

Validity of Retrospective Time-Use Reports in Old Age

PETRA L. KLUMB* and MARGRET M. BALTES

Research Unit Psychological Gerontology, Free University Berlin

SUMMARY

In a sample of $N = 83$ participants aged between 72 and 97, we assessed the accuracy of time budgets originating from the Yesterday Interview (YI; Moss and Lawton, 1982) by means of comparison with in-situ assessments based on the Experience-Sampling Method (ESM; e.g. Csikszentmihalyi and Larson, 1987). We examined convergent and discriminant correlation patterns from indices of activity participation, location, company, and mood collected via both methods. The correspondence between the two methods appeared acceptable. Furthermore, we explored whether (a) length of the retention interval, (b) congruence with pre-existing knowledge, and (c) congruence of the locations of encoding and retrieval accounted for (in)accuracy of recall. We also analysed the degree to which age and cognitive functioning explained performance differences. While we found discrepancies between YI and ESM to be hardly attributable to context effects or differences in cognitive functioning, there was an age effect. Finally, the sensitivity of the two methods to detect differences between groups was found to be largely equivalent but hardly sufficiently convergent. Copyright © 1999 John Wiley & Sons, Ltd.

INTRODUCTION

A broad range of practical and research questions is tackled on the basis of time-use information. Disciplines as different as anthropology, business administration, economics, psychology, psychiatry, and sociology employ time-use reports (e.g. Curie *et al.*, 1990; Robinson and Nicosia, 1991). Since the inferences drawn depend heavily on the quality of the data used, it is important to know how accurate these are. The aim of the present paper, therefore, is (1) to investigate the validity of 24-hour recall data and (2) to determine sources of inaccuracy of these time-use reports.

Time-use information can be collected in a number of ways ranging from direct observation to retrospective recall over periods up to a lifetime. With regard to research-economic considerations, diaries and interviews about the previous day's activities (e.g. Moss and Lawton's, 1982, Yesterday Interview) lie between these techniques and validation studies have shown their superiority to less expensive substitutes (e.g. Juster, 1986; Robinson, 1985; Scheuch, 1972). But like other retrospective reports, 24-hour recall requires the respondent to (a) select the correct time interval, (b) recall the relevant events, and (c) aggregate the retrieved information into the adequate format (Wheeler and Reis, 1991). It may be that the relative efficiency of

*Correspondence to: P. Klumb, Free University Berlin, Nussbaumallee 38, 14050 Berlin, Germany.
E-mail: klumb@zedat.fu-berlin.de

Contract grant sponsor: German Research Foundation (DFG). Contract grant number: Ba 902/8-1.

assessment still has to be paid with the risk of cognitive and motivational factors compromising the validity of the answers.

Engle and Lumpkin (1992) found that, without cognitive enhancement, college students fail to report 54% of the activities observers saw them perform during a two-hour period the day before they were asked to recall them. The probability of remembering an event is known to be a function of (a) the time interval between encoding and retrieval (associated with forgetting and telescoping, e.g. Loftus and Marburger, 1983; Rubin and Baddeley, 1989), (b) the amount of information supplied by a retrieval cue (e.g. Tulving, 1974), (c) the congruence between encoding and retrieval conditions (e.g. Smith, 1988; Eich and Metcalfe, 1989), and (d) the event's congruence with pre-existing expectations and beliefs (associated with reconstructing additional details from general knowledge, e.g. Burt and Kemp, 1994). Schematic processing can lead to reporting routine activity sequences even though they had not actually been carried out because of the day's atypical course. On the other hand, details of an atypical day may be more readily recalled since whatever made it deviate from routine may have received more thorough processing and may be more salient and, thereby, easily accessible.

Beyond the cognitive aspects, motivational and emotional factors may play a role. Comparing national representative surveys with time diaries, Niemi (1993) suggested that socially desirable activities tend to be overreported. Also, relatedness of an event or an activity to central goals of an individual makes recall more likely (e.g. Cantor *et al.*, 1991; Klinger, 1987; for the formation of goal-related encoding categories see also Ach, 1935). Moreover, the magnitude of both cognitive and motivational biases may be associated with interindividual differences. Age-related performance differences, for instance, are larger for recall than for recognition tasks because recall requires active retrieval processes relying more heavily on processing resources that tend to be reduced in elderly individuals (e.g. Craik and McDowd, 1987). As a case in point, Knäuper and Wittchen (1994) found elderly respondents to be particularly susceptible to effects of prior knowledge. They were more likely than younger ones to circumvent the capacity demands of highly complex standardized diagnostic interview questions by using heuristics.

