ELSEVIER

Social Networks 24 (2002) 1–20

SOCIAL NETWORKS

www.elsevier.com/locate/socnet

Estimating the reliability and validity of personal support measures: full information ML estimation with planned incomplete data

Tina Kogovšek [a,*], Anuška Ferligoj [a], Germa Coenders [b], Willem E. Saris [c]

[a] *Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia*
[b] *University of Girona, Girona, Spain*
[c] *University of Amsterdam, Amsterdam, The Netherlands*

Abstract

Egocentered networks are common in social science research. Here, the unit of analysis is a respondent (ego) together with his/her personal network (alters). Usually, several variables are used to describe the relationship between egos and alters.

In this paper, the aim is to estimate the reliability and validity of the averages of these measures by the multitrait–multimethod (MTMM) approach. This approach usually requires at least three repeated measurements (methods) of the same variable (trait) for model identification. This places a considerable burden on the respondent and increases the cost of data collection.

In this paper, we use a split ballot MTMM experimental design, proposed by Saris (1999), in which separate groups of respondents get different combinations of just two methods. The design can also be regarded as having a planned missing data structure. The maximum likelihood estimation is used in the manner suggested by Allison (1987) of a confirmatory factor analysis model for MTMM-designs specified in Saris and Andrews (1991). This procedure is applied to social support data collected in the city of Ljubljana (Slovenia) in the year 2000.

Keywords: Egocentered networks; Reliability; Validity; Social support

1. Introduction

Egocentered networks have become an important part of social science research. This study deals with the quality of measurement of social network data as measured with a social survey. The types of social network data considered here are egocentered networks

* Corresponding author. Tel.: +386-1-58-05-285; fax: +386-1-58-05-101.
E-mail address: tina.kogovsek@guest.arnes.si (T. Kogovšek).

0378-8733/02/$ – see front matter
PII: S0378-8733(01)00047-8

as opposed to complete networks. A complete network consists of a group of individuals with one or more relations defined among them. In contrast, an egocentered network (also called a personal network) consists of a single individual (usually called ego) with one or more relations defined between him/her and a number of other individuals—the members of his/her personal network (called alters).

Egocentered networks in this study are defined as personal social support networks. In the literature, there exist (e.g. Weiss, 1974; Cobb, 1976; Thoits, 1982; Vaux, 1988; Veiel and Baumann, 1992) many definitions of social support. Earlier definitions stressed the emotional dimension of social support, thus focusing on social support as a feeling of belonging, acceptance and care from important others. More recent conceptualizations (e.g. Burleson et al., 1994) put more stress on social support as an interactive communication process among people.

Many studies (e.g. Weiss, 1974; Hirsch, 1980; Wills, 1985: an overview in Vaux, 1988) also show that social support is a multi-dimensional concept. It can be subdivided into four major groups (dimensions): (1) instrumental support, (2) informational support, (3) emotional support, and (4) social companionship. These dimensions of social support were also the content of the name generators used in this study (see Appendix A). A name generator is a question for eliciting the names of the ego's network members (alters).

Usually, several characteristics (variables) are measured, ones which describe ego's relationships (frequently called ties) with his/her alters and the characteristics of alters themselves. Tie characteristics may involve for instance, the type of relation between the ego and the alter (e.g. friend, parent, spouse), feelings of closeness or importance, duration of the tie and so on. Alter characteristics are usually demographic variables, such as, sex, age, education, income of the alter and other similar factors. These kinds of questions are frequently called name interpreters.

In this paper, the aim is to estimate the reliability and validity of some of the very frequently used name interpreters. Since the data about the characteristics of ties function as important explanatory variables in social support research and are, moreover, usually reported by the ego, it is very important to know, to what extent these data are reliable and valid. However, as the units of analysis here are egocentered networks as a whole and not individual ego–alter ties, the variables are defined as averages on these variables across all egocentered networks. The use of averages is further justified by the fact that averages on these variables are very often used in the substantive research on social support. Therefore, the reliability and validity of the averages of these variables are studied.

The reliability and validity are estimated via the multitrait–multimethod (MTMM) approach. MTMM-designs (Cambell and Fiske, 1959) consist of multiple measures of a set of factors (traits) with the same set of measurement procedures (methods). These designs include $t \times m$ measures, that is the number of methods (m) times the number of traits (t). Validity and reliability are conceptualized in many different ways. However, broadly speaking, validity can be defined in the sense that a measurement instrument (e.g. a questionnaire) is valid if it really measures the concepts (in this case, the characteristics of ties in respondents' social support networks) that it is supposed to measure. Reliability is then defined as the ability of the measurement instrument to produce the same results in a repeated measurement (Carmines and Zeller, 1979). In the context of MTMM-designs, validity is conceptualized as the extent to which measurements are free of the so-called

method effects. Method effects tend to inflate correlations among variables measured with the same method, in comparison with correlations among variables measured with different methods. Validity can then be inferred from the comparison of these two sets of correlations. This is the way how the design manages to separate reliability and validity, even if no replicate measurements of the same trait with the same method are made.

For model identification purposes the MTMM approach usually requires at least three repeated measurements of the same variable (called trait in this approach), using three different methods (Kenny, 1976). This places a considerable burden on the respondent and increases the cost of data collection. To reduce these problems, a split ballot MTMM experimental design (Saris, 1999) is proposed in which separate groups of respondents get different combinations of just two repetitions (methods).

The purpose of this study is to establish which measurement method gives data of the best quality in terms of reliability and validity for the assessment of ties in social support egocentered networks. [1] Egocentered networks and characteristics of ties in them can be measured by different data collection modes, e.g., face-to-face interview, telephone, or via mail or Internet. As the costs of the telephone mode are lower than those of the face-to-face mode, the question is whether egocentered network data are as reliable and as valid when collected on the telephone as with face-to-face interviews. It is also, an open question whether name interpreter questions concerning multiple alters are best organized by alters or by questions. After we obtain the list of alters with name generators, we can ask name interpreter questions in two ways. One way ("by alters") is to take each alter individually and to ask all questions about him/her, going alter by alter until the end of the list of alters. The other way ("by questions") is to take the question and ask this question for all alters on the list, going question by question until the end of the list of name interpreter questions has been reached.

The three different methods in this study are defined as a combination of two factors: data collection mode (face-to-face and telephone interview) and data collection technique (by alters and by questions). The methods are:

1. face-to-face interview/by alters;
2. telephone interview/by alters;
3. telephone interview/by questions.

