Folglich wären, je nach Zeitraster und Merkmal, hohe Retest-Koeffizienten eine unerwünschte Eigenschaft der Skala. Da diese gesuchte Varianz mit den Fehlervarianz der Methode konfundiert ist, sind Retest-Koeffzienten von Stimmungsskalen testmethodisch kaum interpretierbar.

Die Schätzung der Paralleltest-Reliabiliät würde zwei weitgehend äquivalente Skaleninstrumente und ein cross-over design zur Kontrolle der Reihenfolgeeffekte verlangen. Im Vergleich hierzu ist die Konsistenzanalyse leichter durchzuführen.

Die Halbierungs-Reliabilität bezieht sich auf zwei möglichst gleichartige Teile des Tests (erste/zweite Hälfte, gerade/ungerade Item-Nummern), und die Konsistenzanalyse – als Verallgemeinerung – auf alle Items eines Tests. Das Maß der inneren Konsistenz wird meist durch den Koeffizienten α (sog. Cronbach α) angegeben. Hohe Koeffizienten der inneren Konsistenz bedeuten eine hohe Homogenität (psychologische Äquivalenz) der einzelnen Items. Wenn sie tatsächlich als gleichwertige, "parallele" Messungen der zugrundeliegenden Eigenschaft gelten sollen, müssen sie hohe oder sehr hohe Interkorrelationen aufweisen. Faktorenanalytisch betrachtet ist die Kommunalität dieser Items hoch und ihre Spezifität gering. Bei der Konstruktion eines Tests werden in der Regel solche Items ausgewählt, die möglichst zu einer Maximierung der Faktorladungen auf der betreffenden Dimension und zur Maximierung der inneren Konsistenz-Reliabilität führen. Durch die Addition einer Anzahl solcher Items als parallele (und unabhängige?) Messoperationen wird die Reliabilität erhöht bzw. der Messfehler reduziert.

Die einzelnen, sehr homogenen Items eines bestimmten Aufgabentyps, z.B. Symbole zuordnen, aufmerksam reagieren, Analogien erkennen, Zahlen merken, könnten durchaus als Parallelmessungen des jeweils zugrundeliegenden Intelligenzfaktors angesehen werden. Die übliche Itemselektion anhand der Trennschärfeindizes und der Schwierigkeitsindizes begünstigt hier die Auswahl sehr ähnlicher, nahezu redundanter Items mit sehr hoher gemeinsamer Varianz. Automatisch steigt die Halbierungs- und Konsistenz-Reliabilität.

Was für Intelligenz- und Leistungstests adäquat ist, kann bereits bei den viel facettenreicheren Konstrukten der Persönlichkeitstests problematisch sein. Wenn Testautoren und Testanwender sehr hohe Skalen-Konsistenzen mitteilen, kann dies u.U. sogar einen schwerwiegenden konstruktiven Mangel einer Persönlichkeits-Skala oder Klinischen Skala anzeigen.

Konsistenzkoeffizienten können folglich unter zwei Perspektiven interpretiert werden: Im Hinblick auf die Reliabilität einer Skala beschreiben sie die Übereinstimmung der einzelnen Messoperationen. Bei perfekter Korrelation (Parallelität) der Items erreicht die Konsistenz-Reliabilität den Wert 1. – Im Hinblick auf die Operationalisierung eines psychologischen Konstrukts drückt der Konsistenzkoeffizient aus, wie eng oder weit (facettenreich) der Bedeutungsumfang während der Testentwicklung festgelegt wurde. Ein sehr hoher Koeffizient kennzeichnet ein entsprechend eingeengtes, empirisch-semantisch reduziertes Konzept. Die statistische Konsistenz von Items muss also im Kontext der psychologischen Konstruktdefinition betrachtet werden.

Lokale und aggregierte Reliabilität

Begriffliche Differenzierungen wurden von Pawlik und Buse (1992) vorgenommen. Sie unterschieden zwischen zwei Ebenen der psychometrischen Reliabilität (Buse & Pawlik, 1984, 1994;1996) als sie die Korrelationen von Items bzw. von Aggregaten verschiedener Leistungstests und von Befindensmerkmalen sowie physiologischen Messwerten innerhalb und zwischen verschiedenen Settings und Untersuchungstagen analysierten.

Lokale Reliabilität, d. h. Reliabilität einer Messung (Items eines Leistungstests bzw. Items von Befindenslisten) innerhalb derselben Testanwendung bzw. Befindlichkeitseinstufung. Die Berechung erfolgt nach der odd even Methode der Halbierung oder als Konsistenz der einzelnen Testwerte bzw. Einstufungen. Diese Koeffizienten können anschließend über alle Gelegenheiten, d. h. Tage oder Settings/ Situationen gemittelt werden.

Aggregat-Reliabilität, d. h. Reliabilität von Aggregaten (Mittelwerten) über Settings/Situationen, Zeiträume bzw. Wiederholungen. Die Berechnung erfolgt anhand von zwei Hälften, z. B. den Mittelwerten an geraden und an ungeraden Tagen einer Woche oder an den Aggregaten von Itemlisten (Skalenwerten) an verschiedenen Tagen.

Die mitgeteilten Beispiele ermöglichen interessante Vergleiche der Konsistenzkoeffizienten und der Reproduzierbarkeit für bestimmte Settings/Situationen und für aufeinanderfolgenden Tage (Buse & Pawlik, 1996, 2001; Pawlik & Buse, 1994; siehe auch Perrez et al., 2001, Wilhelm & Perrez, 2001). Es muss natürlich beachtet werden, ob es sich um objektive Testdaten oder um Selbsteinstufungen handelt, um Sets einzelner Items oder Aggregate von Items, ob Messungen und Items als Parallelmessungen postuliert bzw. die ermittelten Aggregate aus verschiedenen Protokollen oder Settings als äquivalent aufgefasst werden.

Über die Berechnung einzelner Koeffizienten hinaus sind systematische Generalisierbarkeitsstudien nach Cronbach, aufschlussreich, weil die verschiedenen Varianzquellen geschätzt und verglichen werden können. Diese G-Studien verlangen allerdings viele Bedingungen und Replikationen im Sinne einer Datenbox, um auch praktische Empfehlungen zur Generalisierbarkeit (externen Validität) geben zu können.

Das traditionelle Konzept der Test-"Reliabilität" hat also mehrere Aspekte, die in den Lehrbüchern zur Testmethodik kaum behandelt werden, sondern eher bzw. differenzierter – wie hier – in der Methodik der Verhaltensbeobachtung, und speziell für serielle Beobachtungen in Einzelfallstudien (siehe Suen, 1988; Suen & Ary, 1989) oder in neueren Beiträgen zum "Multimethod Measurement in Psychology" (Eid & Diener, 2004).

Reliabilität, Konstruktvalidität und Kriterienvalidität, Testökonomie

In der Testmethodik gibt es widerstreitende Prinzipien, zwischen denen ein Kompromiss zu suchen ist. Der zu geringe Reliabilitätskoeffizient eines Tests (Anteil wahrer Varianz/Fehlervarianz) limitiert die maximal erreichbaren Validitätskoeffizienten (vorhersagbare Kriterienvarianz). Aber eine hohe innere Konsistenz eines Tests (extreme Item-Homogenität) bedeutet – anders betrachtet – Redundanz, geringere Testökonomie und potentiell einen Verlust an u.U. wesentlichen Facetten des gemeinten theoretischen Konstrukts, u.U. einen Operationalisierungsfehler (partielle Inkompatibilität von hoher innerer Konsistenz und Kriterienvalidität im Sinne des Linsenmodells von Brunswik bzw. Wittmann, vgl. Beauducel et al., 2005).

Zu den Qualitätsstandards gehört es, die Reliabilität und Validität einer psychologischen Untersuchungsmethode zu evaluieren. Offensichtlich handelt es sich in beiden Fällen um multi-referentielle Konzepte. Die Beurteilungen sind nicht in einfachen Koeffizienten zusammenzufassen, sondern verlangen Präzisierungen und methodenbewusste Interpretationen im Kontext der jeweiligen Untersuchungen. Außerdem sind die berechneten Koeffizienten populationsabhängig, wobei eine Einschränkung der intra- und inter-individuellen Varianzen zu erheblichen systematischen Verzerrungen führen kann.

Nicht selten vermitteln Test-Publikationen oder empirische Arbeiten den Eindruck, dass – ungleichgewichtig – mehr über die Reliabilitätsaspekte als über die anderen Gütemerkmale berichtet wird. Die konventionellen Reliabilitätsschätzungen sind wie die internen Item- und Faktorenanalysen, auf einfachste Weise möglich, der Nachweis einer neuen (externen) Kriterienkorrelation (nicht bloß mit ähnlichen Fragebogen) oder sogar eines inkrementellen Nutzens für reale Assessment-Entscheidungen sind unvergleichlich viel schwieriger und aufwändiger, so dass sie oft nicht einmal versucht werden. Wichtig für die Testpraxis sind auch die Testökonomie ("Validität pro Zeiteinheit"), die Zumutbarkeit, Fairness und Akzeptanz des Verfahrens. Bei der testmethodischen Bewertung sind mehrere und sich u.U. widersprechende Gütemerkmale anzuwägen. Die Reliabilität ist also nur einer der wichtigen Aspekte der Test- und Messmethodik.

Zusammenfassung

Auf die Frage, wie die Zuverlässigkeit von Selbsteinstufungen der Befindlichkeitsänderungen bestimmt werden kann, gibt es keine befriedigende Antwort. Bei Ein-Item-Skalen ist eine Reliabilitätsschätzung, abgesehen von einer inadäquaten momentanen Wiederholung, nicht einmal formal möglich. Die Konsistenzanalyse zur Schätzung der lokalen Reliabilität gibt es hier nicht. Für die Items einer Stimmungsskala kann zwar eine Konsistenzanalyse berechnet werden, doch bleibt die Interpretation schwierig. Die Bewertung der Koeffizienten hängt wesentlich von der vorausgegangenen Operationalisierungsentscheidung ab. Wird eine sehr enge konzeptuelle Reduktion bevorzugt oder eine facettenreichere Repräsentation des gemeinten psychologischen Konstrukts? Diese operationalen Festlegungen (und potentiellen Operationalisierungsfehler) bilden sich maßgeblich in der Item-Konsistenz ab.

Die Stellungnahme zu diesem Thema wird folglich von bestimmten Vorannahmen abhängen. Wenn in einem Intelligenz-Untertest Symbole zuzuordnen sind, kann für diese vielen einzelnen Operationen gewiss ein hohe Homogenität behauptet und dementsprechend die Reliabilität bestimmt werden. Das Konstrukt Extraversion hat demgegenüber so viele wichtige Facetten, dass heterogen erscheinende Aussagen kombiniert werden müssen, um diesen "Sekundärfaktor" zu repräsentieren. Eine schematische Itemanalyse zur Maximierung der Konsistenz wäre unangebracht. Sind nun die in der Fachliteratur beschriebenen Dimensionen und Kategorien der Befindlichkeit, wenn über die adäquate inhaltliche Auswahl der Deskriptoren zu entscheiden ist, eher einer elementaren Intelligenzfunktion (Symbole zuordnen) oder einer facettenreichen Persönlichkeitseigenschaft zu vergleichen? Diese psychologische Aufgabe der Operationalisierung kann durch den Formalismus einer Item- oder Faktorenanalyse nicht ersetzt, sondern höchstens unterlegt werden.

Die grundsätzlichen Schwierigkeiten einer Reliabilitätsschätzung für Einstufungen der Befindlichkeit – zweifelhaft bei der Konsistenzanalyse von Items, unmöglich bei Ein-Item-Skalen – wurden hier hervorgehoben. Dieser Sachverhalt kann irritieren, denn in der konventionellen Testdiagnostik werden solche Koeffizienten häufig als erstes mitgeteilt. Hier wird für eine Relativierung der Argumente plädiert: Im Vergleich zu der methodisch kühnen Behauptung, subjektive Auskünfte über innere Zustände überhaupt wissenschaftlich und quantitativ untersuchen zu können, sind die problematischen Aspekte der Reliabilitätsanalyse oder die Details statistischer Tests gewiss nicht unwichtig, aber nachgeordnet.

Diese skeptische Einstellung bzw. die zurückhaltende Interpretation von Reliabilitätsschätzungen bedeutet keinesfalls, dass einzelne Fehlerquellen übersehen oder gering geschätzt werden dürfen. Gerade aus solchen Überlegungen wird die computer-unterstützte Methodik heute oft die Methode der Wahl sein. Die momentane, kontextbezogene und zeitlich protokollierte Selbsteinstufung vermeidet oder reduziert wesentliche Fehlerquellen der retrospektiven Erhebung und zeichnet sich durch eine höhere "technische" Reliabilität aus. Einige der Fehlerkomponenten von Selbsteinstufungen sind ty-

pisch für Fragebogen und können durch die computer-unterstützte Methode weitgehend oder völlig vermieden werden. Dazu gehören u.a. die zeitlich genau protokollierte Dateneingabe und damit die Kontrolle der Compliance, die Unzugänglichkeit der vorausgegangenen Einträge und die Sicherheit des Datentransfers zum Auswertungsrechner.

Strategische Schlussfolgerungen

In der Forschung über Befindlichkeit können sich die Untersucher methodisch nicht oder nur mit wesentlichen Vorbehalten auf die konventionellen Bestimmungsverfahren der Reliabilität stützen. Statistische Entscheidungen über Hypothesen müssen also ohne geeignete Reliabilitäts-Informationen durchgeführt werden. Strategisch folgt daraus zweierlei: (1) sich primär um die Analyse der viel wichtigeren empirischen Validität zu kümmern ("valide Tests sind de facto reliabel") und (2) überzeugende Replikationen (Schweizer, 1989) wichtiger Befunde mit relativ großer Personenzahl systematisch einzuplanen, zunächst innerhalb und dann zwischen den Arbeitsgruppen.

