Ambulatory Assessment – European Network

Statements and Open Peer Commentaries

Self-reported Subjective State – Single Items or Scales like AD-ACL and PANAS?

Jochen Fahrenberg, Freiburg i. Br.

(October 2006)

A basic aim of ambulatory monitoring is to assess the change of subjective state (mood, emotion, affect, well-being) and symptoms (complaints) over the course of the day. There are several reasons why the computer-assisted approach is the best method for ambulatory assessment: Besides ecological validity, this method affords great flexibility in programmed data gathering, measurement of the specific time and context of self-reports, and a high controlled compliance (cf. position paper, Fahrenberg, Myrtek, Pawlik & Perrez, 2007).

The principles and methodological problems encountered in this form of assessment correspond to a large extent with those found in paper-and-pencil self-reports (mood scales), which have been used extensively for over 50 years. Given that the contemporary investigative approaches implemented in ambulatory assessment frequently stem from disciplines other than differential psychology it may be helpful to revisit the principles of psychological test theory and to highlight some important methodological problems.

Current textbooks of test theory and personality assessment largely limit their treatment of these aspects of methodology to (1) aptitude tests, in which the concept of parallel measurement is appropriate, or to (2) personality questionnaires of various kinds, which deal also with relatively stable traits. The diagnosis of state change with its own special considerations, or appropriate assessment strategies and decision making based on a general assessment theory are rarely elaborated on. Contemporary textbooks in Germany and USA (and, probably, in academic teaching generally) have still a long way to go before the special features and merits of ambulatory assessment are given their due attention.

This article is intended to stimulate the exchange of experiences and to encourage further discussion.

Commentaries are welcome at this address:
(1) Homepage European Network for Ambulatory Assessment
www.ambulatory-assessment.org
Webmaster Dr. U. Ebner-Priemer, ZI Mannheim
ulrich.ebner-priemer@zi-mannheim.de
www.uli-ebner.com
or
(2) jochen.fahrenberg@psychologie.uni-freiburg.de

Self-reported Subjective State –
Single Items or Scales like AD-ACL and PANAS?

There are remarkable differences in the methodology used by investigators in the ambulatory assessment of subjective state (mood, emotion, affect). While some investigators select single items, others prefer scales composed of several items. The investigators choice is often influenced by the preferences of other investigators as reported in the most recent American publications. Issues of this kind are not of course encountered in every project because the item content and format, and other details are often determined by the fundamental objective of the project.

Advantages and disadvantages of single-item scales and of multiple-item scales

The use of single items (single-items scales) is advantageous when: (1) a relatively broad range of feelings, moods and emotions should be described with a few items, and (2) the choice of items is easily adapted to the particular question.

German item lists for ambulatory assessment of subjective state have been published by 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).

The *disadvantages of items* are that (1) single-item scales with only few scale points for responding permit no more than a correspondingly small degree of differentiation and often show a skewed distribution of values, and that (2) the conventional estimation of reliability (consistency, item homogeneity) is not possible. The unfavourably small variance can at least be moderated by using a multi-point scaling format such as a suitable visual analogue 21-point scale, and by implementing this format providing some initial training of participants.

The use of scale values rather than item values is beneficial when: (1) a numerically larger variance within and between persons (discrimination) can be achieved with scales of for example 10 items; (2) items with different item means are selected to approach a "normal distribution" of scale values with respect to kurtosis/excess; and, (3) the internal consistency of a number of scale items can be calculated and therewith the coefficient of local reliability.

The *disadvantages of scale values* are that: (1) concerns about the reasonable length of a self-report could restrict the investigator to the use of only one or two scales which, for the sake of test efficiency, does mean having to refrain from measuring other important aspects of subjective state; (2) each scale requires a number of items with highly similar content which could however place some burden on the respondents having to enter ratings over and over again. This in turn could have a negative impact on acceptance and method-related reactivity; (3) it is very doubtful in the domain of emotionality whether the fundamental assumptions in measurement theory are tenable: that the items of a scale are homogenous in content and that they constitute independent (parallel) measures of the underlying concept; (4) the assumption that all of the item-response functions behave in a synchronous (consistent) fashion over time has still to be validated; and, (5) the straightforward summation of item values (ordinal data) to metric scale values is questionable.

