Retrospective Questions or a Diary Method? A Two-Level Multitrait–Multimethod Analysis

Joop J. Hox and Annet M. Kleiboer
Faculty of Social Sciences, Utrecht University, The Netherlands

This study describes a comparison between retrospective questions and daily diaries inquiring about positive and negative support in spousal interactions. The design was a multitrait–multimethod matrix with trait factors of positive and negative support, and method factors of retrospective questions and daily asked questions. Five questions were used for positive support and 2 questions were used for negative support. The data were multilevel, with daily measurement occasions nested within subjects. In addition, the data were ordered categorically. The negative support events proved to be so rare that the original 4-point response scale had to be dichotomized. The resulting model could be estimated using M*plus*, but the model and data complexities set some limits to the analysis. The results showed that at the subject level both positive and negative support could be assessed with sufficient reliability and validity. At the daily measurements level, positive support showed significant but low reliability and validity, but negative support could not be assessed reliably. It was concluded that at the daily level both positive and negative support should be viewed as transient events that do not indicate an underlying latent variable, but that could be modeled as a formative construct.

When participants are asked about events occurring in a certain period, two approaches can be followed: asking general questions that require that respondents generate their own general appraisal, or requesting participants to fill in a diary during a specific period. Although the actual questions may be very similar, asking questions that assess the current day or asking retrospectively about a general appraisal is connected with a different underlying cognitive process (Schwarz, 1999). Both approaches have their own strengths and weaknesses. Asking general ques-

Correspondence should be addressed to Joop J. Hox, Faculty of Social Sciences, Utrecht University, POB 80140, NL–3508 TC Utrecht, the Netherlands. E-mail: j.hox@fss.uu.nl

tions is efficient, but must assume that individuals can correctly recall and summarize the relevant events in a general way. On the other hand, using a diary method is more costly both for the researcher (time and money) and for the respondent (response burden). If the object of the study is to investigate relations between individual characteristics, the diary data may be aggregated to the individual level, and produce individual variables comparable to variables based on questionnaire data. If the object of the study is to analyze a process daily over time, a diary method is the only realistic alternative, and the relations studied are the relations between values at the different time points. This assumes that the variation observed across the different time points is reliable and valid, and not merely random fluctuations around an individual average.

Given the considerable burden to respondents and the added costs for researchers, it is important to assess whether the diary method adds reliable and valid information to the data that can be obtained by the straightforward method of a single questionnaire. This study investigated the reliability and construct validity of measures of social support, both at the level of occasions within individual participants, and at the between-participant level. It compared retrospective data collected by a questionnaire on one single occasion with diary data collected daily over a period of 2 weeks. This design can be viewed as a multitrait–multimethod (MTMM) design, with two traits being positive and negative support, and two methods being the retrospective questionnaire and the daily asked questions. Because the diary data were collected on 14 consecutive days, the data can also be viewed as having a multilevel structure, with up to 14 measurement occasions nested within individuals. Finally, the data were nonnormal; the frequency of both positive and negative support events was asked using a four-category answer scale, and especially for negative support events, the response distribution was skewed.

The research questions were as follows:

1. What is the amount of systematic (reliable) variance for positive and negative support at the between-subject level, and how do questionnaire data compare with aggregated diary data?
2. What is the amount of systematic (reliable) variance for positive and negative support at the within-subjects level?
3. What is the convergent and discriminant validity of positive and negative support at both the between-subject and the occasion within-subjects level?

To answer these questions, two types of models were used. The amount of participant-level variance of the positive and negative measures and their reliability were assessed using a multilevel congeneric test model, and the convergent and discriminant validity were assessed using confirmatory factor analysis (CFA). The next section describes the available data and the models.

METHOD

Participants and Measurements

Participants

Two-hundred seventeen individuals participated in the study. Data were collected as part of a study examining spousal support in couples dealing with multiple sclerosis (MS), a chronic progressive disorder, and healthy couples. Therefore, the participants were all members of a couple dealing with MS or a healthy couple. The study included 61 couples dealing with MS (61 patients and 59 partners) and 50 healthy couples (97 individuals). MS couples were recruited through the patient files of one MS center and the neurology department of one hospital in the Netherlands. Healthy couples were a convenience sample. Inclusion criteria were (a) reading and writing the Dutch language, (b) being a couple living together and having a heterosexual relationship for at least 1 year, and (c) being at least 18 years of age. Of the couples with MS, one partner had to be diagnosed with MS. Fifty percent (n = 109) of the participants were male and 50% (n = 108) were female. Participants' average age was 44.4 years (SD = 11.1), 19% of the partners had completed the lowest level of secondary (vocational) education only, 40% of the participants had completed middle to higher levels of secondary (vocational) education, and 41% of the participants had a college degree or higher. The couples had been married (79%) or cohabiting (21%) for an average of 19.7 years (SD = 11.1).

