ELSEVIER

Social Networks 21 (1999) 111–130

SOCIAL NETWORKS

www.elsevier.com/locate/socnet

Evaluation of social network measurement instruments

Anuška Ferligoj *, Valentina Hlebec [1]

Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, Ljubljana 1001, Slovenia

Abstract

This paper evaluates the reliability and validity of network measurement instruments for measuring social support. The authors present and discuss the results from eight experiments which were designed to analyze the quality of four measurement scales: (1) binary, (2) categorical, (3) categorical with labels, and (4) line production, as well as two measurement techniques for listing alters (free recall and recognition). Reliability and validity were estimated by the true score multitrait–multimethod (MTMM) approach. Meta-analysis of factors affecting the reliability and the validity of network measurement was done by multiple classification analysis (MCA). The results show that the binary scale and the first presentation of measurement instruments are the least reliable. Surprisingly, the two data collection techniques (free recall and recognition) yield equally reliable data.

1. Introduction

The quality of survey data from complete social networks can be affected by many characteristics of the measurement instrument. Some of the factors that can affect response behavior, e.g., are (a) the wording of questions, (b) the response scale, (c) the data collection method, and (d) the context of the questions. The overall quality of measurement in a given network thus depends upon several factors whose impact on data quality has seldom been thoroughly analyzed and controlled. This state of affairs results from a lack of systematic research into the issues of data quality in the field of social network analysis. Research work on measurement issues mainly focuses on the questions of measurement validity, reliability, accuracy and measurement error (Wasserman and Faust, 1994, p. 56).

There are, however some studies which have focused mainly on the question of network data quality (Hammer, 1984; Sudman, 1985, 1988; Hlebec, 1993; Brewer and

* Corresponding author. Tel.: +38-6-61-1681-461; fax: +38-6-61-1685-330; E-mail: anuska.ferligoj@uni-lj.si

[1] E-mail: valentina.hlebec@guest.arnes.si.

0378-8733/99/$ - see front matter
PII: S0378-8733(99)00007-6

Webster, 1997). These papers are especially important for the experimental design used in this study, which focuses on the test–retest stability of complete networks. Each of these studies has compared two basic data collection techniques frequently used in the survey collection of network data: free recall and recognition. In all above mentioned studies, free recall and recognition yielded different measured egocentric networks. Previous research indicated little difference (when the networks were small) between the free recall and recognition techniques in the assessment of (a) important ties, (b) most recent contacts, and (c) most frequent contacts. The recognition technique gave much better results when networks were larger and ties were weaker. One aim of this study is to test whether the recognition data collection technique is more stable than that of free recall, since differences in reliability could significantly alter the interpretation of results.

Much work has been done on topics such as respondent accuracy (e.g., Killworth and Bernard, 1976; Bondonio, 1998; Casciaro, 1998), characteristics of the measured networks (e.g., Burt, 1984; Marsden, 1987; Wellman and Wortley, 1990), comparison of the measured networks using different network generators (e.g., Bernard et al., 1987, 1990; Campbell and Lee, 1991), and characteristics of the measured ties (e.g., Marsden and Campbell, 1984; Burt, 1986). Owing to the complexity of the data structure, there is still need for an extensive and systematic evaluation of survey measurement instruments in terms of test–retest reliability of measurement in the field of social network analysis. Nevertheless, in evaluating survey measurement instruments when measuring variables, there are some approaches which are applicable to social network analysis. The first such evaluation of survey measurement instruments, by two stage meta-analysis, was by Andrews (1984), who analyzed the quality of American and Canadian surveys. Together with Willem Saris and several other European social scientists, they established an international group on methodology and comparative survey research (IRMCS). Their extensive and fruitful work (Saris and van Meurs, 1990; Ferligoj et al., 1995; Saris and Münnich, 1995; Scherpenzeel, 1995) contributed substantially to knowledge about the quality of survey measurement instruments. Their results inspired the work of Ferligoj and Hlebec (1995, 1998), who first used the multitrait–multimethod (MTMM) approach to estimate the reliability of complete network measurements. They reported that the binary scale is the least reliable scale, at least when compared with an 11-point ordinal scale or a line drawing scale, regardless of the order of presentation. In the present study, two five-point ordinal scales have also been included.