If a consistency coefficient like Cronbach's Alpha-coefficient is to be regarded as a measure of reliability then all items of a scale must be shown to represent parallel measures of the construct (see above). For this reason a schematic application of the consistency analysis is problematic. A precise evaluation of which facets of the construct are represented and the evaluation of whatever associations there are between item number, redundancy and test efficiency is imperative. In addition to this, some indication of the validity should be provided for each type of scale and for each instrument on the basis of criteria correlations and, albeit more difficult, for informed decision-making in a practical assessment task.

Many investigators have therefore preferred to select single items that they regard as relevant to the question. The lack of standardisation of methodology renders a comparison of the research results

from different research groups difficult, and the emergence of a standard method in this field looks unlikely in the foreseeable future. This of course places all the more value on a research strategy that encompasses consistent replication of important research findings, a strategy that should at the very least be pursued within a research group.

Commentary on AD-ACL and PANAS

The majority of published mood scales have a multi-dimensional concept. The large number of scales and items makes them unsuitable for short-term repeated application. The suggestion to reduce the diversity of subjective states to a few dimensions (factors, basic emotions) has found many voices ever since Wundt's three-dimensional theory of emotions. Various instruments with one or two scales for measuring mood (affect) have been developed.

The AD-ACL Activation-Deactivation Adjective Checklist comprises four subscales: General Activation, High Activation, General Deactivation, and Deactivation-Sleep. This checklist has been propagated for the Psychophysiological Research (Thayer, 1970) to investigate correlative relationships on the basis of change measurements. In a subsequent broader concept Thayer (1978) distinguished dimension A (energetic – sleepy) from dimension B (tense – placid, still). The AD-ACL was published in the seventies and appears to arouse little interest today.

The PANAS Positive Affect – Negative Affect Scales (Watson, Clark & Tellegen, 1988) has been applied on various occasions and in the form of a number of German adaptions in past years (e.g. Krohne, Egloff, Kohlmann & Tausch, 1996). Originally, the authors explicitly regarded PANAS as being complementary to rather than rivalling multi-dimensional concepts (Watson & Tellegen, 1985, p. 220). The authors' far-reaching postulate to have found a "consensual structure of mood", perhaps even a "real" structure, encouraged other investigators to use this method. For this reason the PANAS has been chosen as an example on the basis of which to illustrate typical test-methodological problems and to draw attention to the serious flaws that need to be addressed.

Watson und Tellegen (1985) intended to create as simple, parsimonious and generally consensual description as possible of self-reported and observer-reported affects: Positive Affect PA und Negative Affect NA. The basic dimensions in PA and NA were proposed to provide a basis for consensus in the contradictory literature. Dimensions such as Arousal (Activation) und Potency (Dominance, amongst others) were mentioned by the authors, but only superficially. The authors claimed that by re-analysis of different datasets they could prove that the dimensions PA and NA represent the dominating "secondary factors" in factor analyses. These two dimensions (scales) are supposedly independent of one another.

From a theoretical point of view the original work reveals an astonishing lack of careful and thoughtful consideration, and it is propped up largely by factor-analytic technical arguments. The authors appear to be oblivious however to the inner dependencies of their methodological preliminary decisions and statistical results. The test-methodological problems and flaws are:

(1) It is not discernable from the authors' primary research approach that the instrument basically deals with measurements of change rather than trait dimensions of personality questionnaires. The scale construction was not based on intraindividual variance of state changes (– this matter of adequacy is otherwise seldom raised).
(2) The authors fail to recognise their fundamental bias in their primary item pool studies and in their choice of studies cited by them, that is, the absence of a random sample from the universe of emotion descriptors. This is a fundamental requirement – though empirically hard to achieve – in order to be able to achieve a general taxonomy grounded on factor analysis.
(3) The data were collected from simple paper-and-pencil investigations and are therefore much more doubtful than data from computer-based investigations that were already feasible in the eighties (see Pawlik & Buse, 1982, 1996).
(4) The test-statistical and factor-analytical consequences of the different item (co-)variances and the distribution of the item values (inter- and above all intra-individual), particularly in the case of the NA items, has been given too little attention.