Procedure

The self-report questionnaire was completed prior to the diary part of the study at home. Next, 1 to 4 weeks after the self-report questionnaire had been returned, the couples were visited at home to provide additional information and instructions about the diary part of the study and to install an electronic diary on the couple's computer. Participants who did not own a computer were provided with a computer from the university. Computer software was developed especially for this study. The electronic diary was user friendly and easy to complete. Even participants who had little or no experience with computers were able to use the program after they were given clear instructions. Participants were instructed to complete the electronic diary every evening before going to bed for 2 weeks, starting the following day. Participants were explicitly asked not to exchange the answers with their partner while completing the diaries. The diary was designed to be completed in 5 to 10 min. Recordings were saved on a floppy disk each night and the participants did not have access to their reports after they were saved on the floppy disk. After 2 weeks, the couples returned the floppy disk containing each night's responses by mail.

To verify compliance, the date and time of recordings were saved on the floppy disk. Participants were allowed to fill out the diary the next morning if they did not manage to do it at night. In total, 9.6% of the diaries were completed the next morning. If the diaries were completed too late (after 2 p.m. on the next day) or too early (before 2 p.m.), they were excluded from further analyses because they were considered unreliable. This was the case for 2% of the recordings. Across the 14-day period, participants reported an average of 12.9 days of recordings.

Measures

Retrospective assessment of support was assessed with 7 items: Five items assessed positive support and 2 items assessed negative support. The items were selected from a larger scale for spousal positive support and spousal negative support, respectively, the 34-item Social Support List–Interactions and the 7-item Social Support List–Negative Interactions (SSL–I and SSL–N; van Sonderen, 1991). Participants were asked to complete a larger questionnaire; however, in this analysis only these 7 items were used. Items for positive support were: Does it ever happen to you that your partner pays you a compliment? Provides you with help in practical everyday things such as household chores, odd jobs? Lends you a friendly ear? Gives you good advice? Is affectionate toward you? Items for negative support were: Does it ever happen to you that your partner makes disapproving remarks to you? Makes unreasonable demands of you? Participants were asked to indicate the amount of positive or negative support they received from their spouse in general on a 4-point scale ranging from 0 (*never or seldom*) to 3 (*very often*).

Daily measures of support were assessed with 7 items similar to the questionnaire items: 5 items that assessed positive support and 2 items that assessed negative support. Each evening, both patients and partners reported if and to what extent they had received support on that day. Items for positive support included: Did it happen today that your partner paid you a compliment? Listened to you? Was affectionate toward you? Offered you practical help? Gave you information or advice? Items for negative support included: Did it happen today that your partner made disapproving remarks of you? Demanded a lot of you? All answers were given on a 4-point scale ranging from 0 (*not at all*) to 3 (*very much*).

ANALYSIS MODELS

With diary data, two separate levels can be distinguished: the within-subjects or occasions level, and the between-subject level. The daily questions produce time-varying within-subjects data, but these can also be aggregated to the subject level. This creates subject-level data based on the aggregated within-subjects measures. In addition there are questionnaire data. The questionnaire inquires about the general frequency over time and therefore produces only subject-level information; basically it relies on the participants' ability to summarize the events. In the within-subjects data there are five positive and two negative items asked of 217 participants at up to 14 occasions. The questionnaire adds at the between-subject level five positive and two negative equivalent general items, asked only once. In this case, the interest is in characteristics observed at both levels, occasions within-subjects level referring to time-varying events, and the between-subject level referring to time-invariant or aggregated time-varying variables across all separate occasions.

The multilevel structural equation modeling (SEM) analyses used M*plus* 4 (Muthén & Muthén, 2006). In the initial analysis, it was found that the extreme skewness of the negative support items produced severe numerical problems. Therefore, the decision was made to dichotomize all negative support items. This leads to a small loss of information; the substantive implications of this decision are taken up later in the discussion.