Saris and Münnich (1995) and Ferligoj and Hlebec (1998) have also reported that factors, such as the order of repetition and the time between two successive presentations of the measurement instrument, can also have substantial impact on the quality estimates. Therefore, the order of presentation and the time between presentations have been included in the experimental design of this study. Finally, the content of the network name generators — social support — was selected on the basis of the characteristics of the experimental groups: eight classes of third year high school students.

This paper presents and discusses the results from eight experiments that were designed to analyze systematically the impact of different measurement characteristics on the reliability and validity of complete network data. In the first phase of this study,

estimates of test–retest reliability, validity and method effects are obtained for each set of relationships in each of eight classes, using the MTMM approach. In the second phase, the effects of the characteristics of the measurement instruments used in different classes are analyzed to explain the variability of the estimates for the reliability, validity and method effects. A secondary analysis of MTMM results is done by multiple classification analysis (MCA). The two-stage procedure described is similar to that used by Saris and Münnich (1995).

2. Method

2.1. Experimental design

In this study, data were collected regarding social support relationships among third year students in a high school in Gimnazija Bežigrad in Ljubljana, the capital city of Slovenia. On average, there were 31 students, aged 17, in each of eight classes. Four name generators [2] (traits) were used — exchange of study materials, exchange of information in the case of long-term illness, invitation to a birthday party, and discussion of important personal matters. These four traits served to measure four types [3] of support: (a) instrumental, (b) informational, (c) social companionship, and (d) emotional. The results of these experiments were presented to and discussed with respondents in June 1998. According to the students, the four network generators adequately measured the dimensions of social support. According to the respondents, exchange of study materials is done with students with good academic abilities, since they are likely to have the best notes. Exchange of information in the case of severe illness is done with others who are good friends and who perhaps live in the neighborhood. The discussion of important personal matters occurs between very close friends, while birthday parties were described as a form of socializing. They also commented that the existence of informal contacts, through extra-curricular activities, should be measured in order to encompass all the dimensions of social support.

All the name generators were repeated in two ways. First, respondents described whom they would ask for a particular exchange (original question), and second, who would ask them for a particular exchange (reversed question). Social support was thus measured in the direction of both giving and receiving. The paper and pencil interviews were carried out in January 1998.

To measure the strength of relationships, four measurement scales [4] were used: (1) a binary scale, (2) a five-point ordinal scale, (3) a five-point ordinal scale with labels, and (4) a line drawing scale. In each class, only three scales were applied in keeping with traditional MTMM design. Within each class, the ordering of three selected scales, the

[2] See question wording in Appendix A.
[3] See Cohen and Wills (1985) and Sarason et al. (1990).
[4] See the description of the scales in Appendix A.

Table 1
Experimental design
Labels: Scale (B — binary, C — categorical, CL — categorical with labels, L — line); scale ordering (1 — first, 2 — second, 3 — third); DCT — data collection technique (recognition — 1, free recall — 2); interview (1 — one repetition per interview, 2 — two repetitions per interview); date 1 — first interview; date 2 — second interview.

Class	Scale	Ordering	Interview	DCT	Date 1	Date 2
1	B (1)	1	2	1	5/1	12/1
1	C (2)	2	2	1	5/1	12/1
1	L (3)	3	1	1	5/1	12/1
2	B (1)	2	2	1	5/1	12/1
2	C (2)	3	2	1	5/1	12/1
2	CL (4)	1	1	1	5/1	12/1
3	B (1)	3	1	1	5/1	12/1
3	L (3)	1	2	1	5/1	12/1
3	CL (4)	2	2	1	5/1	12/1
4	C (2)	1	1	1	5/1	12/1
4	L (3)	2	2	1	5/1	12/1
4	CL (4)	3	2	1	5/1	12/1
5	B (1)	1	1	2	7/1	14/1
5	C (2)	2	2	2	7/1	14/1
5	L (3)	3	2	2	7/1	14/1
6	B (1)	2	2	2	6/1	13/1
6	C (2)	3	1	2	6/1	13/1
6	CL (4)	1	2	2	6/1	13/1
7	B (1)	1	2	2	6/1	13/1
7	L (3)	2	1	2	6/1	13/1
7	CL (4)	3	2	2	6/1	13/1
8	C (2)	1	2	2	6/1	13/1
8	L (3)	2	2	2	6/1	13/1
8	CL (4)	3	1	2	6/1	13/1

time intervals between three repetitions, and the data collection method were varied. The outline of experiment as designed is presented in Table 1.