(5) The fundamental issues and methodological problems of measuring changes as well as of scale properties and scaling procedure is not examined in any further detail despite its paramount importance for the description of subjective and, in particular, rapidly changing states.
(6) The statistical separation of the basic components of variance was not performed: between individuals, within individuals, within and between days (and interactions), either by partitioning of covariance, by multi-level analyses or by modelling according to a latent trait/state concept.
(7) The trivial effects of redundant items on the scale construction and factor analysis was overlooked: doublets or triplets of highly correlated items can – on account of the implicit weighting during item selection process – result in a serious structural bias of scale/factor content.
(8) Most researchers will acknowledge that the orthogonal or oblique factor analytic results represent not more than one of a number of possible mathematic-statistical equivalent systems of description (to be evaluated against external criterion measures). Instead, the "existence" of the PANAS dimensions is taken for granted. The author's comparison of such dimensional analyses with the factoral-analytic intelligence research and the g-factor is distorted for a number of reasons and exhibits a reductionism that is particularly problematic in the area of mood and emotion (subjective experience).
(9) The test efficiency of the long item list of 20 items is not given due thought in terms of the "validity per time unit". This begs the question as to what essential information the investigators neglected because the addition of even more items for repeated self-reports would have simply been too much for the subject. No consideration was given to the pragmatic point: would it not have been incomparably easier to provide the respondents with a visual analogue scale (with for example 21-points) for PA and for NA in order to measure PA and NA more easily, with few presuppositions and considerably faster? Proof must be delivered that the PANAS accounts for more incremental variance in terms of the relevant criteria than a simple single-item scale of the VAS-type.
(10) Careful deliberation and systematic results of the different aspects of local and aggregated reliability and of criterion validity is absent, for the factoral validity has, in the first instance, only formal relevance.
(11) The authors did not develop explicit assessment strategies in the original publication, for what should the PA- and NA-scales, both limited to 10 largely homogenous items per scale, be best used for, nor did they elaborate on a typical application in research or professional practice.
(12) The critical question is for which psychological investigations such reduced dimensionality is actually useful, and for which questions a more differentiated system of description is preferable or even crucial. Further to this, different assessment tasks would draw practical benefit from different methods.

In a more recent paper (Watson & Clark, 1997), the authors try to justify some of their more problematic steps and seek to promote the use of their scales. The authors now expressly refer to a hierarchical structure and to multiple specific emotional states. The key question has still not been addressed: structure stability is not the same as change sensitivity, but the latter is neither explained nor investigated. Besides the factorial construct validity there is no mention whatsoever of the author's own contributions to the empirical validity of the PANAS, let alone a superior criterion validity.

Consistencies of 86. to .90 and .84 to .87 (for 10 Items, respectively) for the PA and NA-scales and factor correlations of .95 and .93 are reported without reference to the obvious redundancy of many items (for the sake of higher homogeneity). The authors claim that the scales are independent, a claim that was factor-analytically deliberate and forced, but this independence varies according to the length of the subjectively assessed time interval under consideration, and it still amounts to .30 (momentary) and .34 (for the day) in the authors' only large within-subject dataset. More recent investigations have failed to substantiate the claim of independence (Schmukle, Egloff & Burns, 2002; Zautra, Berkhof & Nicolson, 2002). As long as the data have been drawn from paper-and-pencil tests these findings are in any case doubtful.