In addition to the nesting of occasions within participants, there were participants nested in dyads (couples). This can be incorporated by adding one more level to the model for couples. Because the groups were dyads and therefore small, this would imply strong restrictions on the size and complexity of the model (cf. Newsom, 2002). However, ignoring this source of nonindependence is known to lead to biased standard errors (Hox, 2002; Kenny, 1995). The important issue in this analysis was to assess the sizes of reliability and validity coefficients, rather than their significance. Therefore, the dyad level was disregarded, and robust standard errors and chi-squares were relied on to provide corrected significance tests (Muthén & Muthén, 2006).

The Multilevel Model for Reliability

The first two research questions were (a) What is the amount of systematic (reliable) variance for positive and negative support at the between-subject level, and how do questionnaire data compare with aggregated diary data? (b) What is the amount of systematic (reliable) variance for positive and negative support at the occasions within-subjects level. These questions could be addressed by estimating the internal consistency reliability of these scales. However, simply calculating the customary Cronbach's alpha would lead to a reliability estimate that is difficult to interpret, because it would be based on a mixture of occasion-level and individual subject-level variance. From the measurement point of view, the total variance needs to be decomposed into an error variance and the systematic variance at both available levels.

Estimating reliability in multilevel data is a standard multilevel regression procedure. The approach is to add an extra level for the variables, which in this case leads to three separate levels: the items, the occasions, and the participants. Next,

variance components were estimated for each level, and these were used to calculate separate reliability coefficients at the occasion and the participant level. Details on this model were given by Raudenbush, Rowan, and Kang (1991) and Raudenbush and Sampson (1999); for an introduction see Hox (2002). However, this approach is quite restricted, because it implies equal loadings and equal error variances; that is, parallel measures (Lord & Novick, 1968). Multilevel reliability analysis can also be carried out using SEM (cf. Raykov & Marcoulides, 2006; Raykov & Shrout, 2002). The basic model is a congeneric test model, which is a model where all items load on a single factor, but both loadings and error variances are free to vary. Given the larger flexibility of the congeneric test model, SEM is used here.

The congeneric test model is given by (cf. Raykov & Marcoulides, 2006, pp. 131–132)

$$Y_i = a_i + b_i\eta + \varepsilon_i \tag{1}$$

with subscript i for items, a_i is the intercept, b_i the loading, and ε_i the residual measurement error. Following the classical test theory of reliability, the reliability ρ_y of the sum score is defined as the proportion of true score variance, which can be expressed as

$$\rho_Y = \frac{\left(\sum b_i\right)^2 \mathrm{Var}(\eta)}{\left(\sum b_i\right)^2 \mathrm{Var}(\eta) + \sum \theta_{ii}} \tag{2}$$

where θ_{ii} are the measurement error variances. Because this was a simple unconditional measurement model, the model in Equation 1 could be specified separately for each set of items; for the diary data once on the occasion and once on the subject level of a two-level model.

Because these data were ordered categories, the appropriate analysis model was to use a generalized linear model with a logit link function. For the reliability analysis, this was problematic, because categorical two-level models do not estimate the measurement variance. Using the more restricted logistic multilevel regression model proposed by Raudenbush et al. (1991), which does allow estimation of item-level variance, was not an option, because the estimation methods currently implemented in multilevel software are known to be inaccurate when the lowest level sample is small and the intraclass correlation (ICC) is high—a situation that exists in these data. Only the reliabilities produced by treating the variables as continuous using a congeneric test model in SEM are presented, as an approximate indication of the amount of systematic variance. This was double-checked by estimating a categorical model using multilevel regression software. This leads to different reliabilities, but not to different substantive conclusions.

The Structural MTMM Model for Validity

The third research question was what the convergent and discriminant validity of positive and negative support is at both the between- and the within-subjects level. This required that a model be specified at the occasion and the participant level that reflected the features of the MTMM design. In an MTMM design several distinct characteristics or *traits* are measured using several distinct measurement *methods*. Traits can be "attributes such as multiple abilities, attitudes, behaviors, or personality characteristics" and methods "refer broadly to multiple test forms, methods of assessment, raters, or occasions" (Marsh & Grayson, 1995, p. 177). The two traits in this study were unsupportive and supportive behavior, and the two methods were the questionnaire and the diary method.