The quality of egocentered network measurements, when the questions are organized by alters or by questions has not been studied yet. However, we suspect that there may be differences in measured name interpreter questions. A different cognitive reference frame operates depending on whether tie and alter characteristics are measured by alters or by questions. In the first case, when the respondent answers all the questions for each alter, the reference frame is the current alter. In the second case, when the respondent answers each question for all alters, the reference frame is, first, the current question. On the other hand, it is also possible that at least part of the entire named personal network is activated (thus

[1] A number of similar data quality studies have been done on complete networks (e.g. Ferligoj and Hlebec, 1995, 1998, 1999). However, measurement issues are to some degree specific whether we measure complete or egocentered networks, so the aforementioned studies are not directly comparable to the present study. Also, factors affecting data quality used in those studies were different. For instance, in the case of complete networks, it was found out that the order of methods, time between repetitions, type of the response scale and domain of social support have a considerable effect on the quality of measurement (especially reliability).

representing the context for that particular question), because the respondent describes a certain tie or alter characteristic for all named alters. It is therefore possible that with each question the respondent actually compares the current alter with, and ranks him/her against one or more of the preceding alters on the list. It is less likely that alters are treated as separate and independent units. Therefore, context effects would be more present in the case of data collection by questions.

Knowledge about important people in a person's life is a form of autobiographical memory. It may be defined as "knowledge and schemata that form the memorial basis of the self" (Baddeley, 1992: 14). It is often divided into the two broader categories of episodic and semantic (also called generic) memory (e.g. Baddeley, 1992; Conway, 1992; Brewer, 1993). Whereas episodic memory refers to particular events (or episodes) in one's life, semantic memory refers to a person's abstract knowledge about oneself and his/her personal history. A number of studies (for an overview see Conway, 1992) have shown that autobiographical memory is organized thematically. There is also evidence (see in Conway, 1992) that this general, thematic autobiographical memory knowledge is very important for the retrieval of information, and in this sense constitutes the basic level of autobiographical knowledge of central importance for accurate, economical and efficient memory retrieval. It seems from the studies (see Conway, 1990) that autobiographical memories are organized hierarchically, with the more general themes at the top of the hierarchy and memories about specific events at the bottom level. Within this model retrieval of information proceeds by searches through the hierarchical structures (where different levels index one another), possibly top down, from more general thematic structures to more specific topics or events. Evidence from both theory and research thus suggests that (autobiographical) memory retrieval is a complex and effortful process (e.g. see Conway, 1990, 1992).

The general theme in this study is about personal relationships and, more specifically, the people that are for various reasons important in the respondent's life. The name generators then set the different subdomains (defined by different social support dimensions) into which these people are "classified" by the respondent. The members of the respondent's personal network belong to different (though to some degree overlapping) semantic contexts. On the next level, each individual alter can be considered a special semantic unit and as setting the context for memory search and forming the response. [2] Therefore, when the "by questions" data collection technique is used, the respondent not only has to perform a complex memory search and form a response, but also has to switch between contexts (alters), repeating this complex cycle of finding an answer for each succeeding alter. This is probably more cognitively demanding and time consuming than with the "by alters" technique, where the context (alter) remains the same for all questions asked. Therefore, in comparison to the "by questions" data collection technique, the "by alters" technique may provide a greater degree of correspondence to the most efficient methods of data retrieval from autobiographical

[2] This is in line with the so-called mental models approach. Broadly, mental models can be defined as complex, integrated sets of information that each person has about the world. A special case of mental models are models of particular individual persons, containing different types of information and dimensions of these persons and relationships with them (for more about this topic see e.g. Holmberg and Holmes, 1993). There is also some evidence (an overview in Kihlstrom and Hastie, 1997) that the memory about persons is organized in such a hierarchical way.

memory. We might expect that every tie or alter characteristic would be more reliably and more validly measured when the question is posed by alters than by questions.

Answering questions about network members and the characteristics of ties involved poses a considerable cognitive task for the respondent. It asks of him/her to process a large amount of information (dimensions of ties, alter characteristics), while adding relatively complex, demanding and potentially time consuming cognitive processes (estimating closeness and importance of the tie, frequency of contact, length of the relationship, etc.). Research shows (Groves and Kahn, 1979; Groves, 1978, 1979, 1989) that certain characteristics of telephone communication (time limit, speed of answering, lack of nonverbal communication) result in shorter answers to open-ended questions, fewer responses to multiple-choice questions: in general, respondents taking less time to answer and more missing values are produced. Survey research also shows, that the amount of effort by respondents affects the quality of recall in surveys (e.g. Krosnick, 1991). The pace of the interview can affect the respondent's effort and therefore recall. Increased response time is shown to increase accuracy of reporting and also correlations between variables. In the Burton and Blair (1991), respondents who were left to their own devices (there was no suggestion on the part of the interviewer that they could take more time with the answer) tended to answer more quickly at the cost of accuracy. We therefore expect that cognitively more demanding questions would be more prone to measurement errors in the telephone than in the face-to-face mode. Therefore, it is expected that demanding questions would be more reliably and more validly measured in face-to-face interviews than by telephone.

2. MTMM models

We approach the problem of assessing data quality from the standpoint of the well-known and widely used MTMM approach (Cambell and Fiske, 1959), which was developed in the late 1950s, but has been further developed over succeeding decades.

Basically, Campbell and Fiske proposed evaluation of measurement instruments to be done by measuring a number of traits with different methods. A correlation matrix (also called an MTMM matrix) is then produced and analyzed by a set of criteria, also proposed by Campbell and Fiske.

However, the MTMM matrix is no longer analyzed in this way. Since the introduction of the path analytic approach in the 1970s, these matrices are usually analyzed by confirmatory factor analysis models, a particular case of structural equation models (e.g. Bollen, 1989).

A number of MTMM models have been formulated and tested (see van Meurs and Saris, 1990; Saris and Andrews, 1991; Ferligoj et al., 1995; Saris and Münnich, 1995; Scherpenzeel, 1995c; Coenders and Saris, 2000) by an international group on methodology and comparative survey research (IRMCS) and others (e.g. Althauser et al., 1971; Alwin, 1974; Werts and Linn, 1970; Browne, 1984, 1985; Marsh, 1989; Marsh and Bailey, 1991).

The model that seems to be the most useful is the true score model as proposed by Saris and Andrews (1991). One of the advantages of this model is that it is possible to estimate reliability and validity separately.