No further elaboration is needed here on the author's attempts at justifying why the important components Fatigue and Serenity are missing in the PANAS or why the authors excluded the components Joviality from PA and Sadness from NA: they did not fit in the intended factor scheme ("these terms failed to enhance the psychometric properties of the PANAS scales" p. 277). In the meantime

PANAS-X has been published with the Sadness scale and PANAS plus with PANAS Happiness Scale, etc.

The account of PANAS is presented in many areas in a circulatory or very one-sided citation style. That the questionnaires developed from the American datasets are culture independent is a peculiar claim. In claiming to have identified dominant dimensions that really exist and are therefore valid for all individuals, the PANAS authors correspond in their own assertions entirely with those of Costa & McCrae and the NEO-FFI Personality Questionnaire, an approach Watson & Tellegen cite as an example to be followed.

Both postulates, extant dimensions (reification) and culture independence, need fundamental qualification, and it is frustrating when a culture independent validity (as a universal truth) is postulated (McCrae & Costa, 1997) on the basis of inadequate one-sided empirical methodology. The past years have seen the publication of several such so-called intercultural studies that follow this kind of simple concept. The translation of American questionnaires into other languages and the presentation of these mainly to students of colleges or universities of western orientation is most certainly inadequate. A fundamentally independent and authentic development of questionnaires and their cross comparison would be much more appropriate (for the ethnological and ethnopsychological criticism of the tendency for psychologists to postulate universal personality traits see, for example, Marsella, Dubanoski, Hamada & Morse, 2000).

In the meantime, the proliferation process of both "universal systems" is however well advanced: there are different versions, there are adaptations with fewer items, even making use once again of bipolar items, and single subscales are being constructed retrospectively – something that comes as a particular surprise considering the previous construction of both instruments and the limited item pool.

Summary of evaluation and critical comments

There is certainly nothing new in the trivial point that positive and negative assessment in self-reported mood is important. A bipolar or valence dimension Pleasant-Unpleasant (lust-unlust) has often been postulated since Wundt. For reasons of semantics and especially because unipolar items have found favour, this perspective has been partitioned into two subscales factor-analytically (and already much earlier in EWL und SKAS). Equally well known is the fact that the respondent generally tends to report a positive mood. In comparison to the current test-methodological and emotion-theoretical literature the PANAS approach can in many regards be said to represent more of a regression than methodological progress, besides which the limited reductionism found in the PA-NA dimensions is very far removed from the multi-system-concepts discussed in the contemporary literature on the neuropsychology of emotion (see Peper, 2006).

On the basis of a critical test-methodological evaluation of the PANAS and similar scales the implementation of such scales in computer-assisted ambulatory assessment cannot be recommended.
A standardisation of the methodology is certainly worth striving for, and the desire to make the investigative methods internationally compatible is understandable. But the objections and concerns relating to test-methodology and test-efficiency are obvious. A more practical approach would therefore be to select appropriate single-item scales on the basis of their content and their statistical parameters.

The ambulatory assessment of subjective state (mood, emotion, affect, well-being) requires adequate studies to advance methodology and standardisation. Such developments should be based primarily on German language items, computer-assisted methodology, and must pay particular attention to individual differences in intra-individual variability. Thus, hopefully, some degree of standardisation can be attained by delineating a set of useful items and item formats. Further progress in this methodology could be achieved by the systematic comparison of cross-sectional and longitudinal data and by taking into consideration psychologically important concepts relating to subjective state and state change, and, not least, behavioural criteria.

Note
Brief theses are often imprecise. For this reason, the reader is referred to
(1) the more detailed accounts of these methodological issues and further references:

Zur Methodik von Selbsteinstufungen (Juli 2006) download pdf-file 764 KB
Available at http://www.jochen-fahrenberg.de/index.php (Ambulantes Assessment/Monitoring)

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

Translation provided by Dipl.-Psych. Marcus Cheetham.