MTMM designs are used to evaluate construct validity, which requires both high convergent and discriminant validity. Convergent validity implies a high overlap between alternative measures that refer to the same construct (Hoyle, Harris, & Judd, 2002, p. 90). Discriminant validity implies that a measure should not correlate highly with other measures that refer to different constructs (Hoyle et al., 2002, p. 92). In addition, convergent validity requires small method effects (Marsh & Grayson, 1995).

CFA is commonly used to decompose the underlying factors in MTMM. Trait factors are defined by different measures of the same trait, and method factors are defined by different constructs assessed with the same method. The basic assumptions are that each measure loads on only one trait and one method factor, and that the covariances between trait and method factors are zero. Large factor loadings for the traits indicate high convergent validity, large method factor loadings imply measurement bias due to method effects, and high correlations between trait factors indicate a lack of discriminant validity.

In an MTMM design, usually each combination of trait and method is represented by only one single measure. In this design, there were several measures for each combination of trait and method: There were five measures for positive support and two measures for negative support, once in the questionnaire method and once in the diary method, for a total of 14 measures. These 14 measures indicated the trait factors of positive support and negative support, and the method factors of questionnaire and diary method. A specific feature following from the multilevel nature of these data (daily diaries nested within participants) was that on the between-subject level all 14 measures were present (both questionnaire and aggregated diary data), whereas on the within-subjects level, only the diary variables were present, as the once-only questionnaire measures were time-invariant at the within-subjects level. The path diagram for the multilevel MTMM model is given in Figure 1.

In multilevel SEM, each variable Y_{hij} (the response of participant j on occasion i to question h) was decomposed into a within-persons component W_{hij} where

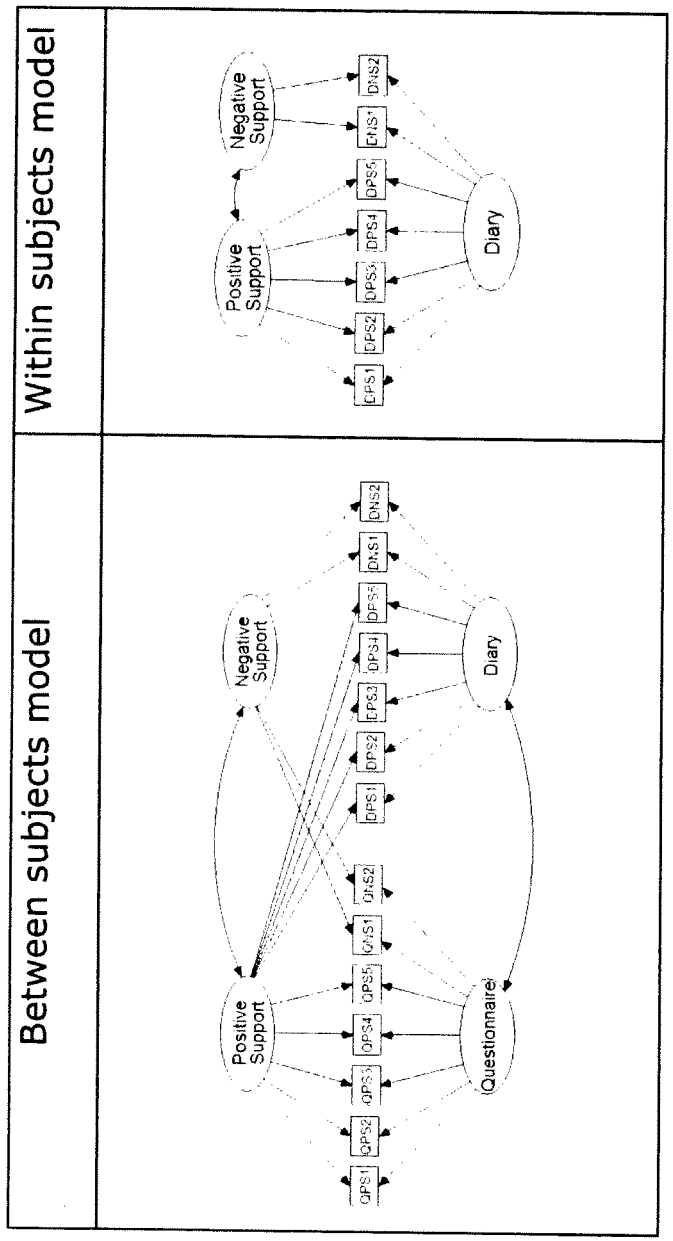

FIGURE 1 MTMM model for between- and within-subjects data.