Zufallskritische Beurteilung von individuellen Zustandsänderungen

Zur statistischen Beurteilung von intra-individuellen Differenzen ist der Standardmessfehler nur mit großen Vorbehalten geeignet, da bei seriellen Daten keine statistische Unabhängigkeit der Messungen existiert. In der Arbeitsrichtung der psychologischen Zeitreihenanalyse sind deshalb verschiedene Methoden vorgeschlagen worden, wie Hypothesen über individuellen Zustandsänderungen geprüft werden können. Dabei soll die serielle Abhängigkeit berücksichtigt werden, u. a. im ARIMA (Auto Regressives Integriertes Moving Average-) Modell. Bei Anwendungen auf psychologische Daten, z. B. in der differentiellen Psychologie und in der Psychotherapieforschung ergaben sich in vielen Fällen große Inkonsistenzen der Modelle und Schätzungen zwischen Personen und zwischen Variablen. Dieser Sachverhalt erlaubte dann statt verallgemeinernder Aussagen nur Einzelfalldarstellungen. Die Zusammenfassung solcher Einzelfälle wird in der Fachliteratur ausführlich behandelt. Am Beispiel von Tagebuchdaten haben u.a. Ott und Scholz (2001) die Möglichkeiten und Schwierigkeiten solcher Aggregationen diskutiert.

Die Nicht-Zufälligkeit der Zunahme oder Abnahme eines Itemwertes in einer individuellen Zeitreihe statistisch zu prüfen, ist das Ziel vieler Überlegungen und Verfahrensvorschläge. Die metrischen Voraussetzungen und die problematischen Entscheidungskriterien einer ARIMA-Analyse lassen diese – in anderen Bereichen genutzten – Verfahren in Hinblick auf *individuell*e Befindlichkeitsänderungen als weitgehend ungeeignet erscheinen. Die Prüfung von Gruppenunterschieden in regelmäßig wiederkehrenden, d.h. durch bestimmte natürliche oder experimentelle Auslösebedingungen induzierten Verläufen, bleibt dagegen gut möglich. Für die statistische Prüfung von Gruppenunterschieden der gemittelten Verläufe stehen verteilungsfreie Verfahren und parametrische, darunter auch multivariate Methoden, zur Verfügung – auf Kosten eines Teils der differentiellen Effekte.

Noch kaum angewendet werden in diesem Bereich die Verfahren zur Konstruktion nichtparametrischer Konfidenzintervalle, die sich auf die Jackknife- und Bootstrap-Methodik stützen, d.h. keine Verteilungsannahmen machen, sondern eigens generierte, problembezogene Verteilungen zur Absicherung bzw. zum Testen statistischer Hypothesen benutzen (siehe Krauth, 1995, Rodgers, 1999).

Im Kontrast zu zweifelhaften statistischen Argumenten könnte im Sinne des "interpretativen Paradigmas" auf die wissenschaftsmethodische Eigenart, d.h. die besondere Qualität der Selbstberichte verwiesen werden. Wenn der (die) Befragte eine Veränderung der Befindlichkeit mitteilt, ist dies eben empirisch bedeutsam und kann intersubjektiv nicht falsifiziert oder als "zufällig" bewertet werden. Diese Argumentation negiert zwar nicht die Möglichkeit von Erinnerungsfehlern, von verzerrenden Urteilsstrategien oder Täuschungen, sieht aber keinen Weg, diese Effekte statistisch als "Fehler" zu separieren, sondern höchstens die zweifelhafte Möglichkeit, im Einzelfall zu explorieren und zu begreifen.

Problematischer Gebrauch der konventionellen Itemanalyse und Faktorenanalyse

Wie die Itemanalyse so ist auch die Methodik der Faktorenanalyse primär für die Intelligenzforschung entwickelt und später auf die Konstruktion von Persönlichkeits-Fragebogen und Stimmungsskalen übertragen worden.

Bei schematischer Anwendung begünstigt die faktorenanalytische Methodik die Entstehung sehr homogener Skalen, d.h. die formale Maximierung der inneren Konsistenz ohne Rücksicht auf die empirische Validität (siehe oben). Bei diesen Analysen können anfänglich enthaltene, d.h. vorgegebene oder noch unzureichend erkannte Dubletten (Tripletts usw.) weitgehend redundanter Items aufgrund ihrer sehr hohen Kommunalität die Ladungsmuster bzw. die Rotation dominieren, mit Folgeschäden für die Evaluation der relativen Varianzanteile und anderer Eigenschaften.

Generell besteht also ein hohes Risiko, dass inhaltlich sehr ähnliche Items technisch aufgrund ihrer höheren gemeinsamen Varianz in den Faktorenanalysen begünstigt werden: die Itemselektion nach Faktorladung bzw. nach Trennschärfeindex führt zwar zu höheren Koeffizienten der inneren Konsistenz, aber damit auch zu einem Verlust an Konstruktfacetten und eventuell auch an empirischer Brauchbarkeit. Eine sehr hohe innere Konsistenz einer Skala kann testmethodisch unerwünscht sein!

Vergleichende Analysen an der Freiburger Wochenstudie (33 Personen x 42 Berichte) zeigten, dass Faktorenanalysen (1) an der Matrix der inter-individuellen Korrelationen (über alle Datenpunkte R x P) und (2) an den intra-individuellen (über Personen gepoolten) Korrelationen keine psychologisch äquivalenten Dimensionalitäten liefern. Die ersten Komponente beider Analysen waren zwar verhältnismäßig ähnlich, doch divergierten die folgenden Ladungsmuster. Bei intra-individueller Analyse waren die Kommunalitäten und die extrahierte Varianz erheblich größer, die Anzahl der markanten Faktoren eher kleiner. Eine noch größere Heterogenität ist zu erwarten, wenn statt der gepoolten Zeitreihen die individuellen Zeitreihen (P-Technik der Faktorenanalyse) verwendet würden. Hier sind gründlichere Methodenstudien angebracht.

Fehlbewertungen faktorenanalytischer Resultate

Aus den skizzierten Gründen kann eine schematische Anwendung faktoren- und itemanalytischer Strategien zu zweifelhaften Skalenkonstruktionen führen. Auch in der theoretischen Interpretation von Faktorenanalysen scheinen sich einige Ansichten zu halten, die kaum zu begründen sind. Dazu gehören Postulate: die basalen Faktoren der Stimmung bzw. der Persönlichkeit gefunden zu haben und damit eine real existierende Entität zu erfassen – statt nur ein mehr oder minder nützliches Beschreibungssystem geschaffen zu haben.

Schon Guilford, einer der Pioniere faktorenanalytischer Intelligenz- und Persönlichkeitsforschung hat vor 50 Jahren darauf hingewiesen, dass höchstens dann über die basale und erschöpfende Anzahl der relevanten Faktoren diskutiert werden kann, wenn das Universum der Items repräsentiert ist. Weder in der Sphäre de Persönlichkeitsmerkmale noch in der Sphäre der Befindensweisen ist eine solche Abgrenzung und Stichprobenziehung praktizierbar, trotz aller lexikalischen Sammelversuche. Da es weitgehend willkürlich ist was zur "Persönlichkeits-Sphäre" gehören soll oder nicht, bleibt es beliebig, außer den Temperamentseigenschaften im engeren Sinn auch andere Bereiche einzubeziehen, d.h. Selbstkonzepte und Kontrollüberzeugungen, Einstellungen, Interessen und Wert-Orientierungen, aber auch Grundstimmungen, körperliche Befindensweisen und alltägliche Beschwerden. Als Konsequenz dieser Beliebigkeit würden sich die Strukturen und Varianzanteile der Komponenten wesentlich verändern. Deshalb müssen alle dieser Reduktionsversuche einen mehr oder minder gravierenden Bias haben. Dies gilt nicht minder für die Sphäre der Befindlichkeit.

Eine definitive Anzahl von Grund-Dimensionen der Befindlichkeit festlegen zu wollen, ist ebenso wenig sinnvoll wie entsprechende Behauptungen im Bereich der Persönlichkeitseigenschaften über eine fixe Anzahl von Basis-Faktoren.

Circumplex-Darstellung

Wenn die hauptsächlichen Emotions-Dimensionen systematisch zusammenhängen, bietet sich statt einer Faktorenanalyse mit schiefwinkliger Rotation der Hauptachsen eher eine Circumplex-Skalierung an. So meinte Russell (1980), dass sich in einer zyklischen Anordnung abbilden lassen: Freude (0), Aufregung (45), Arousal (90), Distress (135), Unlust (180), Depression (225), Schläfrigkeit (270), und Entspannung (315 Grad). Diese Konzeption sei durch verschiedene Analysetechniken und interkulturelle Studien bestätigt worden. Die zyklisch angeordneten Deskriptoren repräsentieren angeblich zwei Dimensionen der Emotionsqualitäten (Lust- Unlust und Grad der Erregung) und ggf. die Länge des Vektors die Intensität des Zustandes.

Die ursprünglich von Guttman vorgeschlagene Circumplex–Darstellung wird ihren methodischen Nutzen in einigen Bereichen der psychologischen Methodik haben. Wenn sie auf die Beschreibung der

subjektiven Befindlichkeit angewendet wird, kommt es zu einer fragwürdigen Vereinfachung, einer zweifellos sparsamen "Procrustes"-Lösung: "Applying this circumplex model for the assessment of mood in ambulatory studies has several advantages. By accounting for most of the qualitative variability in mood states, it provides a parsimonious way of assessing mood "…. (Jacob et al., 1999). Diese Autoren beziehen sich hier auf die Anordnung der Farbtöne im "Farbkreis" als Vorbild einer sparsamen Lösung. Aber ist die Befindlichkeit eines Menschen, die veränderte bzw. gestörte Befindlichkeit eines Patienten, in einem Vektor dieser Art hinreichend erfasst oder bilden solche Circumplex-Darstellungen höchstens eine graphisch formulierte und simplifizierende Merk- oder Kommunikations-Hilfe? Entsprechend kritische Einwände werden sich regen, wenn die Self-Assessment-Manikin Darstellung (Bradley, Greenwald & Hamm, 1993), d.h. die in den USA beliebte Kennzeichnung von Emotionen durch extrem reduzierte Gesichtschemata, verwendet wird.

Fortbestehende semantische Probleme und Alternativen?

Verbale Aussagen aufgrund von Introspektion bzw. Selbstbeobachtung haben einen eigentümlichen wissenschaftsmethodischen Status, sie gelten als Daten, die von anderen Personen grundsätzlich nicht überprüfbar sind und dennoch eine unverzichtbare Informationsbasis der wissenschaftlichen Psychologie geben. Die grundlegenden semantischen Probleme bilden ein überdauerndes Thema der empirischen Psychologie, aber auch der philosophisch-analytischen Reflektion.

Die Aussagen über innere Zustände haben eine logisch-inhaltliche Gültigkeit: Wer sonst sollte solche Auskünfte gegen können? Die empirische Gültigkeit im Sinne eines realen Zutreffens kann grundsätzlich nicht durch Beobachtung von dritter Seite entschieden werden (siehe die in der Analytischen Philosophie des Geistes verbreitete Redeweise von erster bzw. dritter Person). Auskünfte über solche inneren Prozesse sind grundsätzlich nicht falsifizierbar, doch existieren durchaus einige schwächere Bestätigungsweisen. Ein Beispiel für die partielle intersubjektive Prüfbarkeit von introspektiven Aussagen ist die vorhergesagte Verbesserung der Befindlichkeit durch einen Tranquilizer in einem psychopharmakologischen Doppel-Blind-Experiment, eingesetzt zur Validierung einer mehrdimensionalen Stimmungsskala (Hampel, 1972). In der Methodik der Psychologie sind verschiedene Typen der Bestätigung (Konfirmation) solcher Auskünfte über "innere Sachverhalte" entwickelt worden. Dazu gehören u.a.: die ausdruckspsychologischen (Mimik) oder physiologischen (vegetativen, motorischen, elektroenzephalographischen) Vorgänge (Begleiterscheinungen) von erlebten Emotionen. Aber eine wechselseitige Validierung scheint beim gegenwärtigen Stand der multimodalen Untersuchungsmethodik weiterhin unmöglich zu sein, denn zwischen den Beschreibungsebenen von "Emotion" existieren häufig tiefreichende Inkonsistenzen (Diskrepanzen, Desynchronien).

Da sehr viele der deutschen Fragebogen und Skalen Adaptationen angloamerikanischer Verfahren sind, stellt ich die schwierige Frage der angemessenen Übersetzung. Wird eine möglichst treffende lexikalische Übersetzung gesucht oder vielmehr eine psychologisch äquivalente Formulierung, die u.U. einen anderen Inhalt für ein Item verlangt? Aufgrund seiner Erfahrungen in der Konstruktion von Persönlichkeits-Inventaren äußerte H. J. Eysenck gesprächsweise die Ansicht, dass sich seines Erachtens die Mentalität der Menschen in England und in den USA (damals) in psychologisch wesentlichem Ausmaß unterscheiden, sogar die Einstellungen der Engländer und der Deutschen in vieler Hinsicht ähnlicher sein würden. Es mangelt auf dem Gebiet persönlichkeitspsychologischer Deskriptoren an genauen Berichten über die sprachlichen Kompromisse bei der Übersetzung oder der absichtlichen psychologischen Transposition von Iteminhalten. Die neuere kulturvergleichende Forschung ist in dieser Hinsicht offensichtlich kritischer geworden und begnügt sich nicht mit der früher oft praktizierten, schlichten Übernahme englischer Fragebogen.