In Fig. 1, the plan of the study is presented. First, the vectorization of each of 12 relational matrices (4 dimensions of social support $\times$ 3 measurement scales) for each class was performed. Then the reliability and the validity coefficients were estimated for each of 12 vectorized relational matrices within each of the eight classes. In the last phase, a meta-analysis was performed on the data matrix presented in Table 2. This data matrix is discussed later in Section 2.5.

2.2. Estimating reliability and validity of complete networks

Few procedures have been proposed for estimating the reliability of egocentric network measurements, and even fewer for estimating the reliability of complete network measurements. In this paper, the reliability coefficients that were designed to measure the reliability of variables, are also used on the complete network data. This is done by removing the diagonals and vectoring relational matrices.

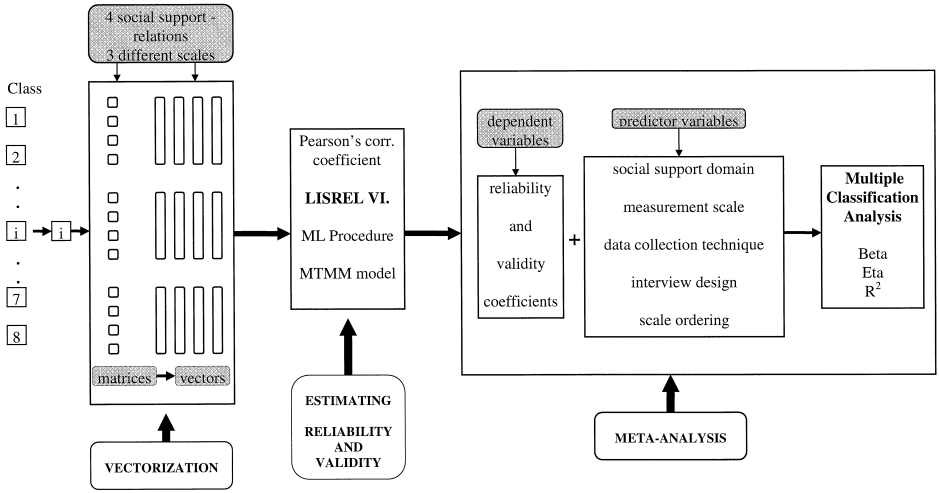

Fig. 1. Plan of the study.

In the following paper, a unit of the analysis is a dyad, and each relation (i.e., vector of dyads) is treated as a variable. There are other methods which could be used to assess

Table 2
A meta-analysis data matrix
Labels: Trait — social support dimension (1 — material support, 2 — informational support, 3 — social companionship, 4 — emotional support); scale (1 — binary, 2 — categorical, 3 — categorical with labels, 4 — line); ordering (1 — first, 2 — second, 3 — third); interview (1 — one repetition per interview, 2 — two repetitions per interview); DCT — data collection technique (recognition — 1, free recall — 2).

Class	Predictor variables					Dependent variables	
	Trait	Scale	Ordering	Interview	DCT	Reliability	Validity
1	1	1	1	2	1	0.763	0.993
1	1	2	2	2	1	0.855	0.986
1	1	3	3	1	1	0.865	0.935
1	2	1	1	2	1	0.757	0.993
1	2	2	2	2	1	0.898	0.987
1	2	3	3	1	1	0.904	0.940
1	3	1	1	2	1	0.801	0.993
1	3	2	2	2	1	0.939	0.988
1	3	3	3	1	1	0.836	0.930
1	4	1	1	2	1	0.786	0.993
1	4	2	2	2	1	0.918	0.988
1	4	3	3	1	1	0.849	0.932
2	1	1	2	2	1	0.710	0.973
2	1	2	3	2	1	0.911	0.989
–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–
8	4	4	3	2	2	0.959	0.981
8	4	4	3	2	2	0.795	0.950