The true score model (Fig. 1) is defined as follows:

$$Y_{jk} = h_{jk} T_{jk} + e_{jk} \tag{1}$$

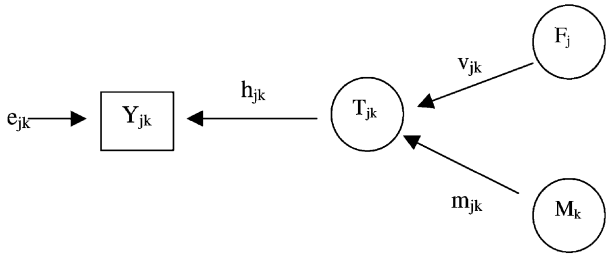

Fig. 1. True score measurement model.

$$T_{jk} = v_{jk}F_j + m_{jk}M_k \tag{2}$$

where

- Y_{jk} is the measured variable (trait F_j measured by the k-th method);
- T_{jk} is the stable component of the response Y_{jk} (also called the true score);
- F_j is the trait and
- M_k is the variation in scores due to the k-th method.

It is assumed that the e_{jk} random error terms are uncorrelated with one another, with method factors and with trait factors. It is assumed that method factors are uncorrelated

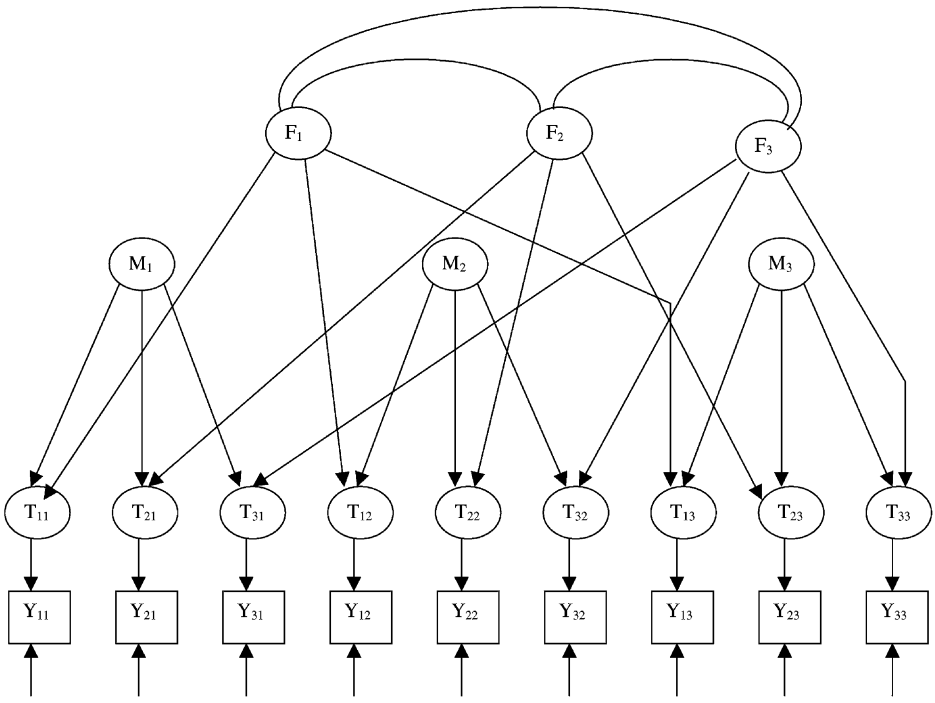

Fig. 2. The path diagram for the 3trait–3method true score model.

with one another and with trait factors. The unstandardized m_{jk} coefficients are constrained to be equal within a method and the unstandardized h_{jk} coefficients are constrained to be equal to 1 to fix the scale of the T_{jk} factors Fig. 1.

If all the variables are standardized the standardized parameters represent the following:

- h_{jk} is the reliability coefficient (h^2_{jk} being the test–retest reliability);
- v_{jk} is the validity coefficient (v^2_{jk} representing the validity of the measure) and
- m_{jk} is the method effect where $m^2_{jk} = 1 - v^2_{jk}$, which means that the method effect is equal to the invalidity of the measure.

At least three traits have to be measured by at least three different methods in order to render the model identified. The path diagram for the 3trait–3method true score model is shown in Fig. 2.

3. The design of the study

Answering three times is a tedious task for respondents. Therefore, we decided to use a form of a split ballot MTMM-design, first proposed by Saris (1999). In his design, respondents were randomly assigned to two groups with different combinations of methods. On the first occasion, all respondents received the first method, but on the second occasion one group received the second and the other group the third method. A somewhat different planned incomplete data MTMM-design, with the aim of similarly reducing respondent burden, was tested by Bunting and Adamson (2000). In their design, respondents were randomly assigned to three groups, by adding a third group with complete three-method data—a design, which was otherwise close to that of Saris (1999). Bunting and Adamson (2000) compared original validity and reliability estimates based on complete real data with the estimates obtained with partly missing data. The results showed that it is possible to reproduce reliability and validity estimates, even when part of the data are missing, with a substantial degree of precision; this result, therefore, shows the usefulness of such designs for MTMM-design in terms of reducing respondent burden and the costs of the survey.

In our study, an intermediate design was used, with three groups, each with two out of the three methods, which is displayed in Table 1. Unlike in Saris (1999), all possible pairs of methods had data, which, for a 3trait–3method design, increased the degrees of freedom (d.f.) by 21 and prevented 9 variances and covariances from being unobserved. Unlike in Bunting and Adamson (2000), no group had complete data, in order to avoid the cost of one extra wave and burdening part of the sample with three repeated measurements. The MTMM correlation matrices used are in the Appendix B.

Table 1
The design of the study

Group	*N*	First interview	Second interview
1	320	Face-to-face/by alters	Telephone/by alters
2	311	Face-to-face/by alters	Telephone/by questions
3	402	Telephone/by alters	Telephone/by questions

A population study using simulated data was carried out to compare the precision of the estimates for the three-group and the two-group designs. Keeping the total sample size and the data collection burden constant (e.g. comparing two groups of $n = 450$ with three groups of $n = 300$; all groups being measured with two methods), the sampling variances of the parameter estimates were on average 3.1 times larger for the two-group design. The results thus showed that much more precise estimates will be obtained if a given total sample size is divided into three groups than if divided into two groups.

Practical considerations also support the use of three groups. Even if the model is theoretically identified with two groups, it proved to be unstable and failed to converge to an admissible solution with our data in one case. [3]

Moreover, in the present study, the true score model was used with three traits and three methods, thus using an MTMM covariance matrix with nine measured variables. Because of the split ballot design, three such covariance matrices were used, one for each group of respondents. In each covariance matrix, the covariances and variances for the omitted variables were missing. The LISREL8 program (Jöreskog and Sörbom, 1993) was used to estimate the models with such deliberately incomplete data following guidelines of Allison (1987) for maximum likelihood.