$W_{hij} = Y_{hij} - \bar{Y}_{hj}$ and a between-person component B_{hj} where $B_{hj} = \bar{Y}_{hj}$. On the within level the model equation reads

$$W_{hij} = \lambda_{hp} P_{ij} + \lambda_{hn} N_{ij} + \lambda_{hd} D_{ij} + e_{ij} \tag{3}$$

where the subscripts ij denote occasion i for individual j, λ_{hp} is the loading of item h on the positive support factor, λ_{hn} is the loading of item h on the negative support factor, and λ_{hd} is the loading of item h on the diary method factor. On the between level the model equation reads

$$B_{hj} = \lambda_{hp} P_j + \lambda_{hn} N_j + \lambda_{hd} D_j + \lambda_{hq} Q_j + e_{ij} \tag{4}$$

where the subscript j denotes for individual j, λ_{hp} is the loading of item h on the positive support factor, λ_{hn} is the loading of item h on the negative support factor, λ_{hd} is the loading of item h on the diary method factor, and λ_{hq} is the loading of item h on the questionnaire method factor.

The data were categorical ordinal for the positive support measures and dichotomous for the negative support items. Multilevel structural models for such data can be estimated with M*plus* (Muthén & Muthén, 2006), which uses a logit link, analogous to the HLM approach. The estimation procedure is numerically demanding, with demands on computer memory and processing time increasing as a power function of the number of latent variables in the model (in this case, seven; cf. Figure 1). For this approach, M*plus* used Monte Carlo numerical integration, which does not produce overall model fit indicators, although it does produce standard errors for the parameter estimates. This problem is not linked to the specific software used. The only publicly available software besides M*plus* that can handle this kind of estimation problem is *glamm* (Skrondal & Rabe-Hesketh, 2004), which also uses numerical integration. A similar modeling procedure for multilevel ordinal data was described by Grilli and Rampichini (2005). Indirect information on the goodness of fit of the multilevel MTMM model in Figure 1 for these data is presented in the next section.

RESULTS

Data Cleaning

The nesting of occasions within participants leads to a data file with 217 participants × 14 occasions = 2,803 data points. Data from 6 participants were deleted listwise because of incomplete questionnaire data. In addition, on 83 occasions the diary data were incomplete, which means that that specific day was deleted from the data file. So, about 2% of the data points were deleted listwise assuming missing completely at random and about 3% of the data points were deleted partially

assuming missing at random (MAR). Under this assumption, this is acceptable because these are considered ignorable types of missing data (Little & Rubin, 2004). In addition, 21 data points (daily measures) were deleted because their Mahalanobis distance score showed that they were extreme multivariate outliers. The resulting data set consists of 2,629 measurement occasions for 211 participants.

In the original metric, all items were measured on a 4-point response scale. All response distributions were to some degree skewed, but the two negative support questions produced strongly skewed distributions: The skewness is larger than 30 for the diary items and larger than 10 for the corresponding questionnaire items. As reported earlier, these large skewness values caused computational problems, and it was decided to dichotomize these responses for both the questionnaire and the daily diaries.

Reliability: Systematic Variance at Occasion and Participant Level

Table 1 presents the results for both the positive and the negative items. Presented are the ICC, average of the estimated population proportion of variance at the individual level, and the reliabilities based on the congeneric test model and continuous data.

Table 1 shows that for the diary data the reliabilities at the participant level were much higher than at the occasion level. Aggregating the measurements (taking the average of the daily measurements) resulted in participant scores that were in fact more reliable than the questionnaire measurement. The reliability at the occasion level was much smaller, especially for the negative items. The difference between the positive and negative items was apparent. The reliability reflects both the effect

TABLE 1
(Average) ICC and Reliability for Positive and Negative Support Items

	Method			
	Diary		*Questionnaire*	
Level	*ICC*	*Reliability*	*ICC*	*Reliability*
Positive items (5 items)				
Occasions	N/A	.68	N/A	N/A
Participants	.46	.89	1.00	.75
Negative items (2 items)				
Occasions	N/A	.27	N/A	N/A
Participants	.22	.67	1.00	.53

Note. ICC = intraclass correlation.

of the size of the systematic variance component and the number of items averaged. However, the ICCs were also much lower for the negative items. This issue is covered further in the discussion.