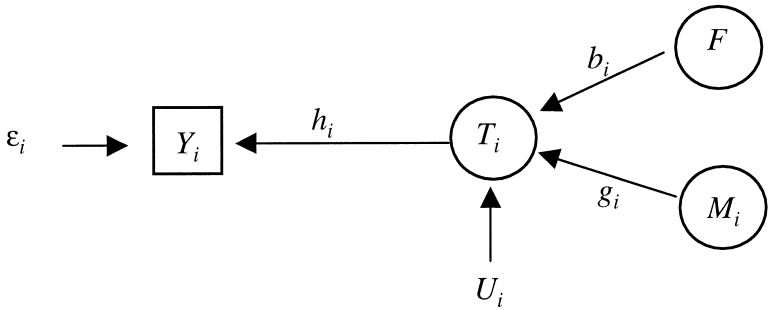

Fig. 2. True score measurement model.

reliability. Ferligoj and Hlebec (1995), e.g., used traditional approaches to estimate the reliability of a composite and single variable. In this paper, as well as in that of Ferligoj and Hlebec (1998), the authors focus exclusively on the reliability and validity of a single variable provided by the true score measurement model as conceived by Saris and Andrews, 1991, pp. 576–583).

The measurement model presented in Fig. 2 can be expressed by the following equations:

$$Y_i = h_i T_i + \varepsilon_i,$$
$$T_i = b_i F + g_i M_i + U_i,$$

where: Y_i is the response or observed variable corresponding to the question measured by the method i; T_i is the stable component when the same question is repeated under exactly the same conditions; ε_i is the random error in the observed variable Y_i; F is the unobserved variable of interest, assumed to be independent of the measurement procedure used; M_i is a method-specific component; U_i is the unique component of the true score T_i.

In this model it is assumed that:

$$E(\varepsilon_i) = 0; \quad E(U_i) = 0; \quad \mathrm{cov}(F, U_i) = 0; \quad \mathrm{cov}(M_i, U_i) = 0;$$
$$\mathrm{cov}(M_i, \varepsilon_i) = 0; \quad \mathrm{cov}(F, \varepsilon_i) = 0; \quad \mathrm{cov}(U_i, \varepsilon_i) = 0; \quad \mathrm{cov}(F_i, M_i) = 0;$$
$$\mathrm{var}(U) = 0.$$

In this measurement model, reliability is defined as the proportion of the variance in Y_i that remains stable across repetitions of the same measure, or:

$$\text{reliability} = \frac{\mathrm{var}(T_i)}{\mathrm{var}(Y_i)} = h_i^2.$$

Validity [5] is defined as the percentage of the variance of the true score explained by the variable of interest, or:

$$\text{validity} = b_i^2.$$

[5] These are not the only possible definitions of reliability and validity (see Saris and Andrews, 1991, pp. 581–582).

Invalidity $(1 - b_i^2)$ can be interpreted as method variance (g_i^2), if $U_i = 0$. Otherwise, invalidity is defined as follows:

$$\text{invalidity} = g_i^2 + \text{var}(U_i).$$

In this model (using one measurement), the reliability, validity and invalidity coefficients cannot be estimated. Therefore, several different approaches with repeated measurements were suggested. In this paper, the authors use the true score MTMM approach proposed by Saris and Andrews (1991) to assess the coefficients. To estimate the reliability and the validity coefficients, four survey questions (traits) have to be repeated at least on three occasions, each time with a different measurement scale.

The analysis of the MTMM model was based on a matrix of Pearson's correlation coefficients. The validity and the reliability coefficients were obtained by the ML procedure in LISREL VI program (Jöreskog and Sörbom, 1986), using the true score MTMM model presented in Fig. 3. For each class, two MTMM matrices were constructed; one for the original four questions, and the other for the reversed four questions measuring social support, i.e., giving and receiving.

2.3. The dependent variables in meta-analysis: quality estimates

The reliability and validity coefficients from the MTMM model were used as the dependent variables in a meta-analysis.