The χ^2-test of the model is sensitive to two kinds of misspecification: different population covariance matrices for the three groups and violations of the assumptions of the model. In case a model is rejected, it is unknown which of the two is the cause. Moreover, a lack of comparability of populations actually invalidates the split ballot design, no matter what model specification may be used. A test of invariance of the variances covariances that are present in more than one group (see Appendix B) must then be carried out first.

4. Data

The data were collected between March and June 2000 by computer-assisted telephone interview (CATI) and computer-assisted personal interview (CAPI) for a representative sample of 1033 inhabitants of the city of Ljubljana, Slovenia. The sampling frame was the telephone directory of Ljubljana. Respondents were randomly assigned to the three groups specified in the design. These respondents produced 7223 alters. [4]

The traits used in this study are represented by two widely used measures of tie strength (for an overview on this topic, see Marsden and Campbell, 1984): the frequency of contact of ego with each alter and feelings of closeness of ego towards each alter. [5] The third trait is

[3] We have estimated the same model with several different trait combinations.

[4] A serious problem with repeated measurements is the memory effect. If the time period between two repetitions is too short, respondents remember their previous answers and therefore artificially increase the reliability of the measurement instrument. In order to avoid memory effects, it has been shown (van Meurs and Saris, 1990) that the minimum time between two repetitions within the same questionnaire is at least 20 min, as long as similar questions have been asked in between and the opinion of the respondents is not extreme. On the other hand, neither should the period between two measurements be too long, in order to avoid a possible actual change in the studied phenomenon, which could then artificially reduce the reliability of measurement. In the present study, the time between the two repetitions was 1 week, a period which can be supposed to minimize both the above mentioned effects.

[5] In several additional analyses also the importance of each alter is used.

represented by a measure of negative aspects of social relationships: the frequency of alter upsetting the ego. [6] The wording of the questions can be seen in the Appendix A.

However, as the units of analysis are egocentered networks as a whole and researchers often use aggregated variables in their analyses, the traits in this study actually represent the average frequency of contact of ego with his/her alters, the average closeness between ego and his/her alters and the average frequency of alters upsetting the ego within these networks. [7]

5. Results

A model, whose only restrictions are equality of covariances and variances across groups, has a χ^2 statistic of 21.2 with 18 d.f. and *P*-value of 0.27, whence the equal covariance hypothesis is tenable in our case.

The hypothesis that the model's constraints would hold exactly for this population was rejected in statistical terms (χ^2 statistic of 89.57, with 39 d.f., *P*-value 0.000). The model is nested into the former equal-covariance model, which allows us to compute the χ^2 change statistic, which at 68.3 (21 d.f.) also leads to the rejection of the model's constraints. However, the power of the χ^2 tests is reckoned to be quite high, due to the fact that the sample sizes and the reliability and validity estimates are all relatively large (Saris and Satorra, 1988). This may have caused the model to be rejected out of minor misspecification errors and calls for the use of descriptive fit indices that quantify the amount of misspecification. According to these indices, the goodness of fit of the model deserves a rather favorable judgement. The standardized root mean squared residual was 0.057. Recent research has shown the Tucker and Lewis non-normed fit index to be independent of sample size and to adequately penalize complex models (Marsh et al., 1996). This index had a value of 0.969 for our model and another widely used index, Bentler's comparative fit index was 0.973, both well above the commonly accepted threshold of 0.95. We proceed then to an interpretation of the standardized measurement quality estimates (Table 2).

As we can see from the patterns of validity coefficients, all three traits are most validly measured by the telephone mode (except the first trait, which is equally valid with all three methods). The first method (face-to-face) gives the least valid measurement.

As for the reliability coefficients, the telephone method by alters seems to be most reliable, except for the first trait, where it is the second most reliable. Telephone by questions is the least reliable method. Face-to-face is the most reliable method for the first trait (frequency of contact), but less reliable for other two traits. Let us note here that we have tried the same model with several different combinations of traits and have found that results were generally

[6] Negative aspects of social relationships and their affect on emotional well-being have become an important part of social support research. Negative exchanges also affect the perceived social support and can be even a stronger predictor of perceived social support than the positive aspects of interactions (for an overview see Rook, 1992; also Pierce et al., 1992; Pierce, 1994).

[7] A further advantage of using averages is that they depart less from a normal distribution than the individual tie variables. For instance, for average frequency of alters upsetting the ego, the skewness coefficients are between 0.20 and 0.39. Other skewness coefficients mostly range from −0.13 to −0.45. The most skewed is average frequency of contact at the second measurement by telephone (−0.90).

Table 2
Validity and reliability coefficients

Method	Contact	Closeness	Upset
Reliability coefficients			
Face-to-face/by alters	0.94	0.80	0.80
Telephone/by alters	0.85	0.88	0.85
Telephone/by questions	0.81	0.76	0.83
Validity coefficients			
Face-to-face/by alters	0.97	0.92	0.93
Telephone/by alters	0.97	0.97	0.96
Telephone/by questions	0.98	0.96	0.97

similar.[8] In particular, the pattern of validity coefficients was quite stable across different combinations of traits, with the telephone by alters being the most and face-to-face by alters being the least valid method. Reliability coefficients were the highest for telephone by alters for all traits, with the exception of frequency of contact, which had the highest reliability coefficient in the face-to-face condition. The lowest reliability coefficients were obtained either by face-to-face or telephone by questions, depending on the trait combination used.[9]

6. Discussion

Possible explanations for these results could include the following factors. Telephone mode may be more valid than face-to-face because it is more anonymous. A person's relationships with other people (especially the closest people, a category which social support name generators tend to elicit) constitute a relatively sensitive topic, one therefore more susceptible to the possibility of producing socially desirable responses. It may be easier to speak about these kinds of relationships and their characteristics over the telephone. It is also less likely that someone would overhear what was said in the interview over the telephone than if the interview was conducted in the respondent's home, with the possible presence of other household members. Some support for this anonymity explanation could be found in the fact that validity coefficients are higher for the telephone than for the face-to-face mode. Also the most sensitive questions about alters (e.g. feelings of closeness and importance, how often the alter upsets the ego) have lower validity with the face-to-face mode than by telephone.[10]

Another very likely explanation could also be that the telephone is a faster means of communication (and time limited; therefore, respondents probably tend to name only the

[8] For instance, some of the other trait combinations used were: frequency of contact, feelings of closeness, feelings of importance; feelings of closeness, feelings of importance, frequency of alter upsetting the ego; frequency of contact, feelings of importance, frequency of alter upsetting the ego. Also a 4trait–3method model was produced with frequency of contact, feelings of closeness, feelings of importance and frequency of alter upsetting the ego as traits.