In the present research, we assessed the accuracy of time budgets originating from the Yesterday Interview (Moss and Lawton, 1982) by means of comparison with *in-situ* assessments via a time-sampling method (cf. Brewer, 1988; Fahrenberg, 1994; Pawlik and Buse, 1982; Totterdell and Folkard, 1992). Furthermore, we explored whether (a) length of the retention interval, (b) congruence with pre-existing knowledge, and (c) congruence of the locations of encoding and retrieval account for (in)accuracy of recall. We also examined the degree to which age and cognitive functioning can explain performance differences. We then analysed convergent and discriminant correlation patterns from indices of activity participation, mood, location, and company collected via both methods. Finally, we examined the sensitivity of the two methods to detect differences between groups.

METHOD

Participants

A sample of 34 women and 49 men ($N = 83$) volunteered for the participation in the intensive time-sampling study. These participants were recruited in the course of the

third measurement occasion of the Berlin Aging Study (BASE). Their age ranged from 72 to 97 years with a mean of 80. With one exception, all were community dwelling, 43 lived alone, 37 with their spouses. Selectivity analyses (Klumb and Baltes, 1995) showed the participants of this study to be a positive selection of the BASE intensive-protocol sample ($N = 516$). The selectivity effects were largest for cognitive functioning (+1.2 *SD*) followed by vision (+1 *SD*) and hearing (+0.8 *SD*) resulting in a restricted range of these measures.

Procedure

Yesterday-Interviews (YI; Moss and Lawton, 1982) were employed to assess the time use of the participants. In the course of this interview, the previous day has to be recalled from waking up to falling asleep regarding the activities the respondent carried out and their durations, the respective locations, and the company present. The general mood level during the first and the second half of the day had to be rated on 5-point scales for two positive (happy, relaxed) and two negative adjectives (bored, lonely). The interview took place in the context where the participants spent most of their time, i.e. in their homes. The respondent could follow the natural sequence of activities in the course of the first round of free recall. After reaching the end of the reconstructed day, the interviewer repeated back to the participant the reported activities along with the time scale in order to enhance the recollection process and to add company and location. The average length of the reported day was about 16 hours ($M = 948$ minutes, $SD = 100$ minutes) and the participants reported an average of 43 activities (range: 14–96). Independent of interviewer and coder ratings of the plausibility of the reported information, all YI were included in the analyses.

To collect reference data for validation purposes, we chose the Experience-Sampling Method (ESM; e.g. Csikszentmihalyi and Larson, 1987), a signal-contingent method (Wheeler and Reis, 1991) for the simultaneous assessment of activities and contexts in everyday situations. The following description is customized to the present context and extended descriptions of design and measures are available from the authors upon request. A portable beeper prompted the participants to fill in an experience-sampling form in a small booklet at five times distributed randomly across the day (software: Lang and Helle, 1994). The items regarding activities and thoughts in the moment the beeper went off were open-ended, as was the question for the location of the participant. If company was present (social context) the first name or initials of the person had to be given together with her relationship to the participant. Eight mood adjectives had to be rated along 5-point scales on which 0 indicated no experience of the particular mood while 4 indicated that the affect was experienced very intensively. Four items stemmed from the Positive-Affect–Negative-Affect Scales (PANAS; Watson *et al.*, 1988) and the remaining four were adopted from the Yesterday-Interview.

Data collected in this way are assumed to be uncompromised by memory, selection, or aggregation problems because of the immediacy of responding. While still based on self-reports, this approach is as close to direct observation as one can get (e.g. Delespaul, 1995) minimizing even social-desirability biases. Two consecutive signals had a minimum distance of 15 minutes and the average inter-signal interval was 150 minutes ($SD = 20$ minutes). This procedure was repeated with different random patterns on six consecutive days, the sampling period. The YI took place on

the first ESM-free day and thereby reconstructed the last day of experience sampling resulting in one day of overlapping measurements. All the analyses reported below are based on this single day.

In contrast to other ESM studies (e.g. Delespaul, 1995), we employed fewer time samples (but see Carstensen *et al.*, in preparation) and each participant could determine period restrictions in order to reduce the intrusiveness of the time-sampling procedure. The start and the end of the daily signalling window were set to equal the average waking-up and falling-asleep times as reported by the participant during the briefing session at the beginning of the sampling period. The beginning of the window was found to be, on average, one hour after the time of waking up reported in the YI ($M = 68$ minutes, $SD = 74$ minutes) and its end was at about one hour before the time of falling asleep reported in the YI ($M = 62$ minutes, $SD = 72$ minutes). During their afternoon naps, the participants kept the beepers out of hearing distance to avoid being disturbed in their sleep. Of the maximally possible 30 forms, between 4 and 40 were handed in, with an average of 26.2 ($SD = 5.3$), i.e. 87%. Independent of their response rates or response latencies, all participants were included in the analyses.