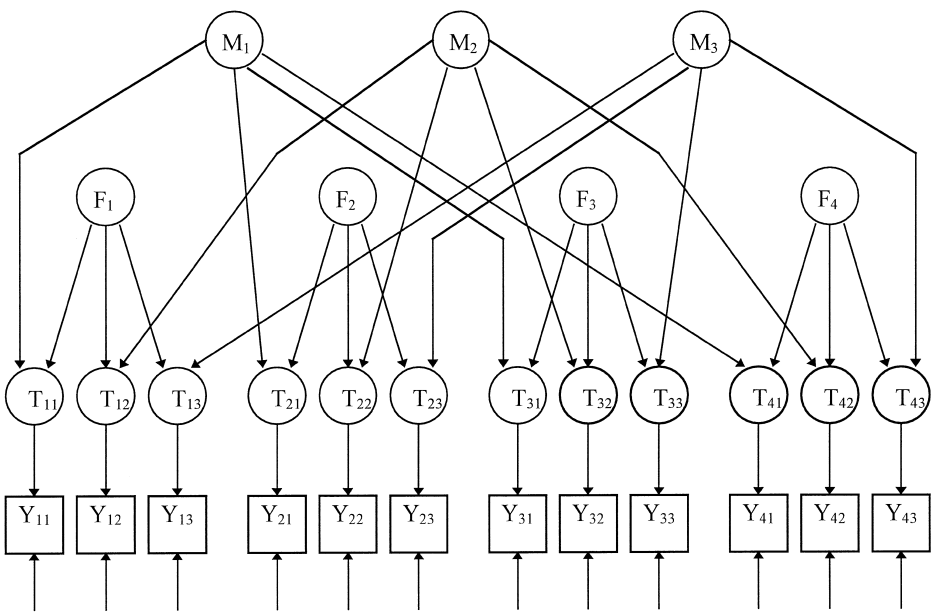

Fig. 3. The MTMM true score model presented in path diagram.

2.4. The predictor variables in meta-analysis: instrument characteristics

The quality estimates obtained in the first two phases of the study are the dependent variables within the meta-analysis. The predictor variables in the third phase are the characteristics of the measurement instruments. Each measurement instrument is characterized by the type of social support measured, its response scale, and the characteristics of the experiment in which it was included. These characteristics are presented below.

2.4.1. Social support

The most important part of a network generator is its content. Within the present study, the content of the network generators is social support among a group of high school students. Dimensions of social support were adjusted to the age of the respondents (17) and to the environment of the survey (the classroom). The dimension of instrumental support involved the provision of material resources, financial aid and needed services. The exchange of study materials (books, notes, etc.) provides a typical example of instrumental support. Informational support included help in defining and coping with problematic events. Respondents were presented with a hypothetical situation in which they would be absent from school owing to an illness during the most important grading month (May); they were then asked to list the schoolmates who would be ready to provide them with the missing school work.

The third dimension of social support in this study was social companionship, a category which was intended to tap the need for affiliation and contact with others. As we did not know in advance which leisure activities our students shared, we chose participation in birthday parties as an activity that all were likely to be involved in at some time during the year. Emotional support was measured with a traditional network generator which listed the names of the persons with whom respondents discussed the matters of importance.

Of course, these four dimensions can only be completely distinguished from one another in theory. However, it is assumed that the four network generators measure the attributed dimension of social support in general. Also some overlap in dimensions is expected, an effect which will be reflected in higher values for some of the correlation coefficients between the network generators.

2.4.2. Response scale

Four different response scales were used to measure the strength of support relations. The binary scale is the most frequently used scale in social network analysis. In this study, the binary scale is compared with two ordinal scales (the five-point ordinal scale without labels, and the five-point ordinal scale with labels) and a line drawing scale.

2.4.3. Data collection techniques

In most, if not all, data collection modes used to analyze social networks, two general approaches can be distinguished. In the first approach — the recognition method — respondents are presented with a list of all members of the group, and are asked to estimate the strength of their relationships with each listed person. In the second approach — the free recall format — respondents are not offered any help in selecting

the names of significant others. In this study, both approaches were used in order to see whether the quality of data is significantly better when respondents are provided with a roster, a technique which is frequently assumed to facilitate responses.

2.4.4. Position in the questionnaire

The position of a measurement instrument in the questionnaire is defined by the ordering of the three scales. Within each class, the presentation of the scales varied, as can be seen in Table 1.