[9] Reliability and validity estimates with other trait combinations are in the Appendix C.

[10] For instance, Költringer (1995) found out that the type of question and sensitivity of the topic had some affect on the quality estimates, especially reliability. If the questions are not sensitive, factual/behavioral questions produce more reliable data than feelings/attitude questions. If the topic is sensitive, the reliability of both types of questions is equally reduced.

most important alters (which are usually better measured precisely because of this characteristic).[11] If we compare only the telephone mode, the formula "by alters" seems to be better, as we expected. Additionally, reliabilities for the "by alters" technique are better than those for the "by questions" technique. With this data collection technique, the salient reference frame is the current alter, and it is more likely that the respondent would make fewer errors. On the other hand, if the data collection technique is "by questions", it is likely that the current question and also part of the named egocentered network are "fighting" for the respondent's working memory space, thus increasing the cognitive burden and making errors more likely. Another consequence of the speed of the telephone mode may also be smaller elicited networks, which would reduce respondent burden and therefore lead to data of better quality. However, this proved not to be the case in this study, since the mean network size was actually smaller in the face-to-face (6.84) than in the telephone mode (7.24) and the difference in mean network size was statistically significant (at 0.05 level).

On the other hand, frequency of contact is less reliably measured over the telephone. The reason may again lie in the speed of telephone communication. Frequency of contact is a type of question different, for instance, from the question about the degree of closeness between the ego and the alter (or the importance of the alter, which was used in another trait combination). The former is a question about actual behavior, whereas, the latter is about subjective, personal feelings toward the alter. Answering the question about frequency of contact is probably more cognitively demanding than answering about the degree of closeness for which the answer may be more readily available. Since these are in most cases the respondents' most important people, the respondents probably have a fairly good idea (a mental model) of what their feelings towards these people are; they can, therefore, answer such a question quickly as well as fairly accurately (which is an important feature if the interview is conducted by telephone). On the other hand, it is not very likely that respondents would have an appropriate answer to most behavioral questions already available at the time of answering the question. Instead, an estimation procedure is used to come up with an answer (whether counting or rate judgement), which is prone to error and can also be time consuming, features which do not work well with fast telephone communication, but may work better with the face-to-face mode which "allows" more time for the respondent to think about the answer (a good recent overview of cognitive processes in surveys appears in Sudman et al., 1996).[12] This result thus gives support to the hypothesis that cognitively more demanding name interpreter questions would be more reliably measured by face-to-face than by telephone mode.

It can also be seen that the reliability of the telephone mode (at least in combination with the "by alters" data collection technique) improves when compared with the face-to-face mode. At this point, we must note that the order of traits in this model is a the same as that in the questionnaire. It is therefore also possible that with questions posed later,

[11] There is some indication about that, but none of the differences are statistically significant. Mean closeness and mean importance of alters across networks are slightly higher in the telephone mode than in face-to-face mode. Also mean coefficients of variation are about the same in both data collection modes.

[12] This argument lines up well with the assertion that people are "cognitive misers", rather following a "satisficing" than "optimizing" strategy, and then tend to truncate the search for the answer as soon as enough information has been retrieved to produce the answer, which is not necessarily very accurate (a well-known study of these processes was made by Krosnick, 1991).

respondents become used to the format for answering questions, and so face-to-face mode loses its initial advantage of potentially longer and more thorough thought about the answer in comparison with telephone mode. A slight trend is evident towards improved reliability of the telephone mode ("by alters") compared with face-to-face with later questions (with some trait combinations). However, this hypothesis should be tested in a design with a different order of questions about alters. If this learning effect is indeed present, it would be advisable to place easier questions about alters and the characteristics of ties first (for instance the easier demographic characteristics such as sex of the alter, type of relationship and so forth), and more demanding questions later, when the respondent has already become used to the responding technique.

Another possible explanation for both telephone methods having higher validity coefficients may be that the usual assumption of independent methods is incorrect. If some of the methods are too similar (in this case they share the data collection mode), they may lead to an overestimation of the validity estimates (Wit, 1995; Wit and Billiet, 1995). On the other hand, the study of Scherpenzeel (1995b) showed that the validity biases in this context are dependent on two factors: the size of the correlation between methods and the sizes of the loadings of the method factors. [13] The results of her study showed that even with a high correlation between two methods introduced into the model (0.50 in her example), the validity coefficients do not change greatly (only by 0.01) if the loadings on the method factors are not too high (in her example 0.20). Scherpenzeel (1995b) suggested trying the model with and without method correlations and seeing to what degree the reliability and validity coefficients change.

In our case, the loadings on method factors for the telephone/by alters and telephone/by questions are still relatively small, ranging from 0.25 to 0.31. The correlation between the two methods was introduced into the model and turned out to be very small. Therefore, the reliability and validity coefficients in this model are not substantially different from the coefficients obtained for the model without a correlation between the two methods. Some of the reliability and validity coefficients differed from −0.01 to 0.01. Therefore, the effect of the correlation between telephone/by alters and telephone/by questions on the reliability and the validity coefficients can be excluded.

Another possible explanation for these results may be the ordering of data collection modes. [14] Face-to-face may be the worst because it was always used for the first interview. Telephone/by questions was always used for the second interview and this may be the reason why it is in most cases better than face-to-face, despite the more difficult data collection technique (by questions). Telephone/by alters was used for the first interview in one group of respondents and for the second interview in another group; yet it still emerged as the most valid and reliable data collection method. If the method order explanation holds, then the method used in the second interview would prove to be best in all instances. Telephone/by questions does not emerge as the best method, despite the fact that it was always used in

[13] Almost no effect was found for the reliability estimates.

[14] For instance, compare to Scherpenzeel (1995a) and Ferligoj and Hlebec (1998). Both studies show that methods used in the first measurement have lower reliability and validity than methods used in subsequent measurements. The explanation is that respondents become familiar with the data collection techniques after the first measurement and can therefore provide more reliable and more valid answers at later measurements.

the second interview. On the other hand, face-to-face/by alters was always the first method used (and should therefore come out as the worst), but it comes out as the better method in some instances (for some traits). So it may be that other factors play more of a role (anonymity and speed of the telephone mode, type of question) than method order itself. However, this problem cannot be tested within the present design and remains a topic for further investigation.

