Chapter XX
MONITOR: Acquisition of Psychological Data by a Hand-held PC

Jochen Fahrenberg, Paul Hüttner, and Rainer Leonhart
Forschungsgruppe Psychophysiologie, Department of Psychology, Freiburg University, Germany

Introduction

Computer-assisted data acquisition

Methods of recording psychological data in everyday life have a long history in differential psychology and clinical psychology. Event recorders for the timed registration of stimuli and responses, "beeper" studies in which a programmable wristwatch prompts the subject to respond to a questionnaire, self-ratings on diary cards, and electronic data logger have all been used for this purpose. The arrival of pocket-sized (hand-held, palm-top) computers has eased the acquisition of data considerably. Many Book Chapters in this volume and in the previous volume refer to this methodology (for reviews, see Fahrenberg, 1996; Fahrenberg & Myrtek, 2001; Pawlik & Buse, 1996; Perrez, 1994; Stone & Shiffman, 1994; see also Hufford et al., this Volume). An overview on the origins and developments of psychological data acquisition was included here (Fahrenberg, this Volume).

Following the progress made in pocket-sized computers, software to facilitate the use of hand-held PCs in field studies has been developed in many institutions, more or less geared to the needs of certain studies. More flexible software systems suited to the requirements of a variety of applications are still an exception (see also, the Behavior Observation System, Noldus Information Technology, AG Wageningen, NL; AMBU for in-field performance testing, cf. Buse & Pawlik, 1996, Hogrefe Verlag, Göttingen, Germany; see also their Chapter in this Volume). At Freiburg University, the course of development progressed from the use of an analog cassette recorder (Fahrenberg, Foerster, Schneider, Müller & Myrtek, 1984) to hand-held PCs (Heger, 1990, Käppler, 1994, Myrtek et al., 1988, Myrtek, Brügner & Müller, 1996) during which time the MONITOR software emerged (Brügner, 1998).

Advantages and Limitations

The application of programmable pocket-PCs in ambulatory assessment has many advantages:

- alarm functions for prompting the subject at predefined intervals and a built-in reminder signal;
- reliable timing of input, delay of input, and duration of input;
- flexible layout of questions and response categories;
- branching of questions and tailor-made sequential or hierarchical strategies;
- concealment of previously recorded responses from the subjects;
- convenience and ease of transfer of data to a stationary PC for statistical analysis.

One can generally assume a higher reliability in the timing and saving of responses and a higher ecological validity of such assessments. Self-reports in the form of paper-and-pencil questionnaires and diaries cannot be designed in such a way that they can be applied flexibly and provide the exactness needed when timing responses. In psychology the hand-held PC so far has been predominantly used for recording self-reports on mood and other aspects of subjective state, including physical complaints and symptoms, that is, as an "electronic diary". There are other kinds of data, which can be obtained in field studies: objective features of a behavior setting, behavior observations, behavior and performance measures (in-field psychometric testing, see Pawlik & Buse, this Volume), self-measurements of various kinds, for example, blood pressure data, and, possibly, ratings of environmental aspects, like ambient noise or temperature. Potential contents of a computer-assisted protocol may further include, for example, individual comments or self-evaluation in connection with events.

The versatility and wide acceptance of computer-assisted data acquisition is evident, although there are limitations and obvious restrictions. All participants of such studies will need sufficient practical training in the basic features of the PC and the program, at least, to avoid malfunctions and missing data. In spite of the obvious increase in computer-literacy within the general population, there are sub-populations which are less familiar with such devices or may experience problems, for example, as in the case of the elderly who were not acquainted with PCs in their job, or in some clinical populations and in patients with sensory or motor disorders. Indeed, the application of questionnaires would likewise prove to be difficult in such samples.

In the past, the issue has been raised whether the use of a hand-held PC will in itself lead to cause a specific method variance. When comparing computer-assisted questioning with paper-and-pencil-type questionnaires, empirical investigations have not confirmed the assumed bias (for example, Finegan & Allen, 1994; Franke, 1998; Hank & Schwenkmezger, 1996). However, the reactivity phenomena may arise in daily life when the participants of such assessments are prompted unexpectedly by a beeper to respond to a number of items; moreover, an enhanced level of self-awareness may be induced by the task of coping and with the Hand-held PC and of giving self-reports several times during the day.

Outline of MONITOR

Computer-assisted self-reports require a hand-held PC with certain features: a large display, easy handling of basic controls, clock, beeper with volume control, sufficient capacity of storage, low power consumption, and a low weight. For many applications, a comparatively large alphanumeric keyboard (complete QWERTY) is also preferable in order to ease recording and especially, to record verbal responses. The latter may involve, for example, recording reports and comments about specific events, or reporting more precisely the occurrence of physical and psychological symptoms, which in either case hardly fit pre-defined categories. For some applications it may suffice to record only “yes” or “no” responses or numbers (Likert Scale type). In this case, a smaller hand-held PC may be preferable, although small keys may present a problem for some subjects or patients.

The Hand-held (palm-top) computers of the Psion™ Series 3 (Psion PLC, London, England) facilitated the development and application of MONITOR particularly well. The computer, weight 275 g, has sufficient memory capacity (16 bit CPU, RAM 512 K,1 MB ROM, 2 MB RAM, card 1 MB), a clock, a piezoresistive beeper, alphanumeric keyboard, and an operating system that allows flexible programming in OPL. A large LCD (125 x 45 mm, 480 x 160 pixel, that is, 80 symbols x 25 lines) is used to display rating scales, multiple choice items or test items which can be answered by pressing a key or by typing in answers. The Psion™ Revo is a successor of the Series 3 using EPOC Release 5 (36 MHz 32 bit RISC CPU, 8 MB ROM, 8 MB RAM, weight 200 g).

Table 1

The MONITOR is a general purpose software which may be used for a variety of applications. In the following, an outline of MONITOR Level 9 is given. MONITOR provides a number of options for input, control commands, and output as depicted in Table 1 (for a detailed description, see the MONITOR Manual, Hüttner, 2001).