2.4.5. MTMM design: time between repetitions in the same interview

As each question (social support dimension) has to be measured three times in order to estimate the MTMM model, we have to make sure that the time intervals between presentations of the same question (with different measurement scales) are long enough to prevent any memory effect. As shown in the previous experiments on the quality of survey measurement instruments (Saris and van Meurs, 1990), respondents can remember their answers [6] to the same question when there is only a short time (less than 20 min) between presentations. The shorter the time between two repeated measures, the higher the correlation between them, and consequently, the higher the estimates of reliability and validity for these measures.

In the present design, the three measures of a social support dimension are presented to respondents in such way that two of them [7] are always positioned within one questionnaire, while the third is presented on a separate occasion. This makes the second measurement instrument especially vulnerable to memory effects since the first and second measures are included in the same interview. The same happens with the third measure when the second and the third presentations appear in the same questionnaire. As the interviews lasted on average 40 min, and since the set of 75 questions on interpersonal relations was positioned between the two measuring occasions, we expect that the memory effect will not affect the estimates of validity and reliability. If a consistent pattern appears in results, indicating higher estimates of validity and reliability, due to the position of measures within the questionnaire, then memory effect will have to be considered as a possible explanation.

2.5. Meta-analysis

In the first stage, 192 estimates of reliability and validity were obtained from the MTMM analyses. Ninety-six estimates were obtained for original questions (4 questions

[6] It should be taken into consideration that these results stand for attitudes and opinions.

[7] The first and second measures or the second and third measures are presented within the same interview at an average of 25 min apart. The separate interview involving only a single measurement took place either 1 week before or 1 week after the joint interview.

about social support × 3 measurement scales × 8 classes) and 96 estimates for the reversed questions. The two databases for the meta-analysis thus comprise 96 cases each. The meta data matrix for the network questions is presented in Table 2. Following the example given in other meta-analyses for explaining the effects on the data quality estimates of different characteristics in the measurement instruments, (Scherpenzeel, 1995), MCA was chosen as the meta-analysis technique. The multivariate (MCA) coefficients indicate how much the validity and reliability estimates deviate from the mean as a result of a given characteristic of the measurement instrument, while controlling for the effects of all other characteristics of the measurement instrument. Two measures of the overall effect of each predictor are obtained, and in addition, the MCA Eta and MCA Beta. The MCA Eta coefficient measures the strength of the bivariate relationship between a quality estimate and a predictor. MCA Beta coefficients, on the other hand, measure the strength of the relationship, controlled for the other predictor variables in the model. The rank order of the Betas indicates the relative importance of the predictor variables in their explanation of the dependent variable. Finally, the multiple R^2, indicating the total proportion of variance explained by all predictors together, is estimated.

3. Results

3.1. Mean levels of data quality

In Table 3, summary statistics for the validity and the reliability coefficients over eight classes are presented. Within each class, 12 reliability and 12 validity estimates were obtained for both the original and the reversed measures. The overall mean reliability of 12 reliability coefficients for the original questions is 0.881, and 0.882 for

Table 3
Mean levels and variation of validity and reliability coefficients

Class	Original				Reversed			
	Validity		Reliability		Validity		Reliability	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	0.972	0.028	0.848	0.061	0.957	0.031	0.856	0.068
2	0.969	0.023	0.862	0.071	0.965	0.010	0.879	0.056
3	0.975	0.019	0.887	0.067	0.971	0.017	0.887	0.056
4	0.985	0.013	0.909	0.044	0.984	0.010	0.914	0.033
5	0.987	0.009	0.898	0.057	0.970	0.022	0.888	0.048
6	0.989	0.007	0.891	0.033	0.988	0.007	0.883	0.044
7	0.988	0.010	0.886	0.060	0.984	0.006	0.883	0.071
8	0.981	0.018	0.867	0.054	0.976	0.012	0.862	0.061
Overall	0.981	0.018	0.881	0.058	0.974	0.019	0.882	0.056

the reversed. The mean overall validity coefficient is 0.981 for original questions, and 0.974 for the reversed.

As is evident in Table 3, the validity estimates are much higher than those for reliability: both the overall mean and the class means are higher in the case of validity. It seems surprising to get such high values for the validity. For these students, preferences in terms of social support are apparently clear and stable, even when measured by different measurement instruments.