In order to minimize noise in the comparisons, coding and entry of the data should be as reliable as possible. The activities collected with both methods were coded into 59 activity categories and 29 subcategories by different coders yielding Kappas between 0.65 and 1.00 with a median of 0.76, i.e. intercoder agreements between 'substantial' and 'almost perfect' according to Landis and Koch (1977). These categories were condensed into three types of activities (see below). The data of the six ESM days were entered twice. The error rate was found to be below 0.5%. As a prerequisite of the comparison, the time windows had to be equivalent for the two methods. Therefore, we matched the YI data to the ESM window for every respondent. Activities that were reported to take place before the start or after the end of the signalling window were excluded from the analyses. A further check considered the distributions of missing data. In the case of ESM, these were sampling times at which a participant did not fill in a sampling form, and in the case of YI, these were time periods for which no activity could be reconstructed. There were no relationships between the number of missing data and age (0.07 and 0.17 for YI and ESM, respectively) or a mental-status dummy (-0.12 and -0.03 for YI and ESM, respectively). The distribution of missings is displayed for both YI and ESM in Figure 1 as a function of time of day. The largest differences arose in the early afternoon hours when a larger number of beeps were missed – probably due to the fact that participants made use of their option to keep the beeper out of reach during their naps.

Measures

Time-use indices

In order to obtain reliable and approximately normally distributed indices of activity participation we aggregated the 59 activity categories into three activity types: self-care; instrumental activities such as shopping and work – performed predominantly because of their outcomes – and autotelic ones such as reading or watching TV – performed predominantly for their own sake. Moreover, the fractions of the day spent in different locations and with differing company were determined.

Activity participation was computed in two different ways from the YI reports: (a) relative duration: time spent in an activity category relative to the length of the

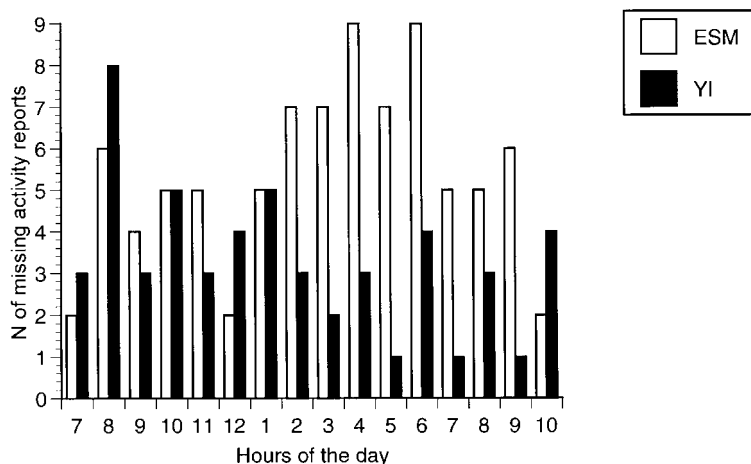


Figure 1. Number of missing activities for Yesterday Interview (YI) and Experience-Sampling Method (ESM) as a function of the hour of the day.

time window and (b) relative frequency: frequency of a category relative to the total number of activities reported in the time window. For ESM, the aggregations had a different logic to them. Since time sampling was based on a complete random procedure every activity had the same probability of being sampled (except daytime naps). This enabled us to compute time budgets in which the relative frequency of an activity can be treated as the portion of the day that activity takes up, i.e. its duration.¹

Affect measures

The positive ESM adjectives interested, active, happy, and relaxed were combined to form the unit-weighted composite 'positive affect' (Alpha on the basis of intra-person averages of the items was 0.90). The negative adjectives depressed, irritable, bored, and lonely were combined to the unit-weighted composite 'negative affect' (Alpha on the basis of intra-person averages of the items was 0.80). Likewise, the retrospectively rated positive adjectives happy and relaxed built the scale 'YI positive affect' (Alpha = 0.74), and the negative adjectives bored and lonely built the scale 'YI negative affect' (Alpha = 0.80).