MONITOR – an example of its application

A typical assessment may include a setting protocol, a number of self-ratings of subjective state, mental performance tests, and specific questions about aspects of acceptance and reactivity in connection with the prompt-to-respond function of the PC (cf. Fahrenberg, 1996, Heger, 1990;

Käppler, 1994). In this example, the setting protocol consists of four questions about location (14 categories), social context (6 categories), posture and motion (4 categories), and present activity (23 categories). The following 14 items relating to subjective state were assessed using seven-point adjective rating scales (from 1 = not at all, to 7 = completely): Is the momentary situation familiar, typical? Is the present situation strenuous, demanding? Is the momentary situation under your control? Do you feel excited, nervous? Is your momentary mood angry, irritated? Is your momentary mood depressed? Do you feel mentally alert? Do you feel weary, exhausted? Do you feel physically well? Did any specific events occur in the preceding interval? (yes – no). Did you experience stress in the preceding interval? When anticipating the outcome of the following tests, how good will your performance be? Having done the tests how good was your actual performance? Was being tested tiresome to you?

Two questions were open, that is, they allowed individual wording of the answer or statement. Following a yes response to whether particular events occurred, the subject was asked to type a short report. When the self-rating indicated that the physical well-being was diminished, that is, a response of 6 or less on the seven-point rating scale, the subject had to specify the complaint or symptom.

Mental performance tests may include tests of attention, for example a Go/No-Go test of reaction time, or a working memory test (adapted by G. Brügner from Zimmermann & Fimm, 1992/93). The Go/No-Go test consists of 60 stimuli. The working memory span test uses 80 stimuli, that is, geometrical symbols: circle, square, diamond, triangle, and cross. The target stimuli was, in each case, the last but one symbol. When a stimulus matched the target stimulus, the subject had to press the key.

MONITOR can be easily adapted to include, for example, retrospective self-ratings for the investigation of recall-error or specific items which relate to the experience of being hindered or distracted by the PC during on-going activities and, as well, of being distracted by external events or persons when responding.

Overview of recent findings

In the following, the findings from a number of recent studies based on MONITOR are reviewed. The hand-held PC is particularly suited to the assessment of diurnal changes in subjective state and performance, for example, in research on aspects of morningness-eveningness. With this methodology an investigation of recall errors with regard to momentary self-ratings and retrospective ratings can be carried out: findings from four studies are available. Repeated ambulatory assessment of behavioral activities and subjective state may contribute to the concurrent validation of psychological tests, for example, measures of personality traits derived from questionnaires.

The MONITOR software was used also in a number of studies reported in this volume (for example, by Käppler & Rieder, by Kubiak & Hermanns, by Schmidt et al.; Stiglmayr et al.).

Mood, attention, and morningness-eveningness

A number of studies have been concerned with the concept of morningness-eveningness. The assumption was that mood, performance, and certain physiological variables like body temperature, skin conductance, and endocrine measures, are dependent on or associated with the individual's basic disposition to exhibit the optimal level of energetic arousal, behavioral activation, and well-being either in the morning or in the evening. A questionnaire to assess this disposition, the Morningness-Eveningness Questionnaire, MEQ, was developed by Horne and Östberg (1976). The empirical findings, however, appear to be rather inconsistent (cf. review by Kerkhof, 1985; Natale & Cigogna, 1996; Tankova, Adan & Buela-Casal, 1994). The circadian regulation of subjective alertness and the individual tendency towards morningness-eveningness may reflect biological rhythms but is also strongly related to socio-cultural factors. Accordingly, sociability and the age-related style of living (life habits) may exert an essential influence on empirical investigations, for example, early and late evening social activities may not only be indicative of eveningness, indeed they may also confound the assessment of mood and performance the next morning.

Nearly all investigations relied on retrospective questionnaires or paper and pencil diaries. Studies that included objective performance tests are an exception. The ambulatory assessment methodology appears to be especially suited to the investigation of the relationships between self-reports on mood, and objective performance tests, morningness-eveningness, and personality dimensions. The present investigation sets its sights on revealing whether morning types and evening types differ in self-reported mood and performance during the course of the day. In particular, it was expected that morning types would exhibit more positive mood and higher performance levels in the morning than evening types (Fahrenberg, Brügner, Foerster & Käppler, 1999).

In this study, 61 university students from various faculties, 24 males and 37 females, participated as paid voluntary subjects. The mean age was 24 years (SD = 3.2, range 20 to 33). In the present study the assessment included: a setting protocol, self-ratings of momentary subjective state, retrospective ratings, and performance in two tests of attention, that is, a Go/No-Go paradigm and a working memory span test (see Section above). The participants received the computer either on Monday or on Wednesday morning at about 9 a.m. and they returned it 48 hours later on Wednesday or Friday morning, respectively. The subjects were trained to operate the PC, paying special attention to the recording of self-ratings according to the different kinds of items and the taking of performance tests. The MONITOR prompted the first and subsequent self-protocols at about 12 a.m., 3 p.m., 6 p.m., and 9 p.m. The data acquisition was automatically resumed at 8 a.m. the following morning.

The participants filled in the adapted German version of the MEQ (Horne & Östberg, 1976). In the present study, two essential items from the MEQ were used instead of the total score of this rather heterogeneous questionnaire. A classification was based on self-ratings of type (Item 17) and statements about "best time of the day" (Item 16): 17 morning types, 24 intermediate types, 20 evening types.