Nur sehr selten wird der naheliegenden methodischen Forderung entsprochen, bei der Adaptation amerikanischer Tests auch die gesamte Konstruktion mit einem weitergefassten deutschen Itempool zu wiederholen und sich nicht allein mit der Erhebung von Testdaten bei deutschen Gelegenheits-Stichproben zu begnügen. – Es sei denn, man wollte die vollständige Angleichung der psychologischen Profile der deutschen an die amerikanische Bevölkerung behaupten.

Es gibt zweifellos ein grobes, sprachlich gelerntes Vorverständnis, was z.B. mit dem Wort "angespannt" gemeint ist. Erst die genauere Analyse bzw. ein Interview der Befragten zeigt die semantischen Schwierigkeiten und die von Fall zu Fall erheblichen Unterschiede im Sprachgebrauch: einige meinen "geistig" angespannt, andere "emotional" aufgeregt oder nur "körperlich" verkrampft. Wird –

wie eingehend versucht – nach allen drei Facetten des Konstrukt gefragt, sind andere Personen irritiert, weil sie hier bisher keinen Unterschied sahen. Die Werte der drei Items korrelierten untereinander nur in der Größenordnung von .60, innerhalb Personen nur ca. 30 (Tagslaufstudie mit N = 52, Heger, 1990). Auch die zur basalen Verankerung von entspannter Ruhe verwendeten Bilder oder Situationen ("ich liege völlig entspannt auf einer grünen Wiese") haben von Person zu Person nicht selten einen bemerkenswert unterschiedlichen Assoziations- und Konnotationsraum.

Aus langen psychologischen Zeitreihenstudien (Fahrenberg, Myrtek, Kulick & Frommelt, 1977; Zimmermann, 1978) und aus vertiefenden Interviews ist zu entnehmen, dass einige Personen über individualcharakteristische Begriffe verfügen, um ihre alltäglichen Befindensänderungen zu beschreiben. Da solche Deskriptoren in den üblichen Listen fehlen werden, ist deren Einzelfall-Validität zugunsten der Standardisierung der Methodik und der Generalisierbarkeit der Ergebnisse gefährdet. Ein nur in wenigen Projekten beschrittener Ausweg ist die sog. Adjective Generation Technique AGT (Allen & Potkay, 1973, 1977), wobei die Deskriptoren von den Befragten weitgehend selber entwickelt werden. In Anlehnung an die Repertory Grid Technik entwickelte Schneider (1982) ein Verfahren zur Einschätzung der Belastungssituationen in psychophysiologischen Experimenten, doch fand er in den generierten Adjektiven in der Regel nur zwei Dimensionen des Situationserlebens im Labor (vgl. auch methodische Ansätze zur Rekonstruktion von Tagesereignissen, Kahneman et. al., 2004).

Ein neuer Ansatz ist die computer-unterstützte Methodik des Affect Grid (Reicherts, 2005), das einen Schritt zur individualisierten Selbstbeurteilung ermöglichen soll, indem Deskriptoren aus einer Liste ausgewählt werden können. Das adaptive System orientiert sich an Russell's Circumplex-Konzeption (Valenz & Aktivierung, plus Intensität) und bietet eine Auswahl aus 30 Deskriptoren an. Diese Items liessen sich, dem Konzept gemäß, clusteranalytisch in zwei bzw. vier Sets gruppieren.

Ein direkter Vergleich zwischen Personen ist wegen der unterschiedlichen Auswahl auf Itemebene nicht mehr möglich, sondern nur noch auf der Ebene der Skalenwerte. Dabei muss jedoch eine sehr hohe Konsistenz der Skalen vorausgesetzt werden, was der Idee der Individualisierung zu widersprechen scheint.

Unter dem Aspekt der missing data sind auch die von anderen Autoren programmierten Verfahren mit flexiblen Abfragen, d.h. mit Verzweigungen und Sprüngen, problematisch, denn die psychologisch wünschenswerte Individualisierung und optimalen intraindividuellen Verlaufsbeschreibung steht der methodischen Standardisierung für interindividuelle Vergleiche entgegen. Verwandte Fragen ergeben sich auch für die u.U., trotz der fraglichen Metrik solcher Daten, gewünschten Normierungen, die (1) inter-individuell, (2) intra-individuell (ipsativ) oder (3) repräsentativer auf den gesamten Datensatz (Personen x Bedingungen) vorgenommen werden kann. Jede dieser Normierungen, durch Eliminierung des Mittelwertes der Verteilung bzw. durch Standardisierung auf die Varianz, hat bestimmte Vorzüge Nachteile.

Insbesondere für Einzelfallstudien oder bei klinisch-psychologischen Fragestellungen kann es nützlich sein, die Liste der Standard-Items zu ergänzen, indem zu wesentlichen Bereichen eine individuelle Texteingabe ermöglicht wird (siehe oben). In der Freiburger Wochenstudie (mit 42 Selbstberichten jeder Teilnehmerin) wurden Texteingaben über (1) Beschwerden und (2) über besondere Ereignisse erbeten. Durchschnittlich gab es 9 Einträge (Variationsbreite 0 – 35) über momentane körperliche Beschwerden, meist Kopfschmerzen, Rückenschmerzen oder Mattigkeit (bei dieser Auswertung wurden 4 der 33 Teilnehmerinnen, die eine akute Erkältung hatten, ausgeklammert). Durchschnittlich 6 Einträge (Variationsbreite 0 – 16) bezogen sich auf besondere Ereignisse, meist positive oder negative soziale Kontakte, Arbeitsanforderungen, Feiern, Zwischenfälle, Nachrichten, sehr selten auch dramatische Episoden. Von den 189 Ereignis-Texten enthielten nur 11 einen direkten Bezug auf die Monitoring-Methode.

Natürlich können mehr freie Texteingaben zu anderen Aspekten, z.B. Setting und Tätigkeit erbeten werden oder auch eine Audio-Eingabe. Zu diesem Zweck wurde früher ein Walkman-Rekorder eingesetzt (Heger, 1990), heute kann der Festspeicher eines hand-held Computer für kürzere Aufnahmen ausreichen. Diese Optionen sind noch weitgehend ungenutzt.

4 Auswahlgesichtpunkte und Übersicht

Historisch-kritische Anmerkung: Vielfalt oder Reduktion?

Bereits Wundt hatte drei allgemeine Kennzeichen von Gefühlen unterschieden: Lust – Unlust, Erregung – Beruhigung und Spannung – Lösung. Die neueren Dimensionssysteme stützen sich auf operationale Definitionen und Instrumentarien, die empirisch anwendbar sind, auf faktorenanalytische, allgemein-sprachanalytische u.a. Ansätze: die neurobehaviorale Konzeption von Aktivierung – Deaktivierung oder von Richtung und Intensität der Aktivierung (Malmo, Duffy u.a.), die sprachanalytische Reduktion auf die Dimensionen Erregung (Aktivität), Valenz und Potenz (Mächtigkeit) wie im Semantischen Differential bzw. Polaritätenprofil (Osgood, Ertel u.a.). Dagegen wurden andere, traditionell untersuchte Aspekte wie Sozialbezogenheit, Kontrolle, Aufmerksamkeit–Rückzug (Zurückweisung, Rejektion) vernachlässigt (vgl. die bis heute gründlichste ideengeschichtliche Darstellung, Bottenberg, 1972; Übersichtsdarstellungen siehe u.a. Scherer, 1990; Larsen & Prizmic-Larsen, 2005; Schmidt-Atzert, 1996). Zunehmend werden auch moderne neurobiologische Konzepte in die emotionstheoretische Diskussion eingeführt (vgl. Peper, 2006; Posner, Russell & Peterson, 2005).

Traditionell wurde in der deutschen Literatur zwischen relativ überdauernden Stimmungen, den erlebnismäßig von diesem Hintergrund abgehobenen Gefühlen und den intensiveren Affekten unterschieden (vgl. Rohracher, 1988). Introspektiv und ausdruckspsychologisch waren also die geringere oder größere Dynamik und die Dauer wesentliche Einteilungsgesichtspunke, hinsichtlich des Affekts außerdem die offensichtliche und auch subjektiv wahrgenommene Beteiligung körperlicher (vegetativer, motorischer) Begleiterscheinungen. Die minutiöse Klassifikation dieser Zustände und ihre psychologisch-phänomenologische Beschreibung erreichten damals eine sprachliche Differenzierung (vgl. Lersch, 1970), die heute weitgehend vergessen zu sein scheint bzw. nicht mehr erreicht wird. Aus der angloamerikanischen Literatur wurde in neurer Zeit zunehmend der Begriff Emotion übernommen, unter dem meist Gefühl und Affekt zusammengefasst werden.

Die traditionellen Unterscheidungsversuche von Stimmung, Gefühl und Affekt, scheinen auf der phänomenalen Ebene einiges für sich zu haben, doch sind operationale Definitionen dieser Kategorien kaum auszumachen. Deskriptiv wäre heute eher an bestimmte *Kontinua* zu denken: die zeitliche Dauer und Häufigkeit sowie die relative Erlebnisintensität, die Dynamik (Prozessgestalt), der Grad der Kopplung mit Körperwahrnehmungen, mit objektiven behavioralen und mit vegetativen Veränderungen und letztlich mit neurobiologischen Systemfunktionen. Näher betrachtet führen alle diese Fragestellungen in schwierige Forschungsgebiete, in denen sich gegenwärtig leichter die Inkonsistenzen und Widersprüche als die konvergenten und replizierten Ergebnisse aufzeigen lassen.

In der Tendenz, die phänomenale Vielfalt auf sehr wenige, oft sogar nur eine oder zwei Dimensionen zu reduzieren, trafen sich verhaltenswissenschaftliche und einige der faktorenanalytischen Arbeitsansätze, z.B. Thayers Activation – Deactivation Adjective Checklist, Langs Reduktion des "affective space" auf Pleasantness und Arousal im Zusammenhang mit der emotionalen Bewertung von Bildern oder Watson's Postive Affect - Negative Affect Scales.

Ein Strukturvergleich von vier Instrumenten wurde von Yik, Russell & Barrett, (1999) berichtet: Russell*'s* (1980) Circumplex, Watson und Tellegen's PANAS, Thayer's zwei Arousal-Dimensionen (tense und energetic arousal) und Larsen und E. Diener's Kombinationen von Pleasantness und Activation. Die Autoren berichten, dass die Befunde einen "großen Überlappungsbereich" aufweisen, und auf zwei bipolare Dimensionen zurückgeführt werden können. Solche bloß formalen, internen Analysen können naturgemäß nur etwas über die gemeinsame Varianz aussagen; Nachweise der einheitlichen oder der differentiellen Validität, Testökonomie usw. verlangen andere Untersuchungen. Interessant wäre auch der Vergleich der jeweiligen Itemlisten mit einer entsprechenden Ein-Item-Skala von vergleichbarer Spannweite. Müssten sich nicht mit zunehmender Homogenität/Konsistenz der Items die beiden Modi bzw. Instrumente immer ähnlicher werden?

Kritisch ist zu fragen, für welche psychologischen Fragestellungen solche reduzierten Dimensionalitäten nützlich sein könnten und für welche Fragestellungen ein differenziertes Beschreibungssystem vorzuziehen oder unverzichtbar ist. Für verschiedene Assessmentaufgaben werden verschiedene Methoden zweckmäßig sein. Bemerkenswert sind Befunde und Hinweise von Zelenski und Larsen (2000), dass vielleicht kategoriale Konzepte eher geeignet sein könnten Zustandsänderungen zu beschreiben und dimensionale Konzepte eher für überdauernde Eigenschaften. Die beiden Perspektiven,

d.h. die dimensionsanalytisch-reduktionistische und die kategorial vielfältig-differenzierende, sind auch in der heutigen Emotionsforschung gut zu erkennen, wenn über die Auswahl der Instrumente (Items bzw. der Skalen) für das ambulante Assessment von Zustandsänderungen entschieden wird. Diese Vorentscheidungen sind keineswegs selbstverständlich und können u.U. unerwünschte Konsequenzen für die Gesamt-Validität einer Untersuchung und die Generalisierbarkeit der Ergebnisse haben – folglich müssen diese Auswahlentscheidungen inhaltlich und methodisch diskutiert und gerechtfertigt werden.

Operationalisierungs-Entscheidungen
In empirisch-psychologischen Untersuchungen werden theoretische Begriffe durch bestimmte Indikatoren (Referenten) dieses Konstrukts operationalisiert. Im konkreten Fall wird ein einzelnes Item (Deskriptor eines Zustandes) oder ein Skalenwert aus mehreren Facetten eines Konstrukts verwendet. Ob diese Methode als adäquat (die Konstruktbedeutung intensional erschöpfend) bzw. als der "richtige Weg zum Ziel" gelten kann, müssen die Untersucher und die Scientific Community entscheiden.

Einige Untersucher verwenden ausschließlich Itemwerte, andere nur Skalenwerte, wobei diese Präferenzen häufig nicht näher begründet sind. Seltener werden beide Ansätze verfolgt. In vielen Untersuchungen wird die Auswahl der Items weitgehend durch die Fragestellung vorgegeben sein, z.B. bestimmte Emotionen oder bestimmte klinische Symptome und deren Kontextbedingungen. Andere Untersuchungen werden eher ein breiteres, möglichst repräsentatives Bild der Zustandsänderungen zu gewinnen versuchen oder zumindest einige solcher Items einfügen, um Trends des Allgemeinbefindens u.a. Aspekte zu erkennen. Oder es werden systematisch-vergleichend Profile des Befindens, der Stimmungen und Emotionen erhoben.