Retention interval

We operationalized the length of the retention interval via the hour of the day in which the interview took place. The majority of the interviews (two-thirds) were carried out in the morning. The average interval duration was fourteen and a half hours ($SD = 3$ hours).

Congruence with pre-existing knowledge

This variable was represented by a dummy variable that took on the value 0 if the reconstructed day was judged to be typical by the participant, and the value 1 if it was

¹Time-use indices being proportion data, we applied arcsine (or angular) transformations (e.g. Cohen and Cohen, 1983) prior to the analyses.

judged atypical. As one would expect from the nature of the concept, only one-quarter of the participants considered the day they had reconstructed to be atypical.

Environmental context

The effect of contexts of encoding and retrieval were approximated via the time a person spent at home, i.e. in the location congruent with that of retrieval.

Cognitive functioning

As an indicator of cognitive functioning, we employed the Digit-Letter-Test score. Stimulus presentation and data collection was supported by a Macintosh SE30 personal computer equipped with a Micro Touch Systems touch-sensitive screen (for a detailed description of the procedure see Lindenberger *et al.*, 1993). Furthermore, the Short Mini Mental State Examination (Folstein, Folstein and McHugh, 1975) was employed.

RESULTS

Comparing time budgets based on YI and ESM data

The comparison between YI-based time-use parameters and ESM-based parameters is displayed in Table 1 for the three activity types and the context variables. For all the domains of comparison, at least one of the YI indicators (duration- or frequency-based) yielded satisfactory measures of equivalence and association. For self-care

Table 1. Time budgets from Yesterday Interview (YI) and Experience-Sampling Method (ESM) and Pearson Correlations

	YI <i>M</i> ± <i>SE</i>	ESM <i>M</i> ± <i>SE</i>	Pearson <i>r</i>	<i>N</i>
Self-care				
Frequency	24% ± 1%	22% ± 2%	0.23	74
Duration	13% ± 1%		0.23	74
Instrumental				
Frequency	36% ± 2%	29% ± 3%	0.36	74
Duration	26% ± 2%		0.39	74
Autotelic				
Frequency	32% ± 2%	44% ± 3%	0.31	74
Duration	50% ± 2%		0.47	74
Alone				
Frequency	59% ± 4%	63% ± 5%	0.66	74
Duration	57% ± 4%		0.67	74
At home				
Frequency	77% ± 2%	76% ± 3%	0.49	74
Duration	80% ± 2%		0.56	74
Positive mood				
Mean	2.69 ± 0.10	2.76 ± 0.1	0.47	65
Negative mood				
Mean	0.28 ± 0.06	0.33 ± 0.05	0.33	65

Note. Relative frequencies and relative durations from the interview are compared with relative frequencies from time sampling on the same day.

activities, relative frequencies seemed to be adequate representations in the YI context, whereas for instrumental and autotelic activities—characterized by longer durations—relative durations seem more appropriate. We will elaborate on this difference below, in the section on method effects. In the following section, along with testing statistical significance and effect sizes of the correspondence between the methods, the sources of discrepancies are analysed.

Sources of discrepancies

Hierarchical multiple regression analyses were employed to test the amount of variance in YI parameters accounted for by ES parameters and the amount of variance in the regressed difference (residual variance) accounted for by the selected covariates. In order to reduce the number of significance tests, we introduced the covariates as sets (Cohen and Cohen, 1983). Relative durations were employed for the YI-indicators, the results are displayed in Table 2.

Up to 45% of the variance in the retrospective measures could be accounted for by the *in-situ* indices. With one exception (self-care activities), time-sampling indices reliably explained at least 10% of the variance in the retrospective measures. For autotelic activities as well as time spent at home the increments in explained total variance (16.6% and 10.0%, respectively) accounted for by the set of covariates were statistically reliable and amounted to 21.1% and 14.4% of the residual variance. Inspecting the individual contributions of each covariate revealed that the effect could be attributed solely to age in both cases.