The reliability estimates varied much more, both between and within classes, than did the validity estimates. Because of the very high mean values for the validity and small variability both between and within classes, the validity estimates are not included within the meta-analysis. In the second stage of the study, only the meta-analysis for the reliability estimates is presented.

3.2. Effects of instrument characteristics

In this section, the results from the meta-analyses are presented. The dependent variables in meta-analyses are the reliability coefficients, and the predictor variables are

Table 4
Predictive power and effects for the social support domain, response scale, data collection method, and interview design on the reliability estimates

	N measures	Original questions (reliability coefficient, mean = 0.881)			Reversed questions (reliability coefficient, mean = 0.882)		
		Bivariate	Multivariate		Bivariate	Multivariate	
		Eta	Beta	Deviation	Eta	Beta	Deviation
Social support							
Material	24			−0.019			−0.014
Illness	24			0.011			0.016
Birthday party	24			−0.006			−0.000
Discussion	24	0.232	0.232	0.014	0.190	0.190	−0.001
Response scale							
Binary scale	24			−0.042			−0.035
Five-point category scale	24			0.022			0.002
Line drawing scale	24			−0.002			0.009
Five-point category scale (labels)	24	0.453	0.453	0.022	0.390	0.390	0.024
Data collection							
Recognition	48			−0.005			0.003
Free recall	48	0.082	0.082	0.005	0.046	0.046	−0.003
Interview design							
One measure	32			−0.027			−0.019
Two measures	64	0.335	0.335	0.014	0.245	0.245	0.010
Multiple R^2			0.378			0.250	

the characteristics of the measurement instruments. In Tables 4 and 5, the reliability estimates for the original questions are presented in the middle columns, and the reliability estimates for reversed questions are shown in the right-hand three columns. The two measures of the predictive power of the instrument characteristics, MCA Eta and Beta, are given for both the original and reversed question wording. The effects of the measurement characteristics (predictors) are presented as deviations from the mean. In the last row of the tables, the R^2 is given for both the reliability of the original and the reliability of the reversed questions estimates.

The results from four separate meta-analyses are presented in Tables 4 and 5. As mentioned before, there are five predictor variables that describe the characteristics of measurement instruments. Two variables (position in the questionnaire and time between repetitions in the same interview) overlap to some degree. The second scale is always presented interchangeably with the first or third scale. This results in a number of empty cells within the MCA design; thus, higher-order interactions between predictor variables cannot be estimated. Therefore, it was decided to obtain estimates for possible interactions between the predictor variables within two separate meta-analyses. In the first meta-analysis (Table 4), four measurement instrument characteristics are included: social support traits, response scales, data collection technique, and interview design. In the

Table 5
Predictive power and effects of social support domain, response scale and presentation ordering on quality estimates

	N measures	Original questions (reliability coefficient, mean = 0.881)			Reversed questions (reliability coefficient, mean = 0.882)		
		Bivariate	Multivariate		Bivariate	Multivariate	
		Eta	Beta	Deviation	Eta	Beta	Deviation
Social support							
Material	24			−0.019			−0.014
Illness	24			0.011			0.016
Birthday party	24			−0.006			−0.000
Discussion	24	0.232	0.232	0.014	0.190	0.190	−0.001
Response scale							
Binary scale	24			−0.042			−0.035
Five-point category scale	24			0.022			0.002
Line drawing scale	24			−0.002			0.009
Five-point category scale (labels)	24	0.453	0.453	0.022	0.390	0.390	0.024
Ordering							
First	32			−0.029			−0.030
Second	32			0.015			0.011
Third	32	0.350	0.350	0.014	0.382	0.382	0.019
Multiple R^2			0.382			0.334	

second analysis (Table 5), three predictor variables were included: social support traits, response scales, and method ordering (position in the questionnaire). The results from the first meta-analysis show that data collection method does not explain much of the variance in the reliability estimates. Thus, this characteristic was excluded from the second meta-analysis.