A MANOVA (3 groups, 2 days, 5 protocols) was conducted to test the hypothesis on relationships between morningness-eveningness and the daily course of mood and attention. In this respect, the interaction term (protocols x MEQ) was of special interest. Several significant main effects ($p \leq .01$) were present. *Between days*: 5 items and test scores; *between protocols*: 6 items and 4 test scores. On day 2, participants rated the situation more "familiar", felt less "excited", less "mentally alert", rated their physical well-being and test performance higher, and had better performance scores in both tests of attention. Diurnal changes were evident in items "strenuous, demanding", "angry, irritated", "mentally alert", and "stress", which all showed an inverted U-shape course, and self-rating of test performance, which reflected an increase over protocols. Performance measures like mean reaction time and standard deviations showed a U-shape, that is, decrease and increase during the day. The *interaction terms* morningness-eveningness x protocols were significant for five measures. Significant effects were found for "strenuous, demanding" (F (8,80)= 2.90, p = .007), "situation under control" (F (8,80) = 3.32, p = .003), self-rating of performance after testing (F (8,80) = 2.64, p < .05), irritated by the repeated questioning (F (8,80) = 3.17, p = .004), and performance index working memory (F (8,92) = 2.07, p < .05). Morning types and evening types differed as predicted.

Within-subject correlation coefficients, pooled across subjects, were moderate to low, for example between "situation under control", "excited, nervous", "angry, irritated" and "stress since last protocol". With the exception of self-ratings of test performance, not a single relationship (r > .20) was found between changes in performance scores and subjective state. This result may be open to question because there were comparatively large intervals, that is, about three hours, between protocols. Furthermore, the statistical distribution of item responses and over-all trends may have affected the results. However, the extension of the assessment over 48 hours and the substantial state variance in most variables appear to support this result.

Discrepancies between subjective state and behavior, although in accordance with many findings in ambulatory assessment research (cf. Fahrenberg & Myrtek, 1996), may not be evident in traditional investigations based on paper-and-pencil questionnaires and retrospective ratings. A direct data acquisition, especially by means of the ecologically more valid ambulatory assessment, is preferable whenever possible.

Negative retrospection effect

The recall-error or retrospection effect has been a methodological issue in several studies. Even though subjects are told to monitor their stress responses, mood and symptoms several times daily, such questionnaire reports are often done from memory. These retrospective ratings are, of course, less accurate. In particular, self-ratings of mood and symptom-reports obtained in the evening may not represent the prevailing state during daytime. A specific retrospection effect may exist when subsequent events and experiences systematically influence and even distort the subjective evaluation and weighting of previous states.

A number of studies have shown discrepancies between actual and retrospective self-ratings. Such discrepancies may be indicative of the subjective evaluation and weighting of previous states, and possibly point to a general response set or bias. (De Beurs, Lange & Van Dyck, 1991; De Longis, Folkman & Lazarus, 1988; De Longis, Hemphill & Lehman, 1992; Hedges, Jandorf & Stone, 1985; Margraf & Jaokobi, 1997; Margraf, Taylor, Ehlers, Roth & Agras, 1987; Shiffman et al., 1997; Smith & Safer, 1993; Stone et al., 1998; Thomas & Diener, 1990).

Käppler, Becker, and Fahrenberg (1993) compared the averages of self-ratings, which were recorded at intervals of about 30 minutes, with the summarizing retrospective ratings, done next morning from memory. The retrospective ratings indicated more negative mood and unease than was to be expected from the actual ratings averaged across the day, that is, a *negative* retrospection bias. The daily course was rated as being more strenuous and the mood more nervous, depressed and weary, , etc., than would have been expected from the averages of repeated self-ratings during the day.

The second study, based on a larger sample of subjects and a 48-hour, allowed for a valid test and generalization of this negative bias in retrospective self-ratings. Sixty-one students participated in this investigation with five self-reports each day for two consecutive days (see morningness-eveningness study). At four particular points of time during the 48 hr monitoring, the subject, after completing the usual protocol, was prompted to provide retrospective ratings about the same set of items, that is, without the performance test, for the whole day. These retrospective summary ratings were obtained in the evening of and in the morning following each of the two days of monitoring.

Once again, a negative bias in self-ratings made in the evening or morning following the day being assessed was evident. The situations were rated as being more strenuous, demanding, less under control, and the mood as being more negative than during the day. These findings substantiated research by De Longis et al. (1988, 1992) and Hedges et al. (1985), and emphasized more clearly the uniformly negative tendency of this retrospection bias. A medium to large effect size was noted here. Hence, the negative retrospection effect, thus, was well replicated. It fits nicely into this construct that a significant correlation between the magnitude of the negative retrospection effect and the personality dimension Emotionality (Neuroticism) was found (Fahrenberg, et al., 1999).

A noteworthy finding was that only one of the ten items, “physical well-being”, did not show a significant deviation between daily average and retrospection. It may be speculated that the participants of this study, healthy students, had little variance in their well-being. However, mean and standard deviation were not particularly conspicuous, except for a tendency to exhibit a ceiling effect, indicating well-being (7 = completely). It appeared worthwhile to further investigate the retrospection bias in these positive items. One possible explanation of this effect may concern the distributions and the noticeable tendency to the mean. The retrospective self-ratings were all nearer to the scale midpoint (4) than the means of the momentary ratings. This finding may be interpreted as a regression to the mean because of a less than perfect reliability or as a heuristic used in a decision situation characterized by a context of uncertainty. This radical interpretation does not appear to be compatible with the finding that is also plausible in personality theory, of a correlation between the negative retrospection effect (magnitude of the distortion) and Emotionality. Individuals with a high degree of Emotionality are known to experience depressed mood, many somatic complaints, and a problematic illness behavior as obvious in large scale representative surveys and large-scale investigations (for example, Fahrenberg, Hampel & Selg, 2001; Myrtek, 1998).

A third study was based on a sample of employed persons with a view to testing and generalizing the previous findings. Special care was taken to also include "positive" items, that is, positively worded adjective rating-scales, and to control for the possible inference aspect on the distribution of means and for the tendency towards the mean (Scheibehenne; Saller, Riemann, and Fahrenberg, 2000). A total of 59 participants from different professions (25 m, 34 f), ages from 19 to 58 (M = 34, SD = 10.6) participated in this monitoring study. Most of them worked in an office and had higher education. MONITOR was used to assess (1) setting and activities (4 variables), and (2) the momentary subjective state (14 adjective rating scales) five times during the course of a typical working day.