The results for the questionnaire items are also presented in Table 1; note that there is no occasion level with the questionnaire data. Again, the reliability for the positive items was much higher than for the negative items. For positive support, both the questionnaire and the averaged diary method provided reliable information, but for negative support only the averaged diary method proved sufficiently reliable. This result is revisited in the discussion.

Validity: Convergent and Discriminant Validity at Occasion and Participant Level

Table 2 presents the factor loadings for the multilevel confirmatory factor model depicted in Figure 1. The item labels describe whether they indicate positive or negative support (PS or NS) and whether the method was diary or questionnaire (D or Q).

Table 2 shows that for questionnaire items the standardized trait loadings are on average a little larger than the method loadings. For the diary items there is a difference between the positive and negative items: For the negative items the diary

TABLE 2
Factor Loadings on the Between Subjects Level

	Trait			*Method*		
Items	*Factor Loading*	*SE*	*Standardized Loading*	*Factor Loading*	*SE*	*Standardized Loading*
Q PS1	1.00	—	0.67	1.00	—	0.43
Q PS2	1.20	0.24	0.71	1.21	0.40	0.46
Q PS3	0.54	0.15	0.41	1.23	0.55	0.60
Q PS4	0.93	0.19	0.66	0.86	0.27	0.39
Q PS5	0.27	0.15	0.26	0.78	0.36	0.48
Q NS1	1.00	—	0.55	−1.04	0.48	−0.55
Q NS2	1.42	0.55	0.56	−0.72	0.37	−0.33
D PS1	2.10	0.47	0.79	1.00	--	0.61
D PS2	1.89	0.46	0.99	0.07	0.13	0.07[a]
D PS3	0.26	0.29	0.15[a]	1.05	0.27	0.98
D PS4	0.83	0.23	0.45	1.02	0.19	0.89
D PS5	0.15	0.18	0.14[a]	0.66	0.12	0.99
D NS1	1.89	0.76	0.97	0.13	0.08	0.23[a]
D NS2	3.65	1.61	0.97	0.23	0.12	0.21[a]

Note. PS = positive support; NS = negative support; D = diary method; Q = questionnaire method.
[a]Not significant at $\alpha = 0.05$.

method produces trait loadings that are clearly higher than the method loadings. For the diary method, several loadings do not reach significance. At the between-subject level, the correlation between PS and NS is –.31 (p = .07). The correlation between the questionnaire and the diary method factor is .04 (p = .56). Given the relatively high method loadings, the convergent validity is only moderate. The low correlation between the two traits indicates good discriminant validity. The method effects are interpreted in the discussion section.

Table 3 presents the factor loadings at the within-subjects level. It shows that at the measurement occasion level the standardized trait method loadings tend to be larger than the method loadings, but smaller than the standardized trait loadings at the between level. The method loadings in the within model are all not significant.

Current software does not produce global model fit indexes when numerical integration is used. To obtain some insight into the fit of the MTMM model, two comparable models are considered for which global fit indexes can be calculated. First, if the categorical data is treated as continuous, a standard multilevel structural equation model can be estimated including a chi-square and other fit indexes. This approach produces a (robust) $\chi^2(74, N = 211/2{,}629) = 171.9$, $p = .00$, and fit indexes of comparative fit index (CFI) = 0.96 and root mean squared error of approximation (RMSEA) = 0.02. Second, if the data are treated as ordered categorical but single level (disaggregated), specifying the data type as complex to take account of the clustering in the data results in $\chi^2(30, N = 2{,}629) = 75.5$, $p = .00$, with CFI = 0.97 and RMSEA = 0.02. This is taken as indirect evidence that this MTMM model fits reasonably well.

The problem with obtaining global model fit indicators when the logit link function with numerical integration is used is that the saturated model cannot be estimated. M*plus* does report a value for the log-likelihood, which makes explicit

TABLE 3
Factor Loadings on the Within Subjects Level

	Trait			*Method*		
Items	*Factor Loading*	*SE*	*Standardized Loading*	*Factor Loading*	*SE*	*Standardized Loading*
D PS1	1.00	—	0.64	1.00	—	0.06
D PS2	1.04	0.26	0.59	8.08	20.89	0.45[a]
D PS3	0.80	0.40	0.35	2.18	6.06	0.77[a]
D PS4	1.04	0.23	0.65	1.27	1.59	0.08[a]
D PS5	0.55	0.16	0.40	6.72	17.10	0.45[a]
D NS1	1.00	—	0.43	2.16	6.32	0.16[a]
D NS2	0.84	0.31	0.36	3.58	10.42	0.26[a]

[a]Not significant at $\alpha = 0.05$.