Convergent and discriminant correlations

According to Campbell and Fiske's (1959) approach, for validation purposes, it is not sufficient to demonstrate high correlation coefficients for traits that have been assessed via different methods. It is also necessary to show lower correlations

Table 2. Hierarchical multiple regressions: proportions of variance in YI-duration parameters accounted for by time-sampling parameters and a set of covariates (age, perceptual speed, length of retention interval, typicality of the reported day, and time spent at home), *F*-tests, increments, *F*-tests

YI variable	IVs	Cum. R^2	F	df	I	F_1	df
Self-care	ES self-care	0.056	4.17*	1,70	0.056	4.17*	1,70
	+Set A	0.082	0.97	6,65	0.026	0.37	5,65
Instrumental	ES instrumental	0.151	12.45***	1,70	0.151	12.45***	1,70
	+Set A	0.212	2.91*	6,65	0.061	1.00	5,65
Autotelic	ES autotelic	0.222	19.96***	1,70	0.222	19.96***	1,70
	+Set A	0.388	6.87***	6,65	0.166	3.53**	5,65
At home	ES at home	0.303	30.06***	1,70	0.303	30.06***	1,70
	+Set A	0.403	8.76***	6,65	0.100	2.70*	5,65
Alone	ES alone	0.447	54.94***	1,70	0.447	54.94***	1,70
	+Set A	0.499	10.47***	6,65	0.052	1.32	5,65
Positive mood	ES positive mood	0.224	18.16***	1,63	0.224	18.16***	1,63
	+Set A	0.280	3.77**	6,58	0.056	0.91	5,58
Negative mood	ES negative mood	0.105	7.19**	1,63	0.105	7.19**	1,63
	+Set A	0.162	1.81	6,58	0.057	0.76	5,58

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

between different traits measured either with the same or different methods. The multitrait-multimethod matrix of Table 3 displays these three types of correlations. Seven indices stemming from the YI were related to the corresponding indices based on ESM in the validity diagonal: fraction of self-care activities, fraction of instrumental activities, fraction of autotelic activities, mean positive mood, mean negative mood, fraction alone, and fraction at home. The values of the coefficients ranged from 0.24 to 0.67, their mean being 0.46.² The mono-method block on the upper left displays the correlations among the seven indices resulting from YI. The values ranged from -0.59 to 0.26 with a mean r of -0.08 . The coefficients of the mono-ESM block on the lower right were between -0.65 and 0.25 with a mean r of -0.08 . As required, the hetero-trait-hetero-method block on the lower left showed the lowest coefficients: ranging from -0.37 to 0.31 , they had a mean r of -0.04 . The negativity of the coefficients in the mono-method blocks reflects a property of compositional or proportion data, that is, the sum constraint. Two shares that add up to one will always be negatively correlated (negative bias, Aitchison, 1986).

Method effects

Time-sampling data are not a 100% reliable criterion—their assessment may also be bound up with some error. Depending on the sampling schedule and the temporal characteristics of the relevant events, time sampling may under- or overestimate the prevalence of an activity (Mann *et al.*, 1991). Particularly short and rare events are liable to being misrepresented. For this reason, we adopted the criterion from Mann and colleagues (1991) and examined whether group differences established by one method could be replicated by the other. Table 4 shows the results with regard to gender differences, age differences, and differences between participants differing in Short Mini Mental State scores.

In 14 out of 21 comparisons both methods led to equivalent decisions but only two of the significant differences converged. Gender-related differences in time spent alone were detected by both methods: women spent more time alone than men. A difference between the gender groups in instrumental activities based on the ESM data had no correspondence in the YI data. Differences between young-old and old-old participants in positive mood were detected by both methods: young-old participants reported a higher intensity of positive mood than old-old ones. The age differences in autotelic activities (YI) as well as in time spent at home (YI) had no equivalent in the alternative method. While the YI data showed unmatched differences in instrumental activities and time spent at home, there was no correspondence in the YI data with regard to the difference between the mental-status groups in autotelic activities and positive mood that was present in the ESM data. The directions of the group differences were equal for both methods in all those group comparisons in which a significant difference was found on the basis of one of the two methods.

DISCUSSION

Bearing in mind the low number of daily time samples, the comparisons between time budgets computed with ESM and YI data revealed considerable agreement

²Correlations were averaged after being transformed into Fisher's z s and the means were retransformed into r s.

Table 3. Correlations between indicators of self care (-CARE), instrumental and autotelic activities (-INST, -AUTO), mean positive mood (-PS), mean negative mood (-NG), time alone (-ALO) and time spent at home (-IN) assessed via Yesterday-Interview (YI, duration-based) and Experience-Sampling Method (ES)