The findings indicated a highly significant retrospection effect in five of the ten mood items and in four items concerning the situation and distracting influences when responding (Figure 1). In retrospection, the day was valuated more negatively: the situations less familiar, more strenuous, less under control, more distracted; the mood was rated more excited, nervous; angry, irritated; depressed; mentally alert. In eight items, the item means tended toward the scale midpoint ("4"); the only exception was "active, ready to go". The "positive" items, that is, "physical well-being", and the newly included "balanced", "self-confident, self-assured", and "in good mood", showed neither a retrospection effect nor a tendency to the scale midpoint.

Figure 1

The personality trait "Emotionality" (FPI-R, Fahrenberg et al., 2001) was correlated with the magnitude of the retrospection effect in two items: strenuous, demanding r = .28 (p = .03, df = 57) and excited, nervous r = 0.27 (p = .04). Furthermore, significant correlation coefficients were noted, for example, between questionnaire scales *Physical Complaints* and "mentally alert" 0.35(p =.006) and "strenuous, demanding" .37 (.004), and FPI-R scale *Strain* and "strenuous, demanding" .44 (p =.000) and "excited , nervous" .30 (p= .023). Thus, the basic findings of the previous study were replicated. However, the inclusion of more "positive" items was generally not very revealing here. A noteworthy exception was the item "active, ready to go" which exhibited a positive retrospection effect and thereby a large deviation from the scale midpoint. In the retrospective protocol, the participants answered a few questions relating to this task (the distribution of "yes" and "no" are given in parenthesis): Was your retrospective rating guided by your typical answers given during the day ? (46 % yes), by the momentary mood ? (31 % yes), by specific events during the day ? (56 % yes). Are you inclined to evaluate days in retrospection differently ? (39 % yes). The self-rating of being inclined to evaluate days in retrospection differently, concurs with a smaller difference (that is, smaller retrospection effect) in certain items: "excited, nervous", "angry, irritated", "depressed", "tense", and less "balanced" (r significant, p < .05). The correlation coefficients indicate that some participants actually were aware of their retrospective interpretation and weighting of daily events.

This series of investigations replicated the previous finding that the recall-error in such self-reports was biased, that is, self-ratings were negatively distorted when summarized in the evening or the following morning. Obviously, neither the average rating nor the retrospective rating can be considered the "true" value since both reflect psychological aspects of state and state change. Nonetheless, the retrospection bias may undermine the validity of self-ratings. A direct data acquisition, especially by means of the ecologically more valid ambulatory assessment, is preferable, whenever possible.

Personality questionnaire FPI-R and MONITOR data

Test performance measures and self-ratings of activities and mood, repeatedly assessed by hand-held PC, provide an adequate means of investigating the variability and consistency between and within subjects and between settings. Pawlik and Buse (1982) were the first to utilize the ambulatory assessment methodology in research on the situationism – interactionism – traitism controversy. The previously used data from questionnaires, for example, concerning anxiety responses in various situations (done from memory), may be heavily biased due to recall errors and a general lack of ecological validity.

The feasibility of computer-assisted self-reports on activities and mood in daily life suggests also a new approach in the evaluation of personality tests, especially regarding the concurrent validation of trait scores derived from personality inventories. Preliminary findings on relationships between scores from the *Freiburger Persönlichkeitsinventar* FPI-R and ambulatory monitoring data on subjective state (Fahrenberg et al., 1999; Heger, 1990; Käppler, 1994) led to further investigations. The FPI-R is a German 138-item personality inventory comprising ten personality scales, namely, *Life Satisfaction, Social Orientation, Achievement Orientation, Inhibitedness, Excitability, Aggressiveness, Strain, Somatic Complaints, Health Concern, Frankness, and two further scales, Extraversion* and *Emotionality*, which can be considered to be equivalent to the corresponding scales of the personality questionnaires developed by Eysenck. Normative data are available from a large sample, N = 3740, representative for the German population (Fahrenberg et al., 2001).

An ambulatory monitoring study was designed to investigate the concurrent validity of trait measures and averages of self-ratings level averaged across one day. A total of 59 participants from different professions participated in this monitoring study (sample and general design, see above, study on negative retrospection effect). MONITOR was used to assess: (1) setting and activities (4 variables), and (2) the momentary subjective state (14 adjective rating scales) five times during the course of a typical working day.

The statistical analysis revealed a large proportion of significant ($p < .05$) relationships between personality scores and means of self-ratings: 35 % of a total of 168 coefficients (12 trait scales, 14 state items). Noteworthy was the distribution of these coefficients. Substantial coefficients ($r > .50$) were found, for example, for *Emotionality* and “depressed”, *Life Satisfaction* and “well-balanced” and “in a good mood”, *Strain* (trait) and experienced “stress” (state), *Health Concern* and “excited, nervous”. Most of the significant relationships were with FPI-R Scales *Emotionality, Life Satisfaction, Physical Complaints, Health Concern,* and *Excitability*; only negligible coefficients of correlation were found regarding *Social Orientation, Aggressiveness,* and *Extraversion* (Fahrenberg et al., 2001; Scheibehenne et al., 2000).

Activities and subjective state across the period of one week

A 7-day ambulatory assessment was made to investigate the validity of FPI-R scales to predict the averages and, in particular, the variability of activities and of certain mood items in daily life. An extended monitoring would allow for more data regarding interindividual differences in intraindividual variability. The intention was to replicate the previously observed relationships concerning personality trait scales and item averages and to investigate the predictability of the standard deviation of self-ratings during the week. The hypotheses were, for example, that trait measures of *Emotionality, Strain, Physical Complaints, Health Concern*, would predict the level and, possibly, the variability in the aforementioned items. Furthermore, a 7-day-assessment could be used to test the negative retrospection effect once more, whereby daily and weekly retrospection could be compared.