Ausgewählte Verfahren (Übersicht)
Als (Fragebogen-) Skala wird in der Regel eine testmethodisch konstruierte Zusammenstellung von Items bezeichnet. Der Skalenwert wird durch Summation über die Itemwerte gebildet.

Die Konstruktion erfolgt:
(1) konventionell durch Itemanalyse und Itemselektion anhand von Schwierigkeitsindizes und Trennschärfeindizes der Items sowie Verteilungsform und Reliabilitätsprüfung der Skalenwerte;
(2) faktorenanalytisch zur Klärung der Dimensionalität und Reduktion auf einige Markier-Items der Faktoren;
(3) selten durch Clusteranalysen, d.h. weniger voraussetzungsreich als im Verfahren der Faktorenanalyse, oder durch Multidimensionale Skalierung, Circumplex-Skalierung u.a. Ansätze.

Das Niveau der testmethodischen Konstruktion und Evaluation in diesem Methodenbereich scheint vielfach schwächer ausgebildet zu sein scheint als etwa hinsichtlich der Fähigkeits-Tests oder Persönlichkeits-Fragebogen, für die ein Kanon von testmethodischen Beurteilungskriterien und testkritischen Rezensionen weitgehend üblich geworden ist (Qualitätskontrolle).

Die verschiedenen Formen von Verfahren der Selbsteinstufung können u.a. unterteilt werden nach:
(1) inhaltlichen Schwerpunkten,
(2) Anzahl der Items und Dimensionen,
(3) Skalentyp bzw. Skalierungsverfahren,
(4) Antwortmodus,

und bewertet werden nach
(5) dem Umfang der empirischen Basis und testmethodischen Arbeiten gemäß den Gütekriterien für psychologische Tests und den speziellen Anforderungen beim Assessment von "interindividuellen Unterschieden der intraindividuellen Variabilität".

Eindimensionale Einstufungsskalen
z.B. Ratingskalen wie Angst (Furcht-) Thermometer, Schmerz-Thermometer, erlebte körperliche Anstrengung (exertion) (Borg, 1970); arousal und fatigue Levi (1972), Skala Allgemeiner zentraler Aktiviertheit (Bartenwerfer, 1963), Befindlichkeit Bf-S (von Zerssen, 1976). Weit verbreitet sind Adaptationen der State-Trait-Anxiety Scale STAI (Laux, Glanzmann, Schaffner & Spielberger, 1981).

Die Tabelle 2 zeigt, dass es eine Anzahl von Skalen mit ähnlichem Gültigkeitsbereich, d.h. inhaltlichem Zielbereich, gibt. Die ersten in Deutschland auf breiter empirischer Basis konstruierten Verfahren sind die Eigenschaftswörter-Liste EWL und die Skalen zur Selbsteinschätzung der aktuellen Stimmung SKAS (= SES). Mehrdimensionale Verfahren sind in der angloamerikanischen und deutschen Literatur zahlreich publiziert worden. Keine hat sich jedoch zu einem Standard oder einer Referenzmethode entwickelt, aus der nach Bedarf einzelne Subskalen entnommen werden könnten.

Tabelle 2: Typische Skalen zur Beschreibung von Befinden, Stimmungen, Emotionen (Auswahl), ohne Symptomlisten und körperliche Beschwerden

Abkürzung	Name	Items	Skalen	Autor(en)	Jahr
MACL	Mood Adjective Check List	145	12	Nowlis & Nowlis	1956, 1970
MAACL	Multiple Affect Adjective Checklist	132	3	Zuckerman	1965, 1976
MARS	Manifest Affect Rating Scale	87	4	Jacobs	1966
AD-ACL	Activation-Deactivation Checklist	28	4	Thayer	1967, 1971
EWL	Eigenschaftswörterliste	161	15	Janke & Debus	1964, 1978
POMS	Profile of Mood States	65	6	Mc Nair et al.	1971
SKAS (SES)	Skalen zur Selbsteinschätzung der aktuellen Stimmung	2 x 42	6	Hampel	1972, 1977
DES (DAS)	Differential Emotion Scales	30	10	Izard et al. Merten & Krause	1974, 1982 1993
EZ	Eigenzustandsskala	40	8	Nitsch	1974, 1976
EMI	Emotionalitätsinventar	70	7	Ullrich de Muynck u.a.	1975, 1977
BfS	Befindlichkeitsskala	2 x 28	1	Zerssen	1976
STAI	State-Trait-Anxiety Inventar	2x20	2	Spielberger et al.	1970
AD–ACL	Activation-Deactivation Checklist	28	4	Thayer	1967, 1970
PANAS	Positive and negative Affect Scales	20	2	Watson u.a.	1985, 1988,
EMO 16	Emotionsskalen	16	16	Schmidt-Atzert & Hüppe	1996

Anmerkungen: Literaturhinweise siehe Übersichten bei Brähler, Schumacher & Strauß, 2002; Fahrenberg, 1983; Larsen & Prizmic-Larsen, 2005; Testzentrale Hogrefe Verlag - Testkatalog 2006/07; Westhoff, 1993).

Die Existenz der zahlreichen und – zumindest auf den ersten Blick – konkurrierenden Konzeptionen (in einem vergleichsweise sehr kleinen Ausschnitt der Forschung über emotionale Befindlichkeit) spricht für sich und gegen die Hoffnung auf eine absehbare, überzeugende Konvergenz der Untersuchungsmethoden.

Nicht zu übersehen ist, dass die Mehrzahl der Instrumente im angloamerikanischen Bereich geschaffen wurde. In praktischer Hinsicht und in diplomatischer Hinsicht auf mögliche Publikationen in US-Journals scheint dieser Sachverhalt viele deutsche Autoren zu motivieren, sich anzupassen und solche Skalen zu übernehmen. Eine schlichte Übersetzung ohne vollständige deutschsprachige Nach-Konstruktion und testmethodische Überprüfung ist hier jedoch ebenso zu kritisieren wie im Falle anderer Fragebogen-Importe. – Der verhältnismäßig rasche Wechsel solcher Präferenzen für das eine oder das andere Instrument lässt durchaus etwas "modische" bzw. publikations- und zitierbedingte Eigenheiten erkennen. Aus dem bisherigen Verlauf ist es wahrscheinlich, dass in einigen Jahren eine andere Skala populär sein wird. Es liegt wohl in der Natur der subjektiven Phänomene, dass ein völlig überzeugendes und bleibendes Beschreibungssystem, ähnlich einer naturwissenschaftlichen oder behavioralen Messvorschrift nicht erwartet werden kann.

5 Einstufung der Befindlichkeit – Einzelne Items oder Skalen wie AD-ACL und PANAS?

Beim ambulanten Assessment der Befindlichkeit existieren bemerkenswerte Unterschiede der Methodik: Einige Untersucher wählen einzelne Items aus, andere Untersucher bevorzugen aus mehreren Items zusammengesetzte Skalen. Dabei spielen oft aktuelle amerikanische Publikationen eine Rolle. Auswahlprobleme dieser Art stellen sich natürlich nicht für alle Projekte, denn die Items, Symptome u.a. Details sind oft durch die inhaltliche Zielsetzung festgelegt.

Vorteile und Nachteile von Ein-Item-Skalen und Skalen

Für die Verwendung einzelner Items ("Ein-Item-Skala") spricht: (1) mit wenigen Items ist ein relativ breiter Bereich von Befindensweisen, Stimmungen und Emotionen zu beschreiben und (2) die Auswahl ist leicht der jeweiligen Fragestellung anzupassen.

Deutsche Item-Listen zum ambulanten Assessment der Befindlichkeit wurden publiziert u.a. von Ebner (2004), Fahrenberg et al. (1984, 2002a, 2002b), Heger (1990), Jain (1995), Käppler (1994**),** Kinne (1997), Kubiak (2003), Myrtek, Foerster & Brügner (2001), Pawlik & Buse (1982), Perrez & Reicherts (1989), Stiglmayr (2003), Triemer (2003), Perrez, Schoebi & Wilhelm (2004).

Nachteilig ist, dass (1) die Ein-Item-Skalen mit wenigen Stufen eine dementsprechend geringe Differenzierungsmöglichkeit bieten, häufig auch eine schiefe Verteilung der Werte aufweisen und (2) die konventionelle Schätzung der Reliabilität (Konsistenz, Item-Homogenität) entfällt.

Zumindest der Nachteil geringer Varianz kann jedoch abgeschwächt werden, wenn ein vielstufiges Format, z.B. eine geeignete visuelle Analog-Skala mit 21 Stufen verwendet wird, und dieses Format beim ohnehin zweckmäßigen Skalierungstraining zu Untersuchungsbeginn eingeführt wird.

Für die Verwendung von Skalenwerten statt Itemwerten spricht: (1) mit Skalen von z.B. 10 Items ist eine numerisch größere Varianz innerhalb und zwischen Personen (Diskrimination) zu erreichen; (2) die Verteilungsform kann durch die Auswahl von Items mit unterschiedlichen Mittelwerten ("Schwierigkeitsindices") beeinflusst werden, (3) bei Skalen kann die innere Konsistenz einer Itemliste berechnet und damit ein Koeffizient der lokalen Reliabilität gewonnen werden.

Nachteile von Skalenwerten sind: (1) im Hinblick auf die zumutbare Länge eines Selbstberichts bedeutet die Präferenz für eine (oder zwei?) Skalen testökonomisch den Verzicht auf wichtige andere Aspekte der Befindlichkeit; (2) für jede Skala wird eine Anzahl inhaltlich recht ähnlicher Items benötigt, was für den Befragten bei wiederholten Einstufungen besonders lästig wird, ggf. mit Folgen für die Akzeptanz und für die methodenbedingte Reaktivität; (3) das zugrundeliegende testtheoretische Postulat, dass die Items der Skala inhaltlich homogene, unabhängige Parallelmessungen darstellen, ist im Bereich der Befindlichkeit besonders fragwürdig, (4) die Annahme, dass sich die Item-Response-Funktionen der Items einer Skala synchron (konsistent) über die Zeit verhält ist in der Regel ungeprüft, und (5) die einfache Addition der Itemwerte (Ordinaldaten) zu metrischen Skalenwerten bleibt fragwürdig.

Falls ein Konsistenzkoeffizient, z.B. Cronbachs Alpha-Koeffizient, als Maß der Reliabilität gewertet werden soll, muss behauptet werden können, dass alle Items der Skala parallele Messungen des Konstrukts bilden (siehe oben). Deswegen ist eine schematische Anwendung der Konsistenzanalyse problematisch. Notwendig bleibt die genaue Evaluation, welche Facetten des Konstrukts repräsentiert sind, und welche Zusammenhänge zwischen Itemzahl, Redundanz und Testökonomie bestehen. Darüber hinaus wären für jeden Skalentyp bzw. für jedes Instrument Validitätshinweise durch Kriterienkorrelationen und – anspruchsvoller –als Entscheidungsnutzen in einer praktischen Assessmentaufgabe wichtig.

Viele Untersucher haben sich folglich für eine fragestellungs-nahe Auswahl von Einzel-Items entschieden, d.h. gegen lange, u.U. pseudohomogene oder z.T. redundante Skalen.

Mangels Standardisierung der Methodik wird ein Vergleich der Forschungsergebnisse aus verschiedenen Arbeitsgruppen schwierig bleiben. Voraussichtlich wird es auf diesem Gebiet in absehbarer Zeit noch keine Standardmethoden geben. Umso wertvoller wird forschungsstrategisch die konsequente Replikation wichtiger Befunde wenigstens innerhalb einer Arbeitsgruppe sein.

AD-ACL oder PANAS und andere "Super-Skalen"?

Die Mehrzahl der publizierten Stimmungsskalen ist mehrdimensional konzipiert. Wegen der großen Anzahl von Skalen und Items sind sie für kurzfristig wiederholte Anwendungen ungeeignet. Seit Wundts dreidimensionaler Gefühlstheorie ist wiederholt vorgeschlagen worden, die Vielfalt der subjektiven Zustände auf wenige Dimensionen (Faktoren, Basisemotionen) zu reduzieren. Verschiedentlich wurden Instrumente mit nur einer oder zwei Skalen zur Erfassung von Befindlichkeit (Stimmung) entwickelt.

Die AD-ACL Activation-Deactivation Adjective Checklist besteht aus vier Subskalen: General Activation, High Activation, General Deactivation, Deactivation-Sleep. Sie wurde u.a. für die psychophysiologische Forschung propagiert (Thayer, 1970), um korrelative Beziehungen anhand von Veränderungswerten zu untersuchen. In einer später erweiterten Konzeption unterschied Thayer (1978) zwischen der Dimension A (energetic – sleepy) und Dimension B (tense – placid, still). Die in den 70er Jahren publizierte AD-ACL scheint heute kaum noch Interesse zu finden.

Die PANAS Positive Affect – Negative Affect Scales (Watson, Clark & Tellegen, 1988) sind in den vergangenen Jahren verschiedentlich verwendet worden, auch in deutschen Adaptionen (siehe u.a. Krohne, Egloff, Kohlmann & Tausch, 1996). Die Autoren schrieben zwar ursprünglich, dass sie in den PANAS keine Konkurrenz zu den mehrdimensionalen Konzepten sehen würden, sondern einen komplementären Ansatz (Watson & Tellegen, 1985, p. 220). Durch die weitreichenden Postulate, eine "consensual structure of mood" gefunden zu haben, ja eine real existierende Struktur, wurden andere Untersucher angeregt, diese Methode zu verwenden. Deswegen wird die PANAS als Beispiel ausgewählt, um typische testmethodische Probleme zu diskutieren und auf gravierende Mängel aufmerksam zu machen.

