Ambulatory Assessment – European Network

Statements and Open Peer Commentaries

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

Jochen Fahrenberg, Freiburg i. Br.

(Stand Juli 2006)

Vorbemerkung

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.

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 ausgeführt. 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.

Dieser Beitrag möchte den Erfahrungsaustausch und eine weiterführende Diskussion anregen. Kommentare sind willkommen an die Adresse: jochen.fahrenberg@psychologie.uni-freiburg.de

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 amerikanischen 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 aus 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. 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.

Methodenkritischer Kommentar zu AD-ACL und PANAS

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. Verschiedent-

lich 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:

(1) 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 intraindividuelle Varianz der Zustandsänderungen verwendet (– diese Adäquatheitsfrage wird allerdings auch sonst selten gestellt).
(2) 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.
(3) 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).
(4) 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.
(5) 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.
(6) 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.

(7) 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.
(8) 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.
(9) 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. 21 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.
(10) Es fehlen Überlegungen und systematische Ergebnisse zu den verschiedenen Aspekten der lokalen und aggregierten Reliabilität und zur Kriterienvalidität, denn die faktorielle Validität hat ja zunächst nur formale Bedeutung.
(11) 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.
(12) Kritisch ist zu fragen, für welche psychologischen Fragestellungen solche reduzierten Dimensionalitäten nützlich sein könnten und für welche Fragestellungen ein differenzierteres 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 "Universalien" 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).

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, d.h. von Ein-Item-Skalen, zweckmäßiger sein.

Offensichtlich müsste, um vielleicht eine Standardmethodik für Selbsteinstufungen der Befindlichkeit zu erreichen, eine neue Konstruktion auf der primären Basis deutschsprachiger Items und mit Blick auf die intra-individuelle Variabilität unternommen werden. Durch den systematischen Vergleich der querschnittlichen und der längsschnittlichen Daten und im Hinblick auf psychologisch wichtige Verlaufshypothesen und Kriterien wären Fortschritte dieser Methodologie möglich.

Anmerkung

Kurzgefasste Thesen sind oft missverständlich. Deswegen wird auf die ausführlichere Darstellung dieser Methodenprobleme mit den entsprechenden Literaturangaben hingewiesen:
Ambulantes Assessment von Befinden, Stimmungen, Emotionen – Zur Methodik von Selbsteinstufungen (Selbstberichten) (Juli 2006)
auf der Homepage http://www.jochen-fahrenberg.de/index.php (Ambulantes Assessment)

und auf die breitere Darstellung in:

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