The participants in this study were 28 first-year students (27 f, 1 m, mean age M = 23.4, SD = 4.6) enrolled in Psychology courses. They received credits for participating in a 7-day monitoring. The data were recorded by MONITOR which was programmed to obtain six records, whereby the first prompt in the morning was timed to occur half an hour after the usual rising time of the subject. MONITOR prompted the subsequent self-protocols at about 11 a.m., 1 p.m., 4 p.m., and 9 p.m. The assessment began at different days of the week in order to balance the assumed weekend effect: 8 subjects began their self-reports on a Monday, 9 subjects on Wednesday, and11 subjects on Friday. The participants could use the start option, in case they were unable to answer the prompt immediately. Only 3 % of the 28 x 7 x 6 entries were missing. At the beginning, the subjects were given the Freiburger Persönlichkeitsinventar FPI-R (Fahrenberg et al., 2001).

The findings showed relationships between personality measures and weekly averages and variability of several items, for example: *Emotionality* and "excited, nervous" Mean (.52, p = .004) and SD (.43, p = .004); "depressed" Mean (.47, p = .01) and SD (.44, p= .02); "weary, exhausted" Mean (.47, p = .01); "in a good mood" Mean (-.49, p = .01); "physically well" Mean(-.44, p = .02). A number of significant relationships were also found between personality trait *Irritability* and items like "angry, irritated", "depressed", and "exhausted"; and for personality traits *Strain* , and *Health Concern*. However, *Extraversion* was only related to the SD of the item "positive changes since previous self-report". It was concluded that the personality test score *Emotionality* predicted the general level and variability of subjective state under naturalistic conditions substantially. The relationship between personality measures and the frequency distributions of certain settings variables, for example, location, social contact, and activities, will be investigated.

As previously noted, a negative retrospection bias was present in 8 of the 10 items used in this study (p < .01 in 8 of these items; based on 190, that is 28 x 7, days). The items means changed towards the scale midpoint in 7 items and to the opposite direction in 2 items. The magnitude of this effect was not related to the personality trait *Emotionality* in this study.

Perspectives, developments, and applications

Evaluation and selection of items for diaries

The selection of appropriate items and the psychometric evaluation of the resulting test scales are basic tasks in test construction. This strategy may generally be valid for performance tests used in ambulatory assessment (Buse & Pawlik, 1996; Pawlik & Buse, 1996). However, there are some methodological aspects at issue and, as far as the assessment of subjective state is concerned, there may be essential deviations from the conventional strategies of item analysis and test construction for several reasons: First, an unusually large between- and within-subject variance can be expected for many items assessing state and state change and, therefore, ceiling or floor effects in distributions may often occur; second, the repeated measurement design does not permit the inclusion of more than a minimum of adjective rating scales so that the reliability (consistency), which relates to the concept of parallel measurement and item total correlation coefficients, is hardly practicable; third, since state measures are being sought, retest-reliabilities may be of little value for the evaluation of psychometric properties and the selection of items; fourth, in self-ratings of mood, and especially in performance data, circadian and septemdian (weekly) trends, and moreover, training, satiation and other effects may be obvious; and, fifth, the results of the statistical analyses are subject to controversy, due to such dependencies and the presence of auto-correlation in repeated measurements. Such issues were discussed, for example, by Suen and Ary (1989), Pawlik and Buse (1996), Delespaul (1992), Ott and Scholz (this Volume), Schwartz and Stone (1998), West and Hepworth (1991), and Wilhelm (this Volume).

Even if researchers refrained from performing a statistical test on their hypotheses and were content with descriptive and exploratory analysis, a number of methodological issues nevertheless would persist, especially those dealing with the selection of appropriate items.

Statistical analyses could, however, assist in revealing which items are especially suited to the assessment of *diurnal* changes in subjective state: over-all changes *common to all subjects* and changes *particular to certain subjects*. Other items may be more likely to indicate *habitual* differences in mood. The aforementioned investigation in 61 students may be used as an illustration as to how possible differences in item sensitivity to state changes can be explored.

The 14 items clearly differed in their statistical distributions. Item means ranged between M = 1.85 ("angry, irritated") and M = 5.79 ("situation under control"); SD varied between SD = 1.23 ("depressed") and SD = 1.83 ("testing tiresome"). Accordingly, two items ("situation under control", "physical well-being") showed a ceiling effect and two items ("angry, irritated", "depressed") showed a floor effect. Between days, as revealed by the MANOVA main effect, changes were evident. Participants were more at ease on day 2 and had better performance scores, thus indicating a process of adaptation and training. Significant main effects between protocols indicated that patterns of diurnal changes, common to all subjects, were present (for a further discussion, see Käppler, et al., 2001).

A two-factorial ANOVA (11 protocols x 61 subjects) was used to obtain the relative amount of variance explained by protocols, subjects, and the interaction term protocols x subjects. Between-subject variance was highest, that is, greater than 20 percent, in "situation under control", "angry, irritated", "depressed", "physical well-being", "stress", and "anticipated performance", and was lower, that is, smaller than 10 per cent, in "excited, nervous" and "mentally alert". The between-protocols variance indicating diurnal mood and performance changes common to all subjects was generally low, for example, 3 percent for "strenuous, demanding" and "stress" and even lower for the remaining items.

Substantial interaction variance subjects x protocols was present in most of the items which showed that subjects gave specific self-ratings and performed differently depending on the time of day. The interaction term, which is indicative of diurnal changes specific to individuals, was substantial for "mentally alert" (74 %), "anticipated performance" (75 %), "excited, nervous" (65 %), "testing tiresome" (54 %), and, on the other hand, negligible or low for "under control" (0 %), "depressed" (0 %), and "strenuous, demanding" (12 %). In these items, however, the residual component, which includes higher order interactions that are not separable from error, was substantial.