Watson und Tellegen (1985) hatten die Absicht, eine möglichst sparsame und allgemein zustimmungsfähige Beschreibung selbstberichteter und fremdbeobachteter Affekte zu geben: Positive Affect PA und Negative Affect NA. Mit den erhaltenen Grunddimensionen sei ein Konsens in der widersprüchlichen Literatur zu erreichen. Andere Dimensionen wie Arousal (Activation) und Potency (Dominance u.a.) wurden von Ihnen nur oberflächlich erwähnt. Die Autoren behaupteten, durch Reanalyse verschiedener Datensätze belegen zu können, dass die Dimensionen PA und NA die faktorenanalytisch dominierenden "Sekundärfaktoren" sind. Angeblich sind die beiden Dimensionen (Skalen) unabhängig voneinander.

Die Originalarbeit ist in theoretischer Hinsicht bemerkenswert unreflektiert und stützt sich vor allem auf faktorenanalytisch-technische Argumente. Die Autoren scheinen jedoch nicht die inneren Abhängigkeiten ihrer methodischen Vorentscheidungen und statistischen Resultate gesehen zu haben. Testmethodische Probleme und Mängel sind:

- Der primäre Forschungsansatz der Autoren liess nicht erkennen, dass es grundsätzlich um Veränderungsmessung im Unterschied zu Eigenschaftsdimensionen von Persönlichkeits-Fragebogen geht. Für die Skalenkonstruktion wurde nicht die intraindividuellen Varianz der Zustandsänderungen verwendet (– diese Adäquatheitsfrage wird allerdings auch sonst selten gestellt).
- Der grundsätzliche Bias hinsichtlich des primären Itempools der eigenen bzw. der nur sehr selektiv zitierten Studien wird nicht erkannt, d.h. das Fehlen einer, allerdings empirisch kaum zu erreichenden Zufallsstichprobe aus dem Universum der Befindlichkeits-Deskriptoren, welche allein zu einer gültigen Inventarisierung der Grunddimensionen führen könnte.
- Die Daten stammen aus einfachen Papier-und-Bleistift-Untersuchungen und sind folglich sehr viel zweifelhafter als die Daten computer-unterstützter Untersuchungen, die in den 80er Jahren bereits möglich gewesen wären (siehe Pawlik & Buse, 1982, 1996).

- Die teststatistischen und faktorenanalytischen Konsequenzen der unterschiedlichen Item- (Ko-) Varianzen und der Verteilungsform der Itemwerte (inter- und vor allem auch intra-individuell), speziell auch bei den NA-Items, werden zu wenig beachtet.
- Die gerade bei der Beschreibung subjektiver, u.U. schnell veränderlicher Zustände wesentlichen Fragen und methodischen Probleme der Veränderungsmessung sowie der Skalierung bzw. Skalenqualität werden nicht erörtert.
- Es wird keine statistische Aufgliederung der wesentlichen Varianzkomponenten vorgenommen: zwischen Personen, innerhalb Personen, innerhalb und zwischen Tagen, und Interaktionen, entweder durch Kovarianzzerlegung, Multi-Level-Analysen oder Modellierung nach einem Latent Trait/State-Konzept.
- Der technisch bedingte, triviale Effekt von hochkorrelierten bzw. redundanten Items auf Skalenkonstruktion und Faktorenanalyse wurde übersehen: einzelne Item-Dubletten und Tripletts können aufgrund des impliziten Maximierungsprozesses zu einem gewichtigen Struktur-Bias führen.
- Im Formalismus der Faktorenanalyse und in den orthogonalen oder schiefwinkligen Rotationen wird hier nicht ein mögliches unter mehreren, mathematisch-statistisch gleichwertigen Beschreibungssystemen verstanden, sondern den PANAS-Dimensionen wird eine besondere Realität zugesprochen. Wenn die Autoren solche Dimensionsanalysen mit der faktorenanalytischen Intelligenzforschung und dem g-Faktor vergleichen, ist dies aus mehreren Gründen schief und lässt einen Reduktionismus erkennen, der im Bereich der emotionalen Befindlichkeit besonders problematisch ist.
- Die Testökonomie der langen Item-Liste von 20 Items wird nicht unter den Gesichtspunkt "Validität pro Zeiteinheit" reflektiert. Auf welche anderen Informationen müssen die Untersucher deswegen verzichten, weil noch mehr Items für wiederholte Selbstberichte kaum zumutbar wären? Pragmatisch wurde nicht überlegt: wäre es nicht unvergleichlich viel einfacher, den Befragten für PA und NA je eine visuelle Analogskala (mit z.B. 10 Stufen) vorzulegen, um PA und NA einfach, voraussetzungsärmer und wesentlich schneller zu erfassen? Den Nachweis, dass die PANAS im Hinblick auf relevante Kriterien mehr inkrementelle Varianzaufklärung leisten als eine simple Ein-Item-Skala vom VAS-Typ, müsste noch erbracht werden.
- Es fehlen Überlegungen und systematische Ergebnisse zu den verschiedenen Aspekten der lokalen und aggregierten Reliabilität und zur Kriterienvalidität, denn die faktorielle Konstruktvalidität hat ja zunächst nur formale Bedeutung.
- Die Autoren haben in der Originalpublikation keine expliziten Assessmentstrategien entwickelt oder auch nur referiert, um an typischen Anwendungsbeispiele zu zeigen, wozu diese auf je 10 weitgehend homogene Items beschränkten PA- und NA-Scales gut sein sollen.
- Kritisch ist zu fragen, für welche psychologischen Fragestellungen solche reduzierten Dimensionalitäten nützlich sein könnten und für welche Fragestellungen ein differenziertes Beschreibungssystem vorzuziehen oder unverzichtbar ist. Für verschiedene Assessmentaufgaben werden auch verschiedene Methoden zweckmäßig sein.

In dem neueren Beitrag (Watson & Clark, 1997) versuchen die Autoren einige ihrer problematischen Schritte zu rechtfertigen und für die Anwendung ihrer Skalen zu werben. Doch nun wird ausdrücklich von einem hierarchischen System und multiplen spezifischen Affekten gesprochen. Nach wie vor wird eine Kernfrage nicht erkannt: Strukturstabilität ist etwas anderes als Änderungssensitivität, doch wird jene weder erläutert noch untersucht. Über die faktorielle Konstruktvalidität hinaus werden keinerlei eigene Beiträge zur empirischen, geschweige denn für eine überlegene, empirische Kriterien-Validität der PANAS mitgeteilt.

Für die PA und NA-Skalen werden Konsistenzen von .86. bis .90 bzw. .84 bis .87 (je 10 Items) und Faktor-Korrelationen von .95 und .93 angegeben ohne Seitenblick auf die damit empirisch wahrscheinlich festgestellte Redundanz vieler Items (höherer Homogenität zuliebe). Die behauptete Unabhängigkeit beider Skalen, die ja faktorenanalytisch gewollt und erzwungen war, variiert offenbar in Abhängigkeit von der Länge des subjektiv beurteilten Zeitintervalls und beträgt bei dem einzigen größeren Within-Subject-Datensatz dieser Autoren immerhin .30 (momentan) und .34 (für den Tag). Auch neuere Untersuchungen sprechen gegen die behauptete Unabhängigkeit (Schmukle, Egloff & Burns, 2002; Zautra, Berkhof & Nicolson, 2002); sofern es sich nur um Papier-und Bleistift-Daten handelt, sind die Befunde ohnehin zweifelhaft.

Auf die Rechtfertigungsversuche, weshalb in den PANAS die wichtigen Komponenten Müdigkeit sowie Gelassenheit fehlen oder weshalb die Autoren die Komponenten Freude aus PA und Traurigkeit aus NA ausklammerten, braucht hier nicht weiter eingegangen zu werden: sie passten eben nicht in das beabsichtigte Zweier-Schema ("these terms failed to enhance the psychometric properties of the PANAS scales" p. 277). Inzwischen gibt es PANAS-X mit Sadness Scale und PANAS Plus, mit PANAS Happiness Scale, usw.

Die Darstellung dieser PANAS erfolgt in weiten Bereichen in einem zirkulären bzw. sehr einseitigen Zitationsstil. Eigenartig ist auch der Anspruch, dass die aus den amerikanischen Datensätzen entwickelten Fragebogen kultur-unabhängig gelten. In diesem Anspruch, "die" Grunddimensionen aufgedeckt zu haben, real existierende und deswegen weitgehend für alle Menschen gültig, entspricht der PANAS-Ansatz durchaus den Ansprüchen des NEO-FFI Persönlichkeitsfragebogens von Costa & Mc Crae, der dementsprechend von Watson & Tellegen als Vorbild zitiert wird.

Beide Postulate sind grundsätzlich zu relativieren, und es ist sogar ärgerlich, wenn hier aufgrund unzureichender, einseitiger Empirie eine kultur-unabhängige Gültigkeit (als "Universalien") postuliert wird (Mc Crae & Costa, 1997). Während der vergangenen Jahre wurden mehrere solcher sog. interkulturellen Studien nach diesem schlichten Schema publiziert. Es ist gewiss unzureichend, nur amerikanische Fragebogen in andere Sprachen zu übersetzen und dann hauptsächlich den Studierenden von westlich orientierten Colleges oder Universitäten vorzulegen. Angemessen wären von Grund auf eigenständige, authentische Entwicklungen und deren Vergleich miteinander (siehe die ethnologische und ethnospsychologische Kritik an der Neigung von Psychologen, Universalien zu postulieren, u.a. von Marsella, Dubanoski, Hamada & Morse, 2000).

Inzwischen ist jedoch der Proliferationsprozess bei beiden "universellen" Verfahren fortgeschritten: es gibt Varianten, es gibt Adaptationen mit weniger Items, sogar wieder mit bipolaren Items, und es werden nachträglich einzelne Subskalen gebildet – was angesichts der Konstruktionsgeschichte beider Instrumente und bei dem beschränkten Itempool besonders überraschend ist.

Zusammenfassung der testkritischen Evaluation.

Dass die positive Bewertung und die negative Bewertung bei der Auskunft über die eigenen Befindlichkeit eine wichtige Rolle spielen, ist trivial und gewiss nicht neu. Seit Wundt wurde oft eine bipolare Valenz-Dimension Angenehm-Unangenehm (Lust – Unlust) postuliert. Aus semantischen Gründen und insbesondere seit unipolare Items bevorzugt werden, ist diese Perspektive faktorenanalytisch in zwei Sub-Skalen aufgespalten (wesentlich früher bereits in EWL und SKAS). Dass die Befragten generell eher eine positive Stimmung angeben, war ebenso bekannt. In vieler Hinsicht enthält also der PANAS-Ansatz eher Rückschritte als methodische Fortschritte im Vergleich zum Stand der testmethodischen und der emotionstheoretischen Literatur, ganz abgesehen davon, das dieser auf PA – NA beschränkte Reduktionismus weit entfernt ist von den in der heutigen neurowissenschaftlich-emotionstheoretischen Literatur diskutierten Multi-System-Konzepten (siehe Peper, 2006).

6 Statistische Konzepte

Die statistische Auswertung kann sich im einfachsten Fall univariat auf die Prüfung von Unterschieden der zentralen Tendenz (Median, Mittelwert) aktueller Einstufungen beziehen. In der Regel wird die Änderung der berichteten Zustände interessieren, innerhalb eines Tages und zwischen Tagen. Dabei kann zwischen Niveau (Level), Streuung (Scatter) und Verlaufsgestalt (Shape) der erhaltenen Profile unterschieden werden, bzw. der zeit- (situations-) abhängigen Veränderung von Niveau, Streuung und Gestalt. Speziellere Analysen, auch multivariate Verfahren, sind erforderlich, um korrelierte Zustandsänderungen oder Muster der Veränderung, d.h. besondere Prozess- (Verlaufs-)gestalten zu erfassen.
Varianzanalysen, Kovarianzzerlegungen, Autokorrelationen, Moving Average- und ARIMA-Ansätze zur Effektprüfung, Multi-Level-Analysen (hierarchische lineare Modelle bzw. Regressionsmodelle), Strukturgleichungs-Konzepte und andere Modellierungen sind parametrische Verfahren und setzten durch ihre Rechenoperationen eine Intervallskala voraus und machen Annahmen über die Wahrscheinlichkeitsverteilung der untersuchten Variablen – im Unterschied zu den nicht-parametrischen Tests.

Methoden für Ordinaldaten
Mit Ordinaldaten sind vergleichsweise nur begrenzte Analysen möglich. Das Gebiet der "verteilungsfreien Methoden" hat sich aber zweifellos stark entwickelt (Bortz et al., 2000) und für eine Reihe von statistischen Fragestellungen existieren heute geeignete Verfahren (vgl. ALMO, SPSS u.a. Programmpakete). Ein Handikap der Daten aus Selbsteinstufungen ist, dass wegen der geringen Spannweite (Range) der üblichen Skalen sehr oft gleiche Ränge vergeben werden und deswegen Rangaufteilungen notwendig sind.