	YI-CARE	YI-INST	YI-AUTO	YI-PS	YI-NG	YI-ALO	YI-IN	ES-CARE	ES-INST	ES-AUTO	ES-PS	ES-NG	ES-ALO
YI-INST	-0.20												
YI-AUTO	-0.22	-0.59***											
YI-PS	-0.08	0.00	0.24*										
YI-NG	0.18	0.26*	-0.31**	-0.39**									
YI-ALO	-0.30**	0.25*	-0.13	-0.08	0.12								
YI-IN	0.11	-0.33**	-0.14	-0.18	0.04	0.17							
ES-CARE	0.24*	-0.09	-0.04	-0.01	-0.02	-0.22	-0.17						
ES-INST	-0.21	0.39**	-0.30*	-0.18	0.04	0.24*	-0.26*	-0.26*					
ES-AUTO	-0.04	-0.37**	0.45***	0.08	-0.07	-0.05	0.31**	-0.41***	-0.65***				
ES-PS	-0.02	0.14	0.26*	0.47***	-0.27*	-0.08	-0.20	0.12	-0.12	0.05			
ES-NG	0.04	-0.00	-0.01	-0.32*	0.33**	0.04	0.23	-0.01	-0.13	0.07	-0.07		
ES-ALO	-0.20	0.15	-0.02	-0.16	0.18	0.67***	0.12	-0.08	0.25*	-0.05	-0.03	-0.04	
ES-IN	0.22	-0.25*	0.04	-0.03	0.00	0.06	0.56***	0.09	-0.33**	0.02	-0.20	-0.03	-0.04

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4. Significant mean differences between women and men, young-old and old-old, participants with Short Mini Mental State score (SMMS) greater equal and less than 14, Yesterday-Interview (YI) versus Time-sampling data

	Gender		Age		SMMS	
	YI	ESM	YI	ESM	YI	ESM
Self-care						
Instrumental		*			**	
Autotelic			**			*
Alone	**	*				
At home			**		**	
Positive mood			**	*		*
Negative mood						

* $p < 0.05$; ** $p < 0.01$.

between the two methods. When looking at the different ways of computing YI indices, it seems obvious that self-care activities were better represented with relative frequencies while instrumental and autotelic activities – both characterized by longer average durations – were better represented by relative durations. Retrospective ratings of general mood level corresponded well with indices based on average momentary mood. This accuracy of memory for affect intensity is consistent with evidence reported by Parkinson and colleagues (1995). There was no indication, however, of the positivity bias found by these authors. Rather, independent of their valence, both scales based on *in-situ* ratings tended to have higher means than their retrospective counterparts.

There were, however, also deviations between YI- and ESM-based measures. Can these discrepancies be attributed to memory effects? Before summarizing and discussing the results, one caveat has to be mentioned: in our naturalistic approach, there was no random assignment of participants to conditions. Therefore, a *ceteris paribus* assumption does not apply: differences between variables of interest tend to be confounded with other properties of the person and the situation (e.g. typicality of a day, performed activities, and age). Larson and Delespaul (1992) have pointed out the problem that aggregates over specific naturally occurring situations may be different from aggregates over experimentally controlled situations and, therefore, cannot be interpreted in the same fashion. They recommend looking at the data from several different angles – a strategy we tried to adopt.

Hardly any of the expected biases were found: neither time at home (indicating congruence of the locations of encoding and retrieval), the length of the retention interval, nor congruence with pre-existing knowledge accounted reliably for the variance in the residuals once the variance explained by the ES measures was removed. This can be attributed to the restricted variance in and non-random nature of the covariates. The majority of the Yesterday Interviews took place in the morning, the majority of the days were considered typical, and the largest proportion of the day was spent at home. Furthermore, in the literature, encoding-specificity effects are often small for older adults since they tend to encode fewer contextual details (e.g. Burke and Light, 1981; but see also: Earles *et al.*, 1996). Extrapolating to the area of memory research of which the reconstruction of the previous day is a special case, i.e. autobiographical memory (e.g. Rubin, 1986), larger effects can be expected since,

with regard to the length of the retention intervals and the variety of contexts, 24-hour recall is located at the low end of the dimensions.

Cognitive functioning was not related to accuracy of recall and age effects were found only for the residuals of autotelic activities and time at home. This is not consistent with existing evidence based on subjective as well as objective indicators of memory functioning (e.g. Craik and McDowd, 1987)—it may have resulted from restrictions of range in age and cognitive functioning. Moreover, since actions are automatically encoded during their performance, Kausler (1994) suggested smaller age differences in memory for activities as opposed to verbal material to be explained by the independence of actions from rehearsal. On the other hand, the automaticity of encoding may make monitoring and recalling motor output more difficult (e.g. Bell *et al.*, 1997; Koriat and Ben-Zur, 1988).