Partly as a result of the discussions in the IRMCS group upon encountering this problem in this study, new split ballot MTMM-designs have recently been formulated and tested (Saris et al., 2001). In these designs, the same method is presented twice to the same respondents with the aim of isolating order effects from method effects. Another design effect that may have confounded the results is the fact that not all possible combinations of data collection mode and question ordering have been included in the study. Our method effect coefficients actually estimate the sum of mode and ordering effects. Within the new designs also, method factors could be studied separately. We are also working on an alternative specification of the model with two sets of method factors, one referring to data collection mode (face-to-face versus telephone) and one referring to data collection technique (by alters versus by questions).The results are encouraging although models with so many latent variables tend to be unstable and have large standard errors.

7. Conclusions

Several major conclusions can be drawn. As far as the split ballot MTMM model is concerned, it can be concluded that it is possible to use models with deliberately designed gaps in data. Models converge and give acceptable reliability and validity estimates. There are at least three advantages to such missing data designs: reducing respondent burden, reducing time and financial costs of the survey and shortening the elapsed time between first and last wave and thus making the assumption of stability of the trait scores between the waves more reasonable. It should also be mentioned that it is advisable to use the three-group split ballot MTMM-design due to its greater statistical efficiency in comparison to the two-group design.

Despite the fact that the method order effect was not completely controlled, we can say that contrary to a quite common assumption that the face-to-face mode produces data of better quality, the telephone mode also appears to produce good quality data. Since telephone interviewing is less costly and less time consuming, this is an important finding as to how such studies could be done in the future.

Some directions for further research would be:

1. including new groups into the design to study possible method and/or trait order effects and to study the stability of obtained results;
2. estimating SEM models with more than one set of method factors (e.g. Scherpenzeel, 1995c, Chapter 6) to distinguish one set (face-to-face/telephone) from the other set (by alters/by questions) of method factors;
3. considering new modes of data collection (for example, web survey).
4. dealing (in this paper) with the quality of measurement of tie characteristics on the aggregated level and not of responses about individual ego–alter ties if the researcher

would be interested in the responses on the level of individual alters, multilevel structural equation models could be used (e.g. Snijders and Bosker, 1999) with alters nested within the egos.

Acknowledgements

The authors would like to thank Brendan Bunting, Valentina Hlebec and the anonymous reviewers for valuable comments on earlier drafts of the paper.

Appendix A

A.1. Name generators used in this study

1. From time to time, people borrow something from other people, for instance a piece of equipment, or ask for help with small jobs in or around the house. Who are the people you usually ask for this kind of help? (Material support)
2. From time to time, people ask other people for advice when a major change occurs in their life, for instance, a job change or a serious accident. Who are the people you usually ask for advice when such a major change occurs in your life? (Informational support)
3. From time to time, people socialize with other people, for instance, they visit each other, go together on a trip or to a dinner. Who are the people with whom you usually do these things? (Social companionship)
4. From time to time, most people discuss important personal matters with other people, for instance if they quarrel with someone close to them, when they have problems at work, or other similar situations. Who are the people with whom you discuss personal matters that are important to you? (Emotional support)
5. Suppose you would find yourself in a situation, when you would need a large sum of money, but do not have it yourself at the moment, for instance five average monthly wages (approximately 500.000 tolars). Whom would you ask to lend you the money (a person, not an institution, e.g. a bank)? (Financial support)

A.2. Name interpreters used in this study

1. How frequently are you in contact with this person (personally, by mail, telephone or Internet) (frequency of contact)?

   1.1 Every day.
   1.2 Several times a week.
   1.3 Several times a month.
   1.4 About once a month.
   1.5 Several times a year.
   1.6 Less than once a year.

2. How close do you feel to this person? Please describe how close you feel on a scale from 1 to 5, where 1 means not close and 5 means very close (feelings of closeness).

1	2	3	4	5
Not close				Very close

3. How important is this person in your life? Please describe how close you feel on a scale from 1 to 5, where 1 means not important and 5 means very important (feelings of importance).

1	2	3	4	5
Not important				Very important

4. How often does this person upset you (frequency of alter upsetting the ego)?

   4.1 Often.
   4.2 Sometimes.
   4.3 Rarely.
   4.4 Never.

Appendix B.

Correlation matrices for traits: frequency of contact, feelings of closeness, frequency of alters upsetting the ego.

	M1T1	M1T2	M1T3	M2T1	M2T2	M2T3	M3T1	M3T2	M3T3
Correlation matrix, group 1 ($N = 320$)									
M1T1	1.0								
M1T2	0.264	1.0							
M1T3	0.152	−0.014	1.0						
M2T1	0.758	0.185	0.154	1.0					
M2T2	0.186	0.614	−0.023	0.223	1.0				
M2T3	0.077	−0.082	0.640	0.156	−0.008	1.0			
M3T1	–	–	–	–	–	–	1.0		
M3T2	–	–	–	–	–	–	–	1.0	
M3T3	–	–	–	–	–	–	–	–	1.0
Correlation matrix, group 2 ($N = 311$)									
M1T1	1.0								
M1T2	0.343	1.0							
M1T3	0.227	−0.045	1.0						
M2T1	–	–	–	1.0					
M2T2	–	–	–	–	1.0				
M2T3	–	–	–	–	–	1.0			
M3T1	0.715	0.204	0.088	–	–	–	1.0		
M3T2	0.104	0.519	−0.104	–	–	–	0.247	1.0	
M3T3	0.120	−0.123	0.625	–	–	–	0.087	−0.133	1.0

Appendix B. (*continued*)

	M1T1	M1T2	M1T3	M2T1	M2T2	M2T3	M3T1	M3T2	M3T3
Correlation matrix, group 3 ($N = 402$)									
M1T1	1.0								
M1T2	–	1.0							
M1T3	–	–	1.0						
M2T1	–	–	–	1.0					
M2T2	–	–	–	0.291	1.0				
M2T3	–	–	–	0.137	−0.074	1.0			
M3T1	–	–	–	0.623	0.225	0.101	1.0		
M3T2	–	–	–	0.150	0.634	−0.056	0.221	1.0	
M3T3	–	–	–	0.000	−0.091	0.628	0.125	−0.109	1.0

Labels: M1: face-to-face/by alters; T1: average frequency of contact; M2: telephone/by alters; T2: average feelings of closeness; M3: telephone/by questions; T3: average frequency of alters upsetting ego.