Erwähnenswert sind hier:

- Zufälligkeit/Regelmäßigkeit/Stationarität einer Zeitreihe (Abfolge), Omnibustests für Zufälligkeit der Abfolge sowie Häufigkeitsverteilungen;
- Homogenität von Abfolgen auch im Vergleich mehrerer unabhängiger oder abhängiger Stichproben;
- Verteilungsfreie Sequenzanalyse, um bei binomial verteilten Merkmalen eine möglichst frühzeitige Entscheidung über die zu prüfende Hypothese zu erreichen (Vorzeichentests für die Zufälligkeit einer Abfolge, Reproduzierbarkeit eines dichotomen Merkmals);
- Prüfung auf monotonen Trend oder auf andere Trendhypothesen;
- Verteilung von einzelnen Ereignissen über Abschnitte einer Zeitreihe, Vergleich mehrerer zeitlicher Verteilungen;
- Zusammenhänge von zeitsynchron erhobenen Abfolgen (Konkomitanzen, Mitveränderungen) durch Rangkorrelationen oder Kontingenzkoeffizienten, wobei Verzögerungen (lags) berücksichtigt werden können, Beschreibung multipler Konkomitanzen durch Konkordanzkoeffizienten;
- eine Vielzahl von statistischen Tests für Unterschiedshypothesen bei unabhängigen und abhängigen Stichproben;
- Korrespondenzanalyse u.a. Verfahren für nominale Daten.

Die Verfahren zur Konstruktion nicht-parametrischer Konfidenzintervalle, die sich auf die Jackknife- und Bootstrap-Methodik stützen wurden bereits im Abschnitt 3 erwähnt.

Auch latent class state-trait models wurden für Zeitreihen nominaler Daten entwickelt, um die Varianzquellen Person, Situation und Residuum zu separieren und den Effekt der Situationen im Hinblick auf Stabilität und Variabilität der Befindlichkeit zu evaluieren (Eid & Langeheine, 2003). Es wird jedoch eine sehr große Anzahl von Personen benötigt, um solche Modelle zu testen – mehr Daten als die meisten Untersuchungen mit ambulantem Assessment haben werden.

Modellierungen
Zur formalen Beschreibung von Verläufen, z.B. Tagesverläufen, können Polynome, (Legendre-Polynome), angepasst werden. Außer diesem induktiv-deskriptiven Verfahren ist auch deduktiv die

Anpassung an bestimmte Funktionen, z.B. die e-Funtion (Wachstumsfunktion) möglich, falls es dafür theoretische Gründe geben sollte. Bei praktischen Anwendungsversuchen stellen sich jedoch Schwierigkeiten heraus: Es sollten hinreichend viele Datenpunkte (Intervalldaten) vorhanden sein, d.h. möglichst 10 oder mehr, um einen Verlauf anpassen zu können. Es können erhebliche Approximationsfehler auftreten (vgl. Becker, 1992). Weder die Parameterisierung durch Polynome noch die Anpassung mit e-Funktionen liefern einen einzelnen Index: Es sind zwei oder mehrere Parameter, so dass ein Vergleich zwischen Individuen nicht einfach durchzuführen ist. Der Nutzen liegt folglich eher in der Beschreibung genereller, über Personen gemittelter Verläufe oder in der Exploration möglicher Untergruppen (Verlaufstypen, Reaktionsverläufe), nicht in der Beschreibung individueller Verlaufsgestalten.

Für die Modellierung dynamischer Systeme gibt es eine Anzahl von parametrischen Methoden. Solche Prozessgestalten im Übergang zwischen Gleichgewichtszuständen sind durch induktive und deduktive Modellierung zu beschreiben. Tschacher (1997a, 1997b) erläutert an Beispielen solche Methoden: die lineare induktive Modellierung, Autokorrelation, Arima-Modellierung, Fourier-Analyse, multivariate lineare Modellierung. Auch einige der Grenzen der gegenwärtigen Verfahren werden erwähnt: Die Zeitreihen psychologischer Daten sind in der Regel nur kurz oder sehr kurz (< 200 oder sogar < 100) Messpunkte, die Vorbehalte hinsichtlich der Daten, die nicht Intervallskalenniveau (gleiche Intervalle) haben, sondern nur aus subjektiven Rating-Skalen stammen, und das Fehlen geeigneter nicht-linearer Modelle.

Auf Hinweise zu Anwendungsbeispielen, z.B. zu Multi-Level-Analysen und deren Möglichkeiten und Grenzen, wird hier verzichtet. (Einige Beispiele und Literaturhinweise sind in dem Review von Publikationen (2000 - 2005) zum Ambulanten Assessment enthalten (Fahrenberg, 2006).

Schlussfolgerungen

Die Bewertung dieser Verfahren ist nicht ohne weiteres möglich. Dies liegt nicht nur an den speziellen Voraussetzungen und z. T. an dem Schwierigkeitsgrad der Analysen, sondern auch an den fehlenden Vergleichsmöglichkeiten. Was hätte im konkreten Fall eine konventionelle Auswertung, z. B. ein Messwiederholungsplan bzw. eine multiple Regression-Korrelation erbringen können? Was wäre mit verteilungsfreien statistischen Verfahren zu erreichen, und welche Schlussfolgerungen hätten nicht bereits aus einer einfachen graphischen Darstellung (Scatter-Plot der Daten und der sukzessiven Differenzen und Trends) gezogen werden können? In wie weit sind diese Verfahren auch für induktive Zwecke ergiebig oder nutzen sie hauptsächlich zum statistischen Testen von Interventionseffekten bzw. von einigermaßen regelmäßig wiederkehrenden Prozessgestalten?

Nach einer Phase optimistischer Erwartungen hinsichtlich der propagierten Algorithmen zur sparsamen Datenreduktion auf Grundeigenschaften (Faktorenanalyse), zur automatischen Klassifikation, Clusteranalyse, ARIMA-Modellen, Rasch-Skalierung, Strukturgleichungen usw. sind die jeweiligen Voraussetzungen und die oft sehr großen praktischen Schwierigkeiten solcher Verfahren besser erkannt worden. In empirischen Projekten werden sehr oft, vielleicht sogar in der Mehrzahl der Untersuchungen, die Anforderungen an die notwendige Anzahl von Personen, Messpunkten und Wiederholungen, Skalenart, Existenz sehr weniger und prägnanter Typen ("Dichtezentren im multivariaten Raum") oder gut reproduzierbarer Verlaufsgestalten usw. nicht erreicht. Deshalb ist nicht selten eine Ernüchterung eingetreten. Das Instrumentarium ist jedoch reichhaltiger geworden, eröffnet mehr Auswertungsstrategien, und lässt andererseits auch erkennen, wie begrenzt jeweils der praktische Gebrauchswert eines einzelnen Verfahrens ist.

Abschließende Thesen

Bei kritischer testmethodischer Evaluation können die PANAS oder ähnliche Skalen für das computergestützte ambulante Assessment nicht empfohlen werden. Standardisierung der Methodik ist gewiss erstrebenswert, und es gibt den verständlichen Wunsch, in der Untersuchungsmethodik "international anschlussfähig" zu sein. Doch die testmethodischen und testökonomischen Einwände sind offenkundig. Deshalb wird die inhaltlich und teststatistisch begründete Auswahl einzelner Items (Ein-Item-Skalen) für die meisten Fragestellungnen zweckmäßiger sein.

Offensichtlich müsste, um vielleicht eine Standardmethodik für Selbsteinstufungen der Befindlichkeit zu erreichen, eine neue Konstruktion auf der primären Basis deutschsprachiger Items mit ihrer intra-individuellen Variabilität unternommen werden. Durch den systematischen Vergleich der querschnitt-lichen und der längsschnittlichen Befunde hinsichtlich Muster und Dimensionen, außerdem durch Beschreibung der Änderungssensitivität von Items sowie durch Vergleich der direkten gegenüber der indirekten Beurteilung von Veränderungen wären Fortschritte dieser Methodologie möglich.

Die computer-unterstützte Methodik ist hier durch ihre ökologische und technische Datenqualität hier den bisherigen Skalen-Konstruktionen, die auf der traditionellen und unsicheren Papier-und-Bleistift-Methodik beruhen überlegen. Die Absicht, psychologisch wichtige Hypothesen über die individuelle Befindlichkeit (Gruppenunterschiede, Verläufe, Prozessgestalten) zu prüfen, rechtfertigt gründlichere Methodenstudien und innovative Ansätze.

Die Entwicklung von Prozessanalysen in der differentiellen Psychologie ist eine wichtige Aufgabe. Das alltagsnahe Assessment psychologischer und physiologischer Veränderungen wird künftig noch mehr zu dieser Forschung beitragen können.

(Anmerkung 1) Assessment-Modelle

Typ des Konstrukts: Ist das Konstrukt durch die Variation zwischen Personen, zwischen Settings/Situationen, Variablen oder Kombinationen dieser Modi definiert?
Operationalisierung: In welchem Modus (Person, Setting/Situation, Variablen) manifestiert sich das Konstrukt, d.h. in welchem Bereich sollen die Operationalisierungen stattfinden?
Anwendungsbereich: Auf welche Einheiten des Assessment beziehen sich alle Schlussfolgerungen und in welchen Modi (Person, Setting/Situation, Variablen) bestehen sie?

Tabelle 3: Assessment-Modelle (Stemmler, 1996, S. 262).

Assessment Model	Locus of Construct	Operationalization	Range of Application	Technique	Variance Analyzed
1	Subjects	Variables	Settings/Situations	R	BS
2	Settings/Situations	Variables	Subject	P	BC
3	Settings/Situations	Subjects	Variable	S	BC
4	Variables	Subjects	Settings/Situations	Q	BV
5	Variables	Settings/Situations	Subject	O	BV
6	Subjects	Settings/Situations	Variable	T	BS
7	Subjects x Settings/Situations	Variables	Subject x Settings/Situations		
8	Settings/Situations x Variables	Subjects	Settings/Situations x Variable		
9	Variables x Subjects	Settings/Situations	Variable x Subject		

Techniken nach Cattells Terminologie.
BS = Between Subjects, BC = Between Conditions (Settings/Situations), BV = Between Variable

Am bekanntesten sind Modell 1 (überdauernde Eigenschaften, "Traits", Querschnitt, R-Technik), Modell 2 (Zustände bzw. Prozesse "State", Längsschnitt, P-Technik,) und Modell 4 Klassen ähnlicher Personen (Typen, Q-Technik)

Hier sind die prozessorientierten Modelle hervorheben, bei denen entweder das Konstrukt (Modell 2, 3, 7, 8) oder die Operationalisierung (Model 5, 6) im Setting/Situations Modus definiert sind. Die Aussagen sind jeweils auf einen Modus begrenzt (Range of Application, siehe Tabelle). Wenn darüber hinaus Zusammenfassungen von Befunden (Aggregationen) und andere Formen der Generalisierung beabsichtigt sind, kann dies in methodische Schwierigkeiten führen.

Die Hinweise auf Arbeitsbeispiele beziehen sich der Einfachheit halber auf die Blutdruckforschung. In der psychophysiologischen Forschung hat die konzeptuelle und untersuchungstechnische Differenzierung solcher Musteranalysen seit langem eine besondere Rolle gespielt (s. Stemmler, 1992). Auch im Hinblick auf die neueren Multilevel-Analysen sind solche Überlegungen nützlich.

2 Zustandsänderungen einer Person; z.B. Analyse der Blutdruckänderungen innerhalb einer 24-Stunden-Registrierung.
3 Zustandsänderungen einer Personengruppe im Unterschied zu anderen Personen, z.B. alltägliches Blutdruckverhalten von Hypertonikern und Normotonikern.
7 Zustandsänderungen aufgrund einer Interaktion von Person mit speziellen Settings/Situationen, z.B. differentielle Blutdruckreaktionen in Abhängigkeit von speziellen Anforderungen/ Erlebnissen (interaktionistische Perspektive).
8 Identifikation von Personen, die in bestimmten Settings/Situationen Reaktionsprofile aufweisen, die von den "normalen" anderer Personen abweichen.
5 Identifikation von Reaktionsmustern einer Person über verschiedene Settings/Situationen (transsituationale Konsistenz bzw. individualspezifisches Reaktionsmusters), z.B. "Blutdruck-Reagierer".
6 Identifikation von solchen Settings/Situationen, die einen konsistenten Effekt auf eine bestimmte Variable haben, z.B. bestimmte Arbeitsplätze/Tätigkeitsanforderungen auf den Blutdruck.

(Anmerkung 2 als Historische Fußnote)

Hugo Münsterberg (1863-1916), der Pionier der Angewandten Psychologie, schrieb in seinem Buch "On the Witness Stand: Essays on Psychology and Crime" (1908/Nachdruck 1925, pp. 119-121) zur ambulanten Datenerhebung über Zusammenhänge zwischen der momentanen Stimmungslage und der Genauigkeit kinästhetischer Größenschätzungen:

To begin with a very simple group of processes, we may start with our ordinary movements is of the arm: does feeling influence them? I can give my reply from a little diary of mine. I kept it years ago. It was not the regulation diary -- there was no sentimentality in it, but mostly figures. Its purpose was to record the results of about twenty experiments which took about half an hour's time. [p. 120] I had the material for these little experiments always in my pocket and repeated them three or four times a day throughout several months. I fell to experimenting whenever daily life brought me into a characteristic mental state, such as emotion or interest or fatigue or anything important to the psychologist. One of these twenty experiments was the following: I attached to the bottom of my waistcoat a small instrument which allowed me to slide along an edge between thumb and fore-finger of the right hand, both outwards and inwards. Now I had trained myself to measure off in this way from memory distances of four and eight inches. Under normal conditions my hand passed through these distances with exactitude [*sic*] while the eyes were closed; the apparatus registered carefully whether I made the distance too long or too short. The results of many hundreds of these measurements went into my diary together with a description of the mood in which I was.