The two methods' sensitivities were similar. But whereas about the same numbers of significant differences were obtained on the basis of the ESM and YI, only two of them converged. This level of agreement can hardly be called sufficient and may be attributed to specific characteristics of the two methods. First, because of the smaller number of activity samples of ESM, the standard errors of the respective indicators were larger and hence their reliabilities lower. Second, the methods might be differentially reliable in different contexts. Time outside the home, for instance, might be underestimated on the basis of ESM because higher noise levels outside compared to inside the home could have led to a greater number of beeps being missed.

To summarize, convergent and discriminant correlation patterns from indices of activity participation, mood, location, and company collected via retrospective (YI) and *in-situ* (ESM) self-reports indicated acceptable correspondence between the two methods. Discrepancies between YI and ESM could not be explained by memory biases or interindividual differences in cognitive functioning. For two indicators, however, there was an age effect. The sensitivity of the two methods to detect differences between groups was found to be largely equivalent.

ACKNOWLEDGEMENTS

The research reported here was financially supported by the German Research Foundation (DFG) grant Ba 902/8-1 to Margret M. Baltes. We would like to thank Heiner Maier, Werner Wittmann, and two anonymous reviewers for their helpful comments on an earlier version of this article. Also, thanks are due to Andrea Nies, Antje Hänsch, Astrid Meyer, Barbara Wörmann, Christian Geng, Jens Bisanz, and Martina Junker for assistance in data collection. Franziska Perels and Sigrun Würfel helped with data coding. Moreover, we have to thank our present and past BASE colleagues for their cooperation, and, last but not least, many thanks to our participants for contributing time and energy to this demanding study.

REFERENCES

- Ach, N. (1935). Analyse des Willens [Analysis of the will]. In E. Abderhalden (Ed.), *Handbuch der biologischen Arbeitsmethoden* (Band VI). Berlin: Urban and Schwarzenberg.
- Aitchison, J. (1986). *The statistical analysis of composition data*. London: Chapman and Hall.

- Bell, B. G., Gardner, M. K. and Woltz, D. J. (1997). Individual differences in undetected errors in skilled cognitive performance. *Learning and Individual Differences*, **9**, 43–61.
- Brewer, W. F. (1988). A qualitative analysis of the recalls of randomly sampled autobiographical events. In M. M. Gruneberg, P. E. Morris and R. N. Sykes (Eds.), *Practical aspects of memory: Current Research and Issues* (Vol. 1, pp. 263–268). Chichester: Wiley.
- Burke, D. M. and Light, L. L. (1981). Memory and aging: The role of retrieval processes. *Psychological Bulletin*, **90**, 513–546.
- Burt, C. D. B. and Kemp, S. (1994). Construction of activity duration and time management potential. *Applied Cognitive Psychology*, **8**, 155–168.
- Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, **56**, 81–105.
- Cantor, N., Norem, J., Langston, C., Zirkel, S., Fleeson, W. and Cook-Flannagan, C. (1991). Life tasks and daily life experience. *Journal of Personality*, **59**, 425–451.
- Carstensen, L. L., Pasupathi, M. and Mayr, U. (in preparation). Emotion and everyday life in a lifespan sample.
- Cohen, J. and Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Craik, F. I. M. and McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**, 474–479.
- Csikszentmihalyi, M. and Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease*, **175**, 526–536.
- Curie, J., Hajjar, V., Marqui, H. and Roques, M. (1990). Proposition méthodologique pour la description du système des activités [Methodological recommendation regarding the description of the activity classification]. *Travail Humain*, **53**, 103–118.
- Delespaul, P. A. E. G. (1995). *Assessing schizophrenia in daily life. The experience sampling method*. Maastricht: UPM.
- Earles, J. L., Smith, A. D. and Park, D. C. (1996). Adult age differences in the effects of environmental context on memory performance. *Experimental Aging Research*, **22**, 267–280.
- Eich, E. and Metcalfe, J. (1989). Mood-dependent memory for internal vs. external events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 443–455.
- Engle, P. L. and Lumpkin, J. B. (1992). How accurate are time-use reports? Effects of cognitive enhancement and cultural differences on recall accuracy? *Applied Cognitive Psychology*, **6**, 141–159.
- Fahrenberg, J. (1994). Ambulantes assessment [Ambulatory assessment]. *Diagnostica*, **40**, 195–216.
- Folstein, M. F., Folstein, S. E. and McHugh, P. R. (1975). 'Mini Mental State': A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, **12**, 189–198.
- Juster, F. T. (1986). Response errors in the measurement of time use. *Journal of the American Statistical Association*, **81**, 390–402.
- Kausler, D. H. (1994). *Learning and memory in normal aging*. San Diego, CA: Academic Press.
- Klinger, E. (1987). Current concerns and disengagement from incentives. In F. Halisch and J. Kuhl (Eds.), *Motivation, intention, and volition* (pp. 337–347). Berlin: Springer.
- Klumb, P. L. and Baltes, M. M. (1995). Berlin Aging Study: Selective participation and attrition in a study on everyday competence in old age based on the experience-sampling method. Paper presented at the EURODEP Conference, Dublin, Ireland, 17–19 November.
- Knäuper, B. and Wittchen, H. U. (1994). Diagnostic major depression in the elderly: Evidence for response bias in standardized diagnostic interviews. *Journal of Psychiatric Research*, **28**, 147–164.
- Koriat, A. and Ben-Zur, H. (1988). Remembering that I did it: processes and deficits in output monitoring. In M. M. Gruneberg, P. E. Morris and R. N. Sykes (Eds.), *Practical aspects of memory: Current Research and Issues* (Vol. 1, pp. 203–208). Chichester: Wiley.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.
- Lang, F. and Helle, G. (1994). RC-GEN: Random call generation software.