Note: Missing correlations are denoted with '–'. In the LISREL program, they are set to zero. The main diagonals contain the total variance (i.e. 1) and not test–retest reliability as is sometimes done.

Appendix C

Validity and reliability estimates (traits: frequency of contact, feelings of closeness, feelings of importance).

Method	Contact	Closeness	Importance
Reliability coefficients			
Face-to-face/by alters	0.96	0.88	0.85
Telephone/by alters	0.81	0.94	0.95
Telephone/by questions	0.84	0.88	0.89
Validity coefficients			
Face-to-face/by alters	0.91	0.80	0.84
Telephone/by alters	0.95	0.96	0.97
Telephone/by questions	0.90	0.84	0.88

Validity and reliability estimates (traits: feelings of closeness, feelings of importance, frequency of alters upsetting the ego).

Appendix C. (*continued*)

Method	Closeness	Importance	Upset
Reliability coefficients			
Face-to-face/by alters	0.85	0.81	0.80
Telephone/by alters	0.92	0.93	0.84
Telephone/by questions	0.84	0.86	0.79
Validity coefficients			
Face-to-face/by alters	0.85	0.88	0.86
Telephone/by alters	0.94	0.95	0.93
Telephone/by questions	0.92	0.95	0.93

Validity and reliability estimates (traits: frequency of contact, feelings of importance, frequency of alters upsetting the ego).

Method	Contact	Importance	Upset
Reliability coefficients			
Face-to-face/by alters	0.95	0.83	0.81
Telephone/by alters	0.84	0.84	0.87
Telephone/by questions	0.82	0.76	0.83
Validity coefficients			
Face-to-face/by alters	0.96	0.93	0.90
Telephone/by alters	0.97	0.97	0.96
Telephone/by questions	0.97	0.97	0.97

Validity and reliability estimates (traits: frequency of contact, feelings of closeness, feelings of importance, frequency of alters upsetting the ego).

Method	Contact	Closeness	Importance	Upset
Reliability coefficients				
Face-to-face/by alters	0.96	0.83	0.81	0.79
Telephone/by alters	0.83	0.91	0.92	0.86
Telephone/by questions	0.84	0.83	0.86	0.80
Validity coefficients				
Face-to-face/by alters	0.94	0.86	0.89	0.87
Telephone/by alters	0.96	0.95	0.97	0.95
Telephone/by questions	0.95	0.92	0.94	0.93

References

Allison, P.D., 1987. Estimation of linear models with incomplete data. In: Clogg, C.C. (Ed.), Sociological Methodology. American Sociological Association, Washington DC, pp. 71–103.

Althauser, R.P., Heberlein, T.A., Scott, R.A., 1971. A causal assessment of validity: the augmented multitrait–multimethod matrix. In: Blalock, H.M. Jr. (Ed.), Causal Models in the Social Sciences. Aldine, Chicago, pp. 151–169.

Alwin, D., 1974. An analytic comparison of four approaches to the interpretation of relationships in the multitrait–multimethod matrix. In: Costner, H.L. (Ed.), Sociological Methodology 1973–1974. Jossey-Bass, San Francisco, pp. 79–105.

Baddeley, A., 1992. What is autobiographical memory? In: Conway, A.M., Rubin, D.C., Spinnler, H., Wagenaar, W.A. (Eds.), Theoretical Perspectives on Autobiographical Memory. Kluwer, Dordrecht, pp. 13–29.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Brewer, W.F., 1993. Autobiographical memory and survey research. In: Schwarz, N., Sudman, S. (Eds.), Autobiographical Memory and the Validity of Retrospective Reports. Springer, New York, pp. 11–20.

Browne, M.W., 1984. The decomposition of multitrait–multimethod matrices. British Journal of Mathematical and Statistical Psychology 37, 1–21.

Browne, M.W., 1985. MUTMUM, decomposition of multitrait–multimethod matrices. Department of Statistics, University of South Africa, Pretoria, South Africa.

Bunting, B.P., Adamson, G., 2000. Assessing reliability and validity in the context of planned incomplete data structures for multitrait–multimethod models. In: Ferligoj, A., Mrvar, A. (Eds.), Developments in Survey Methodology, Metodološki zvezki, Vol. 15. FDV, Ljubljana, pp. 37–53.

Burleson, B.R., Albrecht, T.R., Sarason, I.G. (Eds.), 1994. Communication of Social Support: Messages, Interactions, Relationships, and Community. Sage, Thousand Oaks.

Burton, S., Blair, E., 1991. Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. Public Opinion Quarterly 55, 50–79.

Cambell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait–multimethod matrix. Psychological Bulletin 56, 81–105.

Carmines, E.G., Zeller, R.A., 1979. Reliability and Validity Assessment. Sage, Thousand Oaks.

Cobb, S., 1976. Social support as a moderator of life stress. Psychosomatic Medicine 38, 300–314.

Coenders, G., Saris, W.E., 2000. Testing nested additive, multiplicative and general multitrait–multimethod models. Structural Equation Modeling 7, 219–250.

Conway, M.A., 1990. Autobiographical Memory: An Introduction. Open University Press, Milton Keynes.

Conway, M.A., 1992. A structural model of autobiographical memory. In: Conway, A.M., Rubin, D.C., Spinnler, H., Wagenaar, W.A. (Eds.), Theoretical Perspectives on Autobiographical Memory. Kluwer, Dordrecht, pp. 13–29.

Ferligoj, A., Hlebec, V., 1995. Reliability of network measurements. In: Ferligoj, A., Kramberger, A. (Eds.), Contributions to Methodology and Statistics. FDV, Ljubljana, pp. 219–232.

Ferligoj, A., Hlebec, V., 1998. Quality of scales measuring complete social networks. In: Ferligoj, A. (Ed.), Advances in Methodology, Data Analysis, and Statistics, Metodološki zvezki, Vol. 14. FDV, Ljubljana, pp. 173–186.

Ferligoj, A., Hlebec, V., 1999. Evaluation of social network measurement instruments. Social Networks 21, 111–130.

Ferligoj, A., Leskošek, K., Kogovšek, T., 1995. Zanesljivost in veljavnost merjenja. FDV (in Slovene), Ljubljana.

Groves, R.M., 1978. On the mode of administering a questionnaire and responses to open-ended items. Social Science Research 7, 257–271.

Groves, R.M., 1979. Actors and questions in telephone and personal interview surveys. Public Opinion Quarterly 43, 190–205.