Coefficients of test-retest reliability (stability) between protocol 1 and protocol 2, of the first day, were between .08 for "strenuous demanding" and .47 for "physical well-being". Such coefficients were of roughly the same range in protocol 1/ protocol 6, that is, the first protocols, of each day, and they were slightly higher between protocol 2 and 7, that is, the second protocols of each day. Relationships, of course, were substantially higher when averages, protocol 1 to 5 and protocol 6 to 10, were compared: from .44 "situation under control" to .67 "physical well-being".

Within-subject correlation across protocol averages (pooled for 61 subjects) was obtained to evaluate the covariation of measures. Such coefficients were moderate to low, and there was no indication that certain items from this item pool could be considered redundant, that is, a high ($r^2 > .50$) or very high common variance. None of the items should be eliminated on these grounds.

Based on large-scale assessment studies in students and other populations, Buse and Pawlik (1994; Pawlik and Buse, 1996) developed a battery of performance tests and an item pool suited for ambulatory assessment studies. This may signal the beginnings of the standardization in this domain, possibly based on a core set of normative data. Reference data of this kind could be valuable for many studies. At least, a selection of psychometrically proven items (and item formats) to assess activity level, basic mood dimensions, and workload (job stress) are desirable.

However, the research questions and the applications in many other fields of psychology will require specific assessment strategies in dealing with item contents and sampling (persons, settings, hours, days, etc.) and, therefore, the study protocols, diaries, and symptom reports will remain highly diverse. Thus, the collection of reference data within individual research institutions will be an essential task until cross-institutional standardization is hopefully one day achieved. On the whole, the computer-assisted assessment methodology, appears to be far from standardized, a serious shortcoming compared to conventional psychological testing. The flexibility in programming a PC for recording self-reports, on the other hand, allow for new psychological assessment strategies and other applications of PC technology in daily life.

Compliance and acceptance

Ambulatory assessment by means of hand-held computers is a feasible method. This approach appears to be especially suited to the investigation of diurnal changes and, based on a few items or symptoms only, to long-term assessments extending over many days or weeks.

The assumption, that the general compliance must be perfect because of the computer-assisted method of assessment, may not be true in each participant. The signal can be over-heard, unintentionally or intentionally, so that one or even a number of protocols may be missing. Since the frequency of such missing protocols in a series of studies usually is in the order of a few percent only, a *statistical* analysis exploring such effects and possible relationships with other variables may be pointless. From the protocol provided by means of the hand-held PC, individual differences in the delay in responding (signal compliance) and in duration of input can be noted (or, using MONITOR, also the number of intentional delays when on-going activities preclude immediate responding).

However, distraction and irritation may be caused by the interruption of the on-going activity by the beeper signal and the requested responding. This may be disturbing to the participant. Furthermore, an increased state of self-awareness or distinct avoidance behavior, that is, a tendency to evade settings where responding to the items is found to be disturbing for one-self and for others, may occur. The common use of mobile phones will ease much of these reservations. The comments from participants of such monitoring studies indicate that the primary complaint was the frequency of protocols requested by this methodology. The acceptance and compliance will depend also on the number of items, protocols and the length of study (days). From a series of such studies it may be concluded that the motivation may differ between student subjects (with and without credit points or financial rewards) and patients. Patients can be highly motivated, when there is an evident relationship between the assessment and clinical diagnosis or treatment evaluation.

There are two practical approaches to obtaining evaluating comments from participants: A post-monitoring interview or questionnaire and the use of specific questions to be included in the item pool for concurrent responses during the ongoing monitoring protocol.

A written post-monitoring interview was regularly used in our monitoring studies. For example, in one study, 61 students were given a 17-item questionnaire after returning on the third day asking them about compliance, acceptance, possible reactivity induced by such monitoring, and general criticism about this investigation, and a self-evaluation concerning both the role of distracting events during responding and enhanced states of self-awareness. The participants gave their evaluation of the investigation on seven point rating scales (1 = not at all, 7 = completely). Interest in the present study, for example, was rated at 4.9 on average, typicality of the two days compared to one's usual life at 4.8, interference of repeated self-ratings with daily routine at 3.2, enhanced self-awareness at 3.6, and method-dependent (reactive) changes in behavior at 2.3. Nearly all of the subjects were willing to participate in subsequent studies. Corresponding ratings obtained from the study of 59 employed persons showed very similar results. These subjects were asked whether they noticed a heightened awareness of their behavior on that same day (17 % yes) and whether they had modified their daily routine to adapt to and cope with the monitoring (3 % yes). In general, all of the participants volunteered for another ambulatory monitoring.

In two recent studies, specific items were included in MONITOR concerning the role of distracting events during responding. Commencing the list of items, the participant was asked whether distracting or irritating influences were present during answering. The responses indicated some effects (M = 1.98, SD = 1.46), the Mode = 1 showed, however, that in most instances there was no distraction present. For this item, there were no relationships found with the personality measures (FPI-R), however, small coefficients of correlation with concurrent self-ratings “active, ready to go” (r = .27, pooled across subjects, df = 57), “mentally alert” (.23), “excited, nervous” (.22), and “stress” during past period (.21). In the 7-days monitoring study with 28 participants, correlation coefficients between the 42 momentary ratings (M and SD) of being disturbed by the beeper and personality measures from the FPI-R were insignificant.

The compliance and the generally positive attitude seen in participants were noteworthy, but it should be taken into account that in many investigations they were paid volunteers. However, an outstanding degree of compliance in such ambulatory assessment has also been reported in patient groups and other populations (cf. Fahrenberg & Myrtek, 1996).

Conclusions

Ambulatory assessment of subjective state, symptom reports, or job stress by hand-held PCs (PDAs, “electronic diaries”) appears to have many advantages compared to the Experience Sampling Method ESM (de Vries, 1992), which used a beeper and a booklet, that is, paper-and-pencil method. Reviews on self-report and self-monitoring methodologies indicate that questionnaires and diary forms seem to be still the standard methodology to assess self-reports. However, the shortcomings of these methods are also acknowledged (Pawlik & Buse, 1996; Perrez, 1994; Stone et al., 2000; Suls & Martin, 1993; Wheeler & Reis, 1991; Wilz & Brähler, 1997; Hank, Schwenkmezger & Schumann, this Volume).