Note. PS = positive support; NS = negative support; D = diary support.

testing of nested models possible. In this case, one more restricted model was estimated—following Widaman's (1985) suggestions—which was a model with trait factors but omitting the method factors. The log-likelihood for the full MTMM model was –17446.9 (79 free parameters), and for the "no methods" model it was –17919.4 (64 free parameters). The model without between-subject method factors did not converge, which was interpreted as indicating a large misfit between model and data. The log-likelihood of the model without within-subjects methods was —17483.2 (72 free parameters). The "no methods" models were all nested within the basic MTMM model, but explicit model comparison tests were not conducted because it was uncertain that the likelihood would be accurate enough to permit such testing. However, consistent with the patterns of loadings and corresponding significances, it appeared that there was strong evidence for participant-level method effects, and weak evidence for a method effect on the daily occasion level. Both the trait and the method factors were thus needed for a good model fit.

DISCUSSION

At the between-subject level there was a significant systematic variance component both for questionnaire items and for aggregated diary items. The reliability of both methods was good, and the construct validity both for positive and negative support was sufficient but not impressive. The diary method had a higher reliability and appeared to have higher construct validity.

At the within-subjects level both the reliability and validity were low. For the positive support items, the reliability and construct validity appeared sufficient. For the negative items, the reliability was very low. This is only partly the result of including fewer items (five in the positive scale and two in the negative scale) because the trait loadings were also low. In addition, the occurrence of the negative events was so rare that these items had to be dichotomized to make statistical analysis at all possible. The conclusion here is that a model that assumes a continuous latent trait "propensity to provide support" over time (common to both reliability analysis and MTMM) is defensible for positive support, although reliability should be increased by using more questions if possible. In contrast, such a propensity model is implausible for negative support. Negative support events are so rare and have such a low covariation that they are better described as singular events, difficult to model as endogenous variables, but possibly influential as exogenous variables. Furthermore, in our opinion questions on support (both positive and negative) over time may be better treated as formative or causal indicators (Bollen & Lennox, 1991), which can be combined into an index even if the common variance is low.

From a practical point of view, it is attractive, if both retrospective questions and diary data are available, that at the participant level they are combined into a single variable or treated as multiple outcomes in a multivariate analysis. At the within-subjects level, reliability for positive support can be increased by adding more questions. At the same time, if the number of response categories is increased, the amount of information obtained will be increased and the statistical analysis will be facilitated because the data can be treated as continuous. However, for negative support events, this approach is unlikely to succeed, because even with four-category items, the higher categories were almost never used.

ACKNOWLEDGMENTS

We thank Matthieu Brinkhuis, Lotje Cohen, Jojanneke Dirkzwager, and Johanneke Maenhout for their part in the data analyses. Furthermore, we are grateful to Roeline Kuijer and Jozien Bensing for their comments on an earlier draft of this article.

REFERENCES

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Grilli, L., & Rampichini, C. (2005). *Multilevel factor models for ordinal variables.* Retrieved March 20, 2006, from http://www.ds.unifi.it/rampi/multilevel_ordinal_factor.pdf

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hoyle, R. H., Harris, M. J., & Judd, C. M. (2002). *Research methods in social relations.* Belmont, CA: Wadsworth.

Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships, 2,* 67–75.

Little, R., & Rubin, D. (2004). *Statistical analysis with missing data* (3rd ed.). New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marsh, H., & Grayson, D. (1995). Latent variable models of multitrait–multimethod data. In R. Hoyle (Ed.), *Structural equation modeling* (pp. 177–198). Thousand Oaks, CA: Sage.

Muthén, L. K., & Muthén, B. O. (2006). *Mplus' users guide* (3rd ed.). Los Angeles: Muthén & Muthén.

Newsom, J. T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling, 9,* 431–447.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16,* 295–330.

Raudenbush, S. W., & Sampson, R. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observations of neighborhoods. *Sociological Methodology, 29,* 1–41.

Raykov, T., & Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling, 13,* 130–141.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9,* 195–212.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

van Sonderen, E. (1991). *Het Meten van Sociale Steun [The measurement of social support].* Groningen, The Netherlands: Groningen University.

Widaman, K. A. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–21.