Groves, R.M., 1989. Survey Errors and Survey Costs. Wiley, New York.

Groves, R.M., Kahn, R.L., 1979. Surveys by Telephone: A National Comparison with Personal Interviews. Academic Press, New York.

Hirsch, B.J., 1980. Natural support systems and coping with major life changes. American Journal of Community Psychology 8, 159–172.

Holmberg, D., Holmes, J.G., 1993, Reconstruction of relationship memories: a mental models approach. In: Schwarz, N., Sudman, S. (Eds.), Autobiographical Memory and the Validity of Retrospective Reports. Springer, New York, pp. 267–288.

Jöreskog, K.G., Sörbom, D., 1993. New Features in LISREL8. Scientific Software International, Chicago.
Kenny, D.A., 1976. An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. Journal of Experimental Social Psychology 12, 247–252.
Kihlstrom, J.F., Hastie, R., 1997. Mental representations of persons and personality. In: V Hogan, R.J., Johnson, Briggs, S. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, pp. 711–735.
Költringer, R., 1995. Measurement quality in Austrian Personal Interview Surveys. In: Saris, W.E., Münnich, A. (Eds.), The Multitrait–Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press, Budapest, pp. 207–224.
Krosnick, J.A., 1991. Response Strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5, 213–236.
Marsden, P.V., Campbell, K.E., 1984. Measuring tie strength. Social Forces 63, 482–501.
Marsh, H.W., 1989. Confirmatory factor analysis of multitrait–multimethod data: many problems and few solutions. Applied Psychological Measurement 13, 335–361.
Marsh, H.W., Bailey, M., 1991. Confirmatory factor analyses of multitrait–multimethod data: comparison of the behavior of alternative models. Applied Psychological Measurement 15, 47–70.
Marsh, H.W., Balla, J.R., Hau, K.T., 1996. An evaluation of incremental fit indices: a clarification of mathematical and empirical properties. In: Marcoulides, G.A., Schumacker, R.E. (Eds.), Advanced Structural Equation Modeling: Issues and Techniques. Lawrence Erlbaum, Mahwah, NJ, pp. 315–353.
Meurs, A. van, Saris, W.E., 1990. Memory effects in MTMM studies. In: Meurs, A. van, Saris, W.E. (Eds.), Evaluation of Measurement Instruments by Meta-Analysis of Multitrait–Multimethod Studies. North-Holland, Amsterdam, pp. 134–147.
Pierce, G.R., 1994. The quality of relationships inventory: assessing the interpersonal context of social support. In: Burleson, B.R., Albrecht, T.L., Sarason, I.G. (Eds.), Communication of Social Support: Messages, Interactions, Relationships, and Community. Sage, Thousand Oaks, pp. 247–266.
Pierce, G.R., Sarason, B.R., Sarason, I.G., 1992. General and specific support expectations and stress as predictors of perceived supportiveness: an experimental study. Journal of Personality and Social Psychology 63, 297–307.
Rook, K.S., 1992. Detrimental aspects of social relationships: taking stock of an emerging literature. In: Veiel, H.O.F., Baumann, U. (Eds.), The Meaning and Measurement of Social Support. Hemisphere, New York, pp. 157–169.
Saris, W.E., 1999. Forced choice or agree/disagree questions? In: Proceedings of the Meeting of the IRMCS on the Evaluation by the Split Ballot MTMM Experiment, October 1999, pp. 122–146.
Saris, W.E., Andrews, F.M., 1991. Evaluation of measurement instruments using a structural modeling approach. In: Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S. (Eds.), Measurement Errors in Surveys. Wiley, New York, pp. 575–597.
Saris, W.E., Coenders, G., Satorra, A., 2001. New approach for evaluating quality of measurement instruments. In: Proceedings of the Meeting of the IRMCS on Split Ballot MTMM-design, Gent, Belgium, 25–26 May 2001.
Saris, W.E., Münnich, A. (Eds.), 1995. The Multitrait–Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press, Budapest.
Saris, W.E., Satorra, A., 1988. Characteristics of structural equation models which affect the power of the likelihood ratio test. In: Saris, W.E., Galhofer, I.N. (Eds.), Sociometric Research, Vol. 2, Data Analysis. MacMillan, London, pp. 220–236
Scherpenzeel, A., 1995a. Meta analysis of a European comparative study. In: Saris, W.E., Münnich, A. (Eds.), The Multitrait–Multimethod Approach to Evaluate Measurement Instruments, Eötvös University Press, Budapest, pp. 225–242.
Scherpenzeel, A., 1995b. Misspecification effects. In: Saris, W.E., Münnich, A. (Eds.), The Multitrait–Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press, Budapest, pp. 61–70.
Scherpenzeel, A., 1995c. A Question of Quality: Evaluating Survey Questions by Multitrait–Multimethod Studies. Royal PTT Netherlands, Amsterdam.
Snijders, T.A.B., Bosker, R.J., 1999. Multilevel Analysis. Sage, Thousand Oaks.
Sudman, S., Bradburn, N.M., Schwartz, N., 1996. Thinking About Answers: The Application of Cognitive Processes to Survey Methodology. Jossey-Bass, San Francisco.
Thoits, P.A., 1982. Conceptual, methodological and theoretical problems in studying social support as a buffer against life stress. Journal of Health and Social Behavior 23, 145–159.
Vaux, A., 1988. Social Support, Theory, Research, and Intervention. Praeger, New York.

Veiel, H.O.F., Baumann, U. (Eds.), 1992. The Meaning and Measurement of Social Support. Hemisphere, New York.

Weiss, R.S., 1974. The provisions of social relations. In: Rubin, Z. (Ed.), Doing unto Others. Prentice Hall, Englewood Cliffs, pp. 17–26.

Werts, C.E., Linn, R.L., 1970. Path analysis. psychological examples. Psychological Bulletin 74, 193–212.

Wills, T.A., 1985. Supportive functions of interpersonal relationships. In: Cohen, S., Syme, S.L. (Eds.), Social Support and Health. Academic Press, Orlando, pp. 61–82.

Wit, H. De, 1995. Cijfers en hun Achterliggende Realiteit. De MTMM-Kwaliteitsparameters op hun Kwaliteit onderzocht. Doctoral Dissertation, University of Chicago.

Wit, H. De, Billiet, J., 1995. The MTMM-design: back to the founding fathers. In: Saris, W.E., Münnich, A. (Eds.), The Multitrait–Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press, Budapest, pp. 39–60.