The present volume contains a number of investigations which likewise used a computer-assisted methodology to assess psychological data, that is subjective state or behavioral reports, in daily life. Noteworthy applications have been reported in clinical psychology and at the workplace. Data acquisition software like MONITOR will be of practical use. It became evident that computer-assisted methodologies enable innovative strategies in self-monitoring of job stress and of symptoms. Further developments may include new strategies in health care and self-management, for example, of conditions of chronic illness.

Acknowledgements

We acknowledge Miriam Jakob, Benjamin Scheibehenne, Thomas Saller, and Andreas Sieburg, psychology students, for their assistance, especially in the acquisition of data. We thank Marcus Cheetham for improving the English manuscript.

Figure Legends

Figure 1. Negative retrospection effect when the retrospective self-rating in the evening is compared to the five momentary ratings averaged across the day.

Note: ^{1} Sign reflected in this Figure. ** p< .01; *** p< .001, according to t-test and Wilcoxon test, as well.

References

Brügner, G. (1998). MONITOR: Ein flexibles Programm zur Datenerhebung mittels Pocket-PC. *Zeitschrift für Differentielle und Diagnostische Psychologie, 19,* 145–147.

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 performance testing. In J. Fahrenberg & M. Myrtek (Eds.). *Ambulatory Assessment: computer-assisted psychological and psychophysiological methods in monitoring and field studies* (pp. 29–50). Seattle, WA: Hogrefe & Huber.

De Beurs, E., Lange, A. & Van Dyck, R. (1991). Self-monitoring of panic attacks and retrospective estimates of panic: Discordant findings. *Behaviour Research and Therapy, 30,* 411–413.

Delespaul, P.A.E.G. (1992). Technical note: devices and time-sampling procedures. In M. de Vries (Ed.), *The experience of psychopathology. Investigating mental disorders in their natural settings (*pp. 363–373). Cambridge: Cambridge University Press.

de Vries, M.W. (Ed.). (1992). *The Experience of Psychopathology. Investigating Mental Disorders in their Natural Settings.* Cambridge: Cambridge University Press.

DeLongis, A., Folkman, S. & Lazarus, R.S. (1988). The impact of daily stress on health and mood: Psychological and social resources as mediators. *Journal of Personality and Social Psychology, 54*, 486–495.

DeLongis, A., Hemphill, K.J. & Lehman, D.R. (1992). A structured diary methodology for the study of daily events. In F.B. Bryant, J. Edwards, R.S. Tindale, E.J. Posavac, L. Heath, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Methodological Issues in Applied Social Psychology. Social Psychological Applications to Social Issues.* (Vol. 2, pp. 83–109). New York: Plenum.

Fahrenberg, J. (1996). Ambulatory Assessment: Issues and perspectives. In J. Fahrenberg & M. Myrtek (Eds.). *Ambulatory Assessment: computer-assisted psychological and psychophysiological methods in monitoring and field studies* (pp. 3–20). Seattle, WA: Hogrefe & Huber.

Fahrenberg, J., Brügner, G., Foerster, F. & Käppler, C. (1999). Ambulatory assessment of diurnal changes with a hand-held computer: Mood, attention, and morningness – eveningness. *Personality and Individual Differences*, *26*, 641–656.

Fahrenberg, J., Hampel, R. & Selg, H. (2001). *Freiburger Persönlichkeitsinventar FPI-R* (7th ed.). Göttingen: Hogrefe.

Fahrenberg, J., Foerster, F., Schneider, H.J., Müller, W. & Myrtek, M. (1984). *Aktivierungsforschung im Labor-Feld-Vergleich.* München: Minerva.

Fahrenberg, J. & Myrtek, M. (Eds.) (1996). *Ambulatory Assessment: computer-assisted psychological and psychophysiological methods in monitoring and field studies.* Seattle, WA: Hogrefe & Huber.

Fahrenberg, J. & Myrtek, M. (2001). Ambulantes Monitoring und Assessment. In F. Rösler (Ed.), *Enzyklopädie der Psychologie: Themenbereich C Theorie und Forschung, Serie 1 Biologische Psychologie, Band 4 Grundlagen und Methoden der Psychophysiologie (pp. 657–798).* Göttingen: Hogrefe.

Finegan, J.E. & Allen, N.J. (1994). Computerized and written questionnaires: Are they equivalent? *Computers in Human Behavior, 10*, 483–496.

Franke, G.H. (1998). *Computerunterstützte klinisch-diagnostische Selbstbeurteilungsverfahren im Äquivalenztest. Experimentelle Studien.* Lengerich: Pabst Science Publishers.

Hank, P. & Schwenkmezger, P. (1996). Computer-assisted versus paper-and-pencil based self-monitoring: An analysis of experimental and psychometric equivalence. In J. Fahrenberg & M. Myrtek (Eds.), *Ambulatory Assessment. Computer-assisted psychological and psychophysiological methods in monitoring and field studies* (pp. 85–99). Seattle, WA: Hogrefe & Huber Publishers.

Hedges, S.M., Jandorf, L. & Stone, A.A. (1985). Meaning of daily mood assessments. *Journal of Personality and Social Psychology, 48,* 428–434.

Heger, R. (1990). *Psychophysiologisches 24-Stund en-Monitoring. Methodenentwicklung und erste Ergebnisse eines multimodalen Untersuchungsansatzes bei 62 normotonen und blutdrucklabilen Studenten*. Frankfurt. a. M.: Lang.

Horne, J.A. & Östberg, O. (1976). Individual differences in human circadian rhythms. *Biological Psychology, 5,* 179–190.

Hüttner, P. (2001). *MONITOR Manual.* Forschungsgruppe Psychophysiologie. Department of Psychology. University of Freiburg, Germany.

Käppler, C. (1994). *Psychophysiologische Bedingungsanalyse von Blutdruckveränderungen im alltäglichen Lebenskontext.* Frankfurt a. M.: Lang.