When I came to figure up the results after half a year's records I found a definite relation between my feelings and my arm movements. My diary indicated essentially three fundamental pairs of [p. 121] feeling in the course of time. There was pleasure and displeasure, there was excitement and depression, and there was gravity and hilarity. The figures showed that in the state of excitement both the outward and inward movements became too long, and in the state of depression both became too short; in the state of pleasure the outward movements became too long, the inward movements too short; in the state of displeasure the opposite -- the outward movements too short and the inward movements too long. In the case of gravity or hilarity no constant change in the length of the movement resulted, but the rhythm and rapidity of the action was influenced by them.

Münsterberg schrieb, dass diese Untersuchung Jahre zurückläge. Vielleicht geschah es sogar noch in seiner Zeit an der Freiburger Universität (1887-1892)?

Literaturhinweise

Allen, B. P. & Potkay, C. R. (1973). Variability of self description on a day-to-day basis. Longitudinal use of the adjective generation technique. Journal of Personality, 41, 638-652.

Allen, B. P. & Potkay, C. R. (1977). Misunderstandings in the Adjective Generation Technique (AGT). Comments on Bem's rejoinder. Journal of Personality, 45, 334-342.

Asendorpf, J. B. (1996). Die differentielle Sichtweise in der Psychologie. Göttingen: Hogrefe.

Baltisssen, R. & Boucsein, W. (1987). Vergleichende Untersuchungen zur Zustandsskalierung und Veränderungsskalierung bei der Beurteilung subjektiver Stresswirkungen in psychophysiologischen Experimenten. Zeitschrift für Differentielle und Diagnostische Psychologie, 8, 1-23.

Barrett, L. F. & Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. Social Science Computer Review, 19, 175-185.

Bartenwerfer, H. (1963). Über Art und Bedeutung der Beziehung zwischen Pulsfrequenz und skalierter psychischer Anspannung. Zeitschrift für Experimentelle und Angewandte Psychologie, 10, 455-470.

Baumann, U., Fähndrich, E., Stieglitz, R.-D. & Woggon, B. (Hrsg.). (1990). Veränderungsmessung in Psychiatrie und Klinischer Psychologie. München: Profil.

Baumann, U., Sodemann, U. & Tobien, H. (1980). Direkte versus indirekte Veränderungsdiagnostik. Zeitschrift für Differentielle und Diagnostische Psychologie, 1, 201-216.

Baumann, U., Thiele, C. & Laireiter, A. R. (2003). Felddiagnostik (ambulantes Assessment) – insbesondere mittels computerunterstützter Verfahren – als Methode der psychopathologischen Verlaufsforschung. In: M. Soyka, H.-J. Möller & H.-U. Wittchen (Eds.). Psychopathologie im Längsschnitt. Methoden, Analyse, Bewertung (S. 65-87). Landsberg/Lech: ecomed.

Beauducel, A., Biehl, B., Bosniak, M., Conrad, W. Schönberger, G. & Wagener, D. (Eds.). (2005). Symposium and Festschrift on multivariate research strategies – Professor Dr. Werner Wittmann. Aachen: Shaker.

Becker, H.-U. (1992). Die Orthostase-Reaktion: Gruppierung und Parametrisierung individueller Reaktionsverläufe (Forschungsbericht Nr. 86).

Berntson, G.G., Bigger, J.T., Eckberg, D.L., Grossman, P., Kaufmann, P.G., Malik, M., Nagaraja, H.N., Porges, S.W., Saul, J.P., Stone, P.H. & van der Molen, M.W. (1997). Heart rate variability: Origins, methods, and interpretive caveats. Psychophysiology, 34, 623-648.

Bland, J. M. & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 8 (1), 307-310.

Borg, I. & Staufenbiel, T. (1997). Theorien und Methoden der Skalierung. Eine Einführung. (3. Aufl.). Bern: Huber.

Bortz, J., Lienert, G. A. & Boehnke, K. (2000). Verteilungsfreie Methoden in der Biostatistik (2.Aufl.). Berlin: Springer.

Bottenberg, E. H. (1972). Emotionspsychologie, Ein Beitrag zur Dimensionierung emotionaler Vorgänge. München: Goldmann.

Bradley, M. M., Greenwald, M. K. & Hamm, A. O. (1993). Affective picture processing. In: N. Birbaumer & A. Öhman (Eds.). The structure of human emotion. Psychophysiological, cognitive and clinical aspects. Seattle: Hogrefe & Huber.

Brähler, E., Schumacher, J. & Strauß, B. (Hrsg.). (2002). Diagnostische Verfahren in der Psychotherapie. Göttingen: Hogrefe.

Brandstätter, H. (1983). Emotional responses for other persons in everyday life situations. Journal of Personality and Social Psychology, 45, 871-883.

Buse, L. & Pawlik, K. (1984). Inter-Setting-Korrelationen und Setting-Persönlichkeit-Wechselwirkungen: Ergebnisse einer Felduntersuchung zur Konsistenz von Verhalten und Erleben. Zeitschrift für Sozialpsychologie, 15, 44-59.

Buse, L. & Pawlik, K. (1994). Differenzierung zwischen Tages-, Setting- und Situationskonsistenz ausgewählter Verhaltensmerkmale, Maßen der Aktivierung, des Befindens und der Stimmung in Alltagssituationen. Diagnostica, 40, 2-26.

Buse, L. & Pawlik, K. (1996). Ambulatory behavioral assessment and in-field psychological testing. In J. Fahrenberg & M. Myrtek (Eds.). Ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies. Seattle, WA.: Hogrefe & Huber. S. 29-50.

Buse, L. & Pawlik, K. (2001). Computer-assisted Ambulatory Performance Tests in Everyday Situations: Construction, Evaluation, and Psychometric Properties of a Test Battery Measuring Mental Activation. In: J. Fahrenberg & M. Myrtek (Eds.). *Progress in ambulatory assessment* (pp. 3-23). Seattle WA: Hogrefe & Huber Publishers.

de Vries, M. W. (Ed.). (1992). The experience of psychopathology. Investigating mental disorders in their natural settings. Cambridge: Cambridge University Press.

Diener, E. & Larsen, R. J. (1984). Temporal stability and cross-situational consistency of affective, behavioral, and cognitive responses. Journal of Personality and Social Psychology, 47, 871-883.

Ebner, U. (2004). Ambulantes psychophysiologisches Monitoring in der psychiatrischen Forschung. Phil. Diss. Freiburg i.Br. Frankfurt a.M.: P. Lang.

Ebner-Priemer, U. (2005). Borderline und die Herausforderung der Instabilität. Expert Meeting. 30th June 2005 - 2nd July 2005, ZI Mannheim. http://www.ambulatory-assessment.org/Expert_Meetings.html

Egloff, B. & Krohne, H.-W. (2002). PANAS. Positive and Negative Affect Schedule. In E. Braehler, J. Schumacher & B. Strauss (Hrsg.), Diagnostische Verfahren in der Psychotherapie (S. 264-266). Goettingen: Hogrefe.

Egloff, B., Schmukle, S.C., Burns, L. R., Kohlmann, C.-W. & Hock, M. (2003). Facets of dynamic positive affect: Differentiating joy, interest, and activation in the Positive and Negative Affect Schedule (PANAS). Journal of Personality and Social Psychology, 85 (3), 528-540.

Eid, M. & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. Journal of Personality and Social Psychology, 76, 662-676.

Eid, M. & E. Diener, E. (Eds.). (2005). Handbook of multimethod measurement in psychology. Washington, D.C.: American Psychological Association.

Eid, M. & Langeheine, R. (2003). Separating stable from variable individuals in longitudinal studies by mixture distribution models. Measurement: Interdisciplinary Research and Perspectives, 1, 179-206.

Fahrenberg, J. (1983). Psychophysiologische Methodik. In K.J. Groffmann & L. Michel (Hrsg.), Enzyklopädie der Psychologie. Psychologische Diagnostik Bd. 4. Verhaltensdiagnostik (pp. 1-192). Göttingen: Hogrefe.

Fahrenberg, J. (2006). Assessment in daily life. A review of computer-assisted methodologies and applications in psychology and psychophysiology, years 2000 - 2005. Available at http://www.ambulatory-assessment.org/

Fahrenberg, J. (2006/2007). Emotionsforschung im Alltag. In: M. Schmidt-Daffy, G. Debus & W. Janke (Hrsg.). Experimentelle Emotionspsychologie: Methodische Ansätze, Probleme und Ergebnisse.

Fahrenberg, J., Bolkenius, K., Maier, S., Schmidt, M., Foerster, F., Hüttner, P., Käppler, C. & Leonhart, R. (2002a). Evaluation des negativen Retrospektionseffektes. Untersuchungen mit Monitor. Forschungsbericht des Psychologischen Instituts der Universität Freiburg, Nr. 156.

Fahrenberg, J., Foerster, F., Schneider, H. J., Müller, W. & Myrtek, M. (1984). Aktivierungsforschung im Labor-Feld-Vergleich. Zur Vorhersage von Intensität und Mustern psychophysischer Aktivierungsprozesse während wiederholter psychischer und körperlicher Belastung. München: Minerva.

Fahrenberg, J., Foerster, F. & Wilmers, F. (1995). Is elevated blood pressure level associated with higher cardiovascular responsiveness in laboratory tasks and with response specificity? Psychophysiology, 32, 81-91.

Fahrenberg, J., Hüttner, P. & Leonhart, R. (2001). Psychological assessment in everyday life by hand-held PC: Applications of MONITOR. In: J. Fahrenberg & M. Myrtek (Eds.). *Progress in ambulatory assessment* (pp. 93-112). Seattle WA: Hogrefe & Huber Publishers.

Fahrenberg, J., Leonhart, R. & Foerster, F. (2002b). Alltagsnahe Psychologie mit hand-held PC und physiologischem Mess-System. Bern: Huber.

Fahrenberg, J. & Myrtek, M. (2005). Psychophysiologie in Labor, Klinik und Alltag. 40 Jahre Projektarbeit der Freiburger Forschungsgruppe Psychophysiologie – Kommentare und Neue Perspektiven. Frankfurt a.M.: Lang.

Fahrenberg, J., Myrtek, M., Kulick, B. & Frommelt, P. (1977). Eine psychophysiologische Zeitreihenstudie an 20 Studenten über 8 Wochen. Archiv für Psychologie, 128, 242-264.

Fahrenberg, J., Myrtek, M., Pawlik, K. & Perrez, M . (2006/2007). Ambulantes Assessment – Verhalten im Alltagskontext erfassen. Eine verhaltenswissenschaftliche Herausforderung an die Psychologie. Psychologische Rundschau.
Foerster, F. (1995). On the problems of initial-value-dependencies and measurement of change. Journal of Psychophysiology, 9, 324-341.
Gorin, A. A. & Stone, A. A. (2001). Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments. In: A. Baum, T. A. Revenson & J. E. Singer (Eds.). Handbook of health psychology (pp. 405-413). New Jersey: Erlbaum.
Hampel, R. (1972). Entwicklung einer Skala zur Selbsteinschätzung der aktuellen Stimmung (SKAS). Phil. Diss., Universität Freiburg i. Br.
Hampel, R. (1977). Adjektiv-Skalen zur Einschätzung der Stimmung (SES). Diagnostica, 23, 43-61.
Heger, R. (1990). Psychophysiologisches 24-Stunden Monitoring. Methodenentwicklung und erste Ergebnisse eines multimodalen Untersuchungsansatzes bei 62 normotonen und blutdrucklabilen Studenten. Phil. Diss., Universität Freiburg i. Br. Frankfurt a. M.: P. Lang.
Hektner, J. M. & Csikszentmihalyi, M. (2002). The Experience Sampling Method: Measuring the context and content of lives. In: R.B. Bechtel & A. Churchman (Eds.). Handbook of environmental psychology, (pp. 233-243). New York: Wiley.
Hüttner, P. (2001). *MONITOR Manual.* Forschungsgruppe Psychophysiologie. Department of Psychology. University of Freiburg, Germany.
Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10*, 253-294.
Hüttner, P. & Leonhart, R. (2002). Beschreibung von MONITOR-9 http://www4.psychologie.uni-freiburg.de/einrichtungen/Psychophysiologie/
Jacob, R. G., Thayer, J. F., Manuck, S. B., Muldoon, M. F., Tamres, L. K., Williams, D. M., Ding, Y. & Gatsonis, C. (1999). Ambulatory Blood Pressure Responses and the Circumplex Model of Mood: A 4-Day Study. Psychosomatic Medicine, 61, 319-333.
Jain, A. (1995). Kardiovaskuläre Reaktivität im Labor und im Feld. Eine komparative Studie zur Aussagekraft kardiovaskulärer Reaktivitätsparameter unter Feldbedingungen. Phil. Diss., Universität Köln. Münster: Waxmann
Janke, W. & Debus, G. (1978). EWL Die Eigenschaftswörter-Liste. Göttingen: Hogrefe.
Käppler, C. (1994). Psychophysiologische Bedingungsanalyse von Blutdruckveränderungen im alltäglichen Lebenskontext. Phil. Diss. Universität Freiburg i. Br. Frankfurt a. M.: P. Lang.
Käppler, C., Brügner, G. & Fahrenberg, J. (2001). Pocketcomputer-unterstütztes Assessment mit MONITOR: Befindlichkeit im Alltag, Methodenakzeptanz und die Replikation des Retrospektionseffektes. Zeitschrift für Differentielle und Diagnostische Psychologie, 22, 249-266.
Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N. & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. Science, 306, 1776-1780.
Kinne, G. (1997). Interaktives Monitoring von Myokardischämie. Psychophysiologische Zusammenhänge von Ischämie und Angina pectoris im Alltag von Koronarpatienten. Phil. Diss., Universität Freiburg i. Br. Frankfurt a. M.: P. Lang.
Krauth, J. (1983). Bewertung der Änderungssensitivität von Items. Zeitschrift für Differentielle und Diagnostische Psychologie, 4, 7-28.
Krauth, J. (1985). Testkonstruktion und Testtheorie. Weinheim: Beltz.
Krohne, H. W., Egloff, B., Kohlmann, C.-W. & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). Diagnostica, 42 (2), 139-156.
Kubiak, T. (2003). Entwicklung und erste empirische Überprüfung eines stationären Interventionskonzepts zur Behandlung von Typ 1 Diabetikern mit Hypoglykämieproblemen. Phil. Diss. Freiburg i. Br. Frankfurt a.M.: P. Lang.
Larsen, R. J. & Prizmic-Larsen, Z. (2005). Measuring emotions: Implications of a multimethod perspective. In: M. Eid & E. Diener (Eds.). Handbook of multimethod measurement in psychology (pp. 337-352). Washington, D.C.: American Psychological Association.
Larson, R. & Csikszentmihalyi, M. (1983). The Experience Sampling Method. New Directions for Methodology of Social & Behavioral Science, 15, 41-56.