- Larson, R. and Delespaul, P. A. E. G. (1992). Analyzing Experience Sampling data: A guidebook for the perplexed. In M. W. deVries (Ed.), *The experience of psychopathology: investigating mental disorders in their natural settings* (pp. 58–78). New York: Cambridge University Press.
- Lindenberger, U., Mayr, U. and Kliegl, R. (1993). Speed and intelligence in old age. *Psychology and Aging*, **8**, 207–220.
- Loftus, E. F. and Marburger, W. (1983). Since the eruption of Mount St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*, **11**, 114–120.
- Mann, J., Ten Have, T., Plunkett, J. W. and Meisels, S. J. (1991). Time sampling: A methodological critique. *Child Development*, **62**, 227–241.
- Moss, M. S. and Lawton, M. P. (1982). Time budgets of older people: A window on four lifestyles. *Journal of Gerontology*, **37**, 115–123.
- Niemi, I. (1993). Systematic error in behavioral measurement: Comparing results from interview and time budget studies. *Social Indicators Research*, **30**, 229–244.
- Parkinson, B., Briner, R. B., Reynolds, S. and Totterdell, P. (1995). Time frames for mood: Relations between momentary and generalized ratings of affect. *Personality and Social Psychology Bulletin*, **21**, 331–339.
- Pawlik, K. and Buse, L. (1982). Rechnergestützte Verhaltensregistrierung im Feld: Beschreibung und erste psychometrische Überprüfung einer neuen Erhebungsmethode [Computer-based behaviour recording *in situ*: Description and first psychometric test of a new assessment method]. *Zeitschrift für differentielle und Diagnostische Psychologie*, **3**, 101–118.
- Robinson, J. P. (1985). The validity and reliability of diaries versus alternative time use measures. In F. T. Juster and F. P. Stafford (Eds.), *Time, goods, and well-being* (pp. 33–62). Ann Arbor: Institute for Social Research.
- Robinson, J. P. and Nicosia, F. M. (1991). Of time, activity, and consumer behavior: An essay on findings, interpretations, and needed research. *Journal of Business Research*, **22**, 171–186.
- Rubin, D. C. (1986). *Autobiographical memory*. Cambridge: Cambridge University Press.
- Rubin, D. C. and Baddeley, A. D. (1989). Telescoping is not time compression: A model of the dating of autobiographical events. *Memory and Cognition*, **17**, 653–661.
- Scheuch, E. K. (1972). The time-budget interview. In A. Szalai (Ed.), *The use of time* (pp. 69–87). The Hague: Mouton.
- Smith, S. (1988). Environmental context-dependent memory. In G. M. Davies and D. M. Thomson (Eds.), *Memory in context: Context in memory* (pp. 13–34). London: Wiley.
- Totterdell, P. and Folkard, S. (1992). In situ repeated measures of affect and cognitive performance facilitated by use of a hand-held computer. *Behavior Research Methods, Instruments, and Computers*, **24**, 545–553.
- Tulving, E. (1974). Cue-dependent forgetting. *American Scientist*, **62**, 74–82.
- Watson, D., Clark, L. A. and Tellegen, A. (1988). A development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology*, **54**, 1063–1070.
- Wheeler, L. and Reis, H. T. (1991). Self-recording of everyday life events: Origins, types and issues. *Journal of Personality*, **59**, 339–354.