Käppler, C., Becker, H.-U. & Fahrenberg, J. (1993). Ambulantes 24-Stunden-Monitoring als psychophysiologische Assessmentstrategie: Reproduzierbarkeit, Reaktivität, Retrospektionseffekt und Bewegungskonfundierung. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14,* 235–251.

Käppler, C., Brügner, G. & Fahrenberg, J. (in press, 2001). Computer-unterstütztes Assessment mit MONITOR: Befindlichkeit und Aufmerksamkeit im Alltag. *Zeitschrift für Differentielle und Diagnostische Psychologie.*

Kerkhof, G.A. (1985). Inter-individual differences in the human circadian system: A review. *Biological Psychology, 20*, 83–112.

Margraf, J. & Jacobi, F. (1997). Marburger Angst- und Aktivitätstagebuch. In G. Wilz & E. Brähler (Eds.), *Tagebücher in Therapie und Forschung. Ein anwendungsorientierter Leitfaden* (pp. 139–153). Göttingen: Hogrefe.

Margraf, J., Taylor, C.B., Ehlers, A., Roth, W.T. & Agras, W.S. (1987). Panic attacks in the natural environment. *Journal of Nervous and Mental Disease, 175*, 558–565.

Myrtek, M. (1998). *Gesunde Kranke – kranke Gesunde.* Bern: Huber.

Myrtek, M., Brügner, G., Fichtler, A., König, K., Müller, W., Foerster, F. & Höppner, V. (1988). Detection of emotionally induced ECG changes and their behavioral correlates: A new method for ambulatory monitoring. *European Heart Journal, 9* (Supplement N), 55–60.

Myrtek, M., Brügner, G. & Müller, W. (1996). Interactive monitoring and contingency analysis of emotionally induced ECG changes: Methodology and applications. In J. Fahrenberg & M. Myrtek (Eds.). *Ambulatory Assessment: computer-assisted psychological and psychophysiological methods in monitoring and field studies* (pp. 115–127). Seattle, WA: Hogrefe & Huber.

Natale, V. & Cigogna, P. (1996). Circadian regulation of subjective alertness in morning and evening "types". *Personality and Individual Differences, 20*, 491–497.

Pawlik, K. (1995). Persönlichkeit und Verhalten: Zur Standortbestimmung von differentieller Psychologie. In K. Pawlik (Ed.), *Bericht über den 39. Kongreß der Deutschen Gesellschaft für Psychologie in Hamburg 1994* (pp. 31–49). Göttingen: Hogrefe.

Pawlik, K. & Buse, L. (1982). Rechnergestütze Verhaltensregistrierung im Feld: Beschreibung und erste psychometrische Überprüfung einer neuen Erhebungsmethode. *Zeitschrift für Differentielle und Diagnostische Psychologie, 3,* 101–118.

Pawlik, K. & Buse, L. (1996). Verhaltensbeobachtung in Labor und Feld. In K. Pawlik (Ed.), *Enzyklopädie der Psychologie. Differentielle Psychologie und Persönlichkeitsforschung. Band 1. Grundlagen und Methoden der Differentiellen Psychologie* (pp. 359–394). Göttingen: Hogrefe.

Perrez, M. (1994) Felddiagnostik mit besonderer Berücksichtigung der computerunterstützten Diagnostik. In R.-D. Stieglitz & U. Baumann (Eds.), *Psychodiagnostik Psychischer Störungen* (pp. 149–161). Stuttgart: Enke.

Scheibehenne, B. Saller, T., Riemann, D. & Fahrenberg, J. (2000). Befinden im Tageslauf. Zwei Untersuchungen mit MONITOR. Research Reports. Institute of Psychology, University of Freiburg, Germany.

Schwartz, J.E. & Stone, A.A. (1998). Strategies for analyzing ecological momentary assessment data. *Health Psychology, 17,* 6–16.

Shiffman, S., Hufford, M. Hickcox, M., Paty, J.A., Gnys, M. & Kassel, J.D. (1997). Remember that? A comparison of real-time versus retrospective recall of smoking lapses. *Journal of Consulting and Clinical Psychology, 65,* 292–300.

Smith, W.B. & Safer, M.A. (1993). Effects of present pain level on recall of chronic pain and medication use. *Pain, 55,* 355–361.

Stone, A.A. & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine, 16,* 199–202.

Stone, A.A., Schwartz, J.E., Neale, J.M., Shiffman, S., Marco, C.A., Hickcox, M. & Paty, J. (1998). A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of Personality and Social Psychology, 74,* 1670–1680.

Stone, A.A., Turkhan, J.S., Bachrach, C.A., Jobe, J.B., Kurtzman, H.S. & Cain, V.S. (2000). The science of self-report. Implications for research and practice. Mahwah, N.J.: Lawrence Erlbaum.

Suen, H.K. & Ary, D. (1989). *Analyzing quantitative behavioral observational data.* Hillsdale, N.J.: Lawrence Erlbaum.

Suls, J. & Martin, R.E. (1993). Daily recording and ambulatory monitoring methodologies in behavioral medicine. *Annals of Behavioral Medicine, 15,* 3–7.

Tankova, I., Adan, A. & Buela-Casal, G. (1994). Circadian typology and individual differences: A review. *Personality and Individual Differences, 16,* 671–684.

Thomas, D.L. & Diener, E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology, 59,* 291–297.

West, S.G. & Hepworth, J.T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality, 59,* 609–662.

Wheeler, L. & Reis, H.T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality, 59*, 339–354.

Wilz, G. & Brähler, E. (Eds.) (1997). *Tagebücher in Therapie und Forschung. Ein anwendungsorientierter Leitfaden.* Göttingen: Hogrefe.

Zimmermann, P. & Fimm, B. (1992/93). *Testbatterie zur Aufmerksamkeitsprüfung TAP. Department of Psychology. University of Freiburg*: Psytest.