Laux, L., Glanzmann, P., Schaffner & Spielberger, C.D. (1981). Das Sate-Trait-Angst-Inventar (STAI). Weinheim: Beltz.
Lersch, P. (1970). Der Aufbau der Person (11. Aufl.). München: Barth.
Lucas, R. E. & Baird, B. M. (2005). Global self-assessment. In: M. Eid & E. Diener (Eds.). Handbook of multimethod measurement in psychology (pp. 29-42). Washington, D.C.: American Psychological Association.
Marsella, A. J., Dubanoski, J., Hamada, W. C. & Morse, H. (2000). The measurement of personality across cultures. American Behavioral Scientist, 44, 41-62.
Matyas, T. A. & Greenwood, K. M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13*, 137-157.
Mc Crae, R. R. & Costa, P. T. (1997). Personality trait structure as a human universal. American Psychologist, 52, 509-516.
Myrtek, M. (2004). Heart and emotion. Ambulatory monitoring studies in everyday life. Cambridge, MA: Hogrefe & Huber Publishers.
Myrtek, M., Foerster, F. & Brügner, G. (2001). Freiburger Monitoring System (FMS). Ein Daten-Aufnahme- und Auswertungssystem für Untersuchungen im Alltag: Emotionale Beanspruchung, Körperlage, Bewegung, EKG, subjektives Befinden, Verhalten. Frankfurt a. M.: Peter Lang.
Orth, B. (1983). Grundlagen des Messens. In H. Feger & J. Bredenkamp (Hrsg.). Enzyklopädie der Psychologie. Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 3 Messen und Testen (S. 136-180). Göttingen: Hogrefe.
Ott, R. & Scholz, O. B. (2001). Time series analysis of diary data. In: J. Fahrenberg & M. Myrtek (Eds.). Progress in ambulatory assessment (pp. 157-171). Seattle WA: Hogrefe & Huber Publishers.
Palmblad, M. & Tiplady, B. (2004). Electronic diaries and questionnaires: designing user interfaces that are easy for all patients to use. Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 13, 1199-1207.
Pawlik, K. & Buse, L. (1982). Rechnergestützte Verhaltensregistrierung im Feld: Beschreibung und erste psychometrische Überprüfung einer neuen Erhebungsmethode. Zeitschrift für Differentielle und Diagnostische Psychologie, 3, 101-118.
Pawlik, K. & Buse, L. (1992). Felduntersuchungen zur transsituativen Konsistenz individueller Unterschiede im Erleben und Verhalten. In K. Pawlik & K. H. Stapf (Hrsg.), Umwelt und Verhalten, S. 25-69. Bern: Huber.
Pawlik, K. & Buse, L. (1994). "Psychometeorologie": Zeitreihenanalytische Ergebnisse zum Einfluß des Wetters auf die Psyche aus methodenkritischer Sicht. *Psychologische Rundschau, 45,* 63-78.
Pawlik, K. & Buse, L. (1996). Verhaltensbeobachtung in Labor und Feld. In: K. Pawlik (Hrsg.), Enzyklopädie der Psychologie. Differentielle Psychologie und Persönlichkeitsforschung. Band 1. Grundlagen und Methoden der Differentiellen Psychologie (S. 359-394). Göttingen: Hogrefe.
Peper, M. (2006). Neurobiologische Emotionsmodelle. In G. Stemmler (Hrsg.), Enzyklopädie der Psychologie: Themenbereich C Theorie und Forschung, Serie 4, Band 3: Psychologie der Emotion. Göttingen: Hogrefe.
Perrez, M. & Reicherts, M. (1989). Belastungsverarbeitung: Computerunterstützte Selbstbeobachtung im Feld. Zeitschrift für Differentielle und Diagnostische Psychologie, 10, 129-139.
Perrez, M., Schoebi, D. & Wilhelm, P. (2000). How to assess social regulation of stress and emotions in daily family life? A computer-assisted family self-monitoring system (FASEM-C). Clinical Psychology and Psychotherapy, 7, 326-339.
Posner, J., Russell, J. A. & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17(3), 15-734.
Pohl, R. F. (Ed.). (2004). Cognitive illusions. A handbook on fallacies and biases in thinking, judgment and memory. New York: Psychology Press.
Reicherts, M. (2005). The Learning Affect Grid - a computer based monitoring system: first results. Expert Meeting. 30th June 2005 - 2nd July 2005, ZI Mannheim. http://www.ambulatory-assessment.org/Expert_Meetings.html
Rodgers, L. J. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. Multivariate Behavioral research, 34, 441-456.
Rösler, F., Baumann, U. & Marake, H. (1980). Zum Vergleich zwischen globaler und additiver Befindlichkeitserfassung. Diagnostica, 26, 151-164.

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161-1178.
Rohracher, H. (1988). Einführung in die Psychologie (13. Aufl.). Wien: Urban & Schwarzenberg.
Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie, 9, 222-245.
Scherer, K. P. (Hrsg.). (1990). Psychologie der Emotion. Enzyklopädie der Psychologie : Themenbereich C, Theorie und Forschung. Serie Motivation und Emotion. Band 3. Göttingen:Hogrefe.
Scherer, K., Wranik, T., Sangsue, J., Tran, V. & Scherer, U. (2004). Emotions in everyday life: probability of occurrence, risk, factors, appraisal and reaction patterns. Social Science Information, 43, 499-570.
Schmidt-Atzert, L. (1996). Lehrbuch der Emotionspsychologie. Stuttgart: Kohlhammer.
Schmidt-Atzert, L. (1997). Entwicklung und Evaluierung von Skalen zur Erfassung des emotionalen Befindens in den letzten 7 Tagen (EMO-16-Woche). Zeitschrift fuer Differentielle und Diagnostische Psychologie, 18 (3), 182-198.
Schmidt-Atzert, L. & Hüppe, M. (1996). Emotionsskalen EMO 16. Ein Fragebogen zur Selbstbeschreibung des aktuellen emotionalen Gefühlszustandes. Diagnostica, 42 (3), 242-267.
Schmukle, C., Egloff, B. & Burns, L. R. (2002). The relationship between positive and negative affect in the Positive and Negative Affect Schedule. Journal of Research in Personality, 36, 463-475.
Schneider, H. J. (1982). Befindensskalen für Aktivierungsexperimente: Anforderungen, statistische Eigenschaften, Selektionskriterien, Eingewöhnungseffekte, Dimensionalität. (Forschungsbericht Nr. 2). Freiburg i. Br.: Albert-Ludwigs-Universität, Psychologisches Institut.
Schwarz, N. (1990). Assessing frequency reports of mundane behaviors: Contributions of cognitive psychology to questionnaire construction. In C. Hendrick & M. S. Clark (Eds.). Research methods in personality and social psychology (pp. 98-119). Newbury Park, CA : Sage.
Schwarz, N. & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Antwortvorgaben des Fragebogens lernen. Zeitschrift für Klinische Psychologie, 21, 197-208.
Schweizer, K. (1989). Eine Analyse der Konzepte, Bedingungen und Zielsetzungen von Replikationen. Archiv für Psychologie, 141, 85-97.
Stemmler, G. (1992). Differential psychophysiology: Persons in situations. Berlin: Springer.
Stemmler, G. (1996). Strategies and designs in ambulatory assessment. In: J. Fahrenberg & M. Myrtek (Eds.). Ambulatory Assessment: computer-assisted psychological and psychophysiological methods in monitoring and field studies (pp. 257-268). Seattle, WA: Hogrefe & Huber.
Stemmler, G. (2001). Grundlagen psychophysiologischer Methodik. In: F. Rösler (Hg.). Ergebnisse und Methoden der Psychophysiologie. Enzyklopädie der Psychologie, Band 4, Serie I, Themenbereich C (S. 1-84). Göttingen: Hogrefe.
Stemmler, G. & Fahrenberg, J. (1989). Psychophysiological assessment: Conceptual, psychometric, and statistical issues. In: G. Turpin (Ed.). Handbook of clinical psychophysiology (pp. 71-104). Chichester: Wiley.
Stiglmayr, Ch. (2003). Spannung und Dissoziation bei der Borderline-Persönlichkeitsstörung. Phil. Diss. Freiburg i. Br. Frankfurt a. M.: P. Lang.
Stone, A. & Litcher-Kelly, L. (2005). Momentary capture of real-world data. In: M. Eid & E. Diener (Eds.). Handbook of multimethod measurement in psychology (pp. 61-72). Washington, D.C.: American Psychological Association.
Stone, A. A. & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. Annals of Behavioral Medicine, 16, 199-202.
Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. Behavioral Assessment, 10, 343-366.
Suen, H. K. & Ary, D. (1989). Analyzing quantitative behavioral observational data. Hillsdale, N.J.: Lawrence Erlbaum.
Tschacher, W. (1997). Prozessgestalten. Göttingen: Hogrefe.
Thayer, R. E. (1970). Activation states as assessed by verbal report and four psychophysiological variables. Psychophysiology, 7, 86-94.
Thayer, R. E. (1978). Toward a psychological theory of multidimensional activation (arousal). Motivation and Emotion, 2, 1-36.
Totterdell, P., Briner, R. B., Parkinson, B. & Reynolds, S. (1996). Fingerprinting time series: Dynamic patterns in self-report and performance measures uncovered by a graphical non-linear method. British Journal of Psychology, 87, 43-60.

Totterdell, P., Spelten, E., Smith, L., Barton, J. and Folkard, S. (1995). Recovery from work shifts: how long does it take? Journal of Applied Psychology. 80, 43-57.
Triemer, A. (2003). Ambulantes psychophysiologische 24-Stunden-Monitoring zur Erfassung von arbeitsbezogenen Stimmungen und Emotionen. Frankfurt a. M.: Peter Lang.
Triemer, A. & Rau, R. (2001). Stimmungskurven im Arbeitsalltag - eine Feldstudie. Zeitschrift fuer Differentielle und Diagnostische Psychologie, 22 (1), 42-55.
Tschacher, W. (1997). Prozessgestalten. Göttingen: Hogrefe.
Watson, D. (1997). Measurement and mismeasurement of mood: Recurrent and emergent issues. Journal of Personality Assessment, 68, 267-296.
Watson D., Clark L. A. & Tellegen, A. (1984).Cross cultural convergence in the structure of mood a Japanese replication and a comparison with US findings. Journal of Personality Social Psychology, 47, 127-144.
Watson, D., Clark, L. A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54 (6), 1063-1070.
Watson, D. & Tellegen, A. (1985). Toward a consensual structure of mood. Psychological Bulletin, 98, 219-235.
Westhoff, G. (1993). Handbuch psychosozialer Messinstrumente. Göttingen: Hogrefe.
Wilhelm, P. (2004). Empathie im Alltag von Paaren. Akkuratheit und Projektion bei der Einschätzung des Befindens von Paaren. Bern: Huber.
Wilhelm, P. & Perrez, M. (2001). Felddiagnostik. In: R.-D. Stieglitz, U. Baumann & H.J. Freyberger (2001). Psychodiagnostik in Klinischer Psychologie, Psychiatrie, Psychotherapie (S. 169-182). Stuttgart: Thieme
Wilhelm, P., Schoebi, D. & Perrez, M. (2004). Frequency estimates of emotions in everyday life from a diary method's perspective: a comment on Scherer et al.'s survey-study "Emotions in everyday life". Social Science Information, 43 (4), 647-665.
Yamane, T. (1969) Statistics. An introductory analysis (2 nd ed.) New York: Harper & Row.
Yik, M. S. M, Russell, J. A., Barrett, L. F. (1999). Structure of self-reported current affect: Integration and beyond. Journal of Personality and Social Psychology, 77(3), 600-619.
Zautra, A. J., Berkhof, J. & Nicolson, N. A. (2002). Changes in affect interrelations as a function of stressful events. Cognition & Emotion, 16, 309-318.
Zelenski, J. M. & Larsen, R. J. (2000). The distribution of basic emotions in everyday life: A state and trait perspective from Experience Sampling data. Journal of Research in Personality, 34(2), 178-197.
Zerssen, D. von (1976). Die Befindlichkeitsskala. Weinheim: Beltz.
Zimmermann, P. (1978). Zur Zeitreihenanalyse von Stimmungsskalen. Diagnostica, 25, 24-48.