3.2.1. Predictive power

The multiple R^2 given in the last row of Table 4 shows that these four predictors explain 38% of the variance in the reliability estimates of the original questions, and 25% of the variance in the reliability estimates of the reversed questions. There are no differences between bivariate Eta and multivariate Beta coefficients, indicating that bivariate relationships between each of the predictors and quality estimates are not suppressed by their relationship with other predictors.

The largest bivariate and multivariate effects for the reliability estimates of both the original and the reversed questions occur in response scale, interview design and social support domain (Etas and Betas ranging from 0.453 to 0.232 for the original questions and from 0.39 to 0.19 for the reversed questions). In contrast, data collection methods do not seem to have much effect on the reliability estimates. When comparing the original and the reversed questions, one can see that even though the reliability estimates of the reversed questions produce lower variances, the relative power of predictor variables is nevertheless distributed in the same way for both types of questions.

The multiple R^2 in Table 5 shows that three variables explain 38% of the variance in the reliability estimates for the original questions, and only a little less for the reversed questions (33%). The relative predictive power for both the social support domain and the response scale is the same as in the first meta-analysis. The bivariate Eta and multivariate Beta for scale ordering are 0.350 for the original questions and 0.382 for the reversed questions. This result indicates that scale ordering is the second most important explanatory variable.

3.2.2. Specific effects of instrument characteristics

Next to the columns with Beta coefficients in Tables 4 and 5, there are the columns with the deviations, which indicate how much the reliability estimates change as a result of different characteristics, while controlling for the effects of all other characteristics. Thus, the average reliability for the measure of social support among high school students, measured in the exchange of study materials (original question) is 0.881 (the mean reliability) − 0.019 (the effect of this characteristic) = 0.862, controlling for all other effects. Similarly, the average reliability for social support among students, as measured in terms of material exchange (reversed question), is 0.868 (0.882–0.014).

The domain of social support has a substantial effect on reliability estimates. Among the measures of social support, the exchange of materials is the least reliable measure for both the original questions and the reversed questions. One possible reason for these lower reliabilities is that this question was always presented first to the respondents. It is possible that this question was answered less reliably simply because at the beginning of interviews, students were still learning how to answer network questions. In line with this explanation, it could be expected that the reliability of subsequent questions would

increase because of the question ordering. From the results, it can be seen that this cannot be the only explanation for reliability estimates, since the third question (about the birthday party) has a lower reliability than the preceding question (help in the case of illness).

Another explanation as to why the instrumental dimension of social support is less reliable could be its importance to the respondents, in comparison to the other included dimensions of social support. A third possible explanation is the different density of reported complete networks for the four dimensions of social support. The smallest numbers of dyads [8] were reported for informational support (help in the case of illness), and emotional support (discussion of important matters) across all eight classes, regardless of the data collection technique. Sparse networks were measured more reliably than dense ones when the total number of network members was around 30.

When comparing the original and the reversed questions, we need to have in mind that the reversed questions measure the support respondents expect to be asked for from other students in the classroom. Somewhat higher reliability estimates in the case of the reversed questions do not necessarily mean that these questions provide better measures of social support than the original questions. One can conclude only that the perception of social support expected from respondents is as stable as the social support needed by respondents.

The response scale is the most important predictor of the reliability estimates. It appears that the binary scale is the least reliable. This is in agreement with previous results (Ferligoj and Hlebec, 1995, 1998). The five-point category scale is the most reliable, regardless of the labels used in the original questions. For the reversed questions, the scale with labels seems to be the best. The line drawing scale falls somewhere in the middle. A closer look at higher-order interactions [9] of explanatory factors in meta-analysis shows that one of the second-order interactions was significant for both original and reversed questions. An interaction of data collection technique and measurement scale had to be considered as part of any explanation of results. It appears that the binary scale is less reliable compared to the other three scales, only when the recognition data collection technique was used. The percentage of reported dyads when the binary scale was used with the recognition technique was a half to two-thirds of the dyads reported by the other three scales. Cross-tabulations of networks measured by the binary scale with networks measured by other scales showed that, when one is using the binary scales stronger ties tend to be reported. Recognition data collection enhances the reporting of both strong and weak ties, especially when the measurement scales employed can also measure the strength of ties.

In previous research (Ferligoj and Hlebec, 1995, 1998), the binary scale was compared to both the 11-point ordinal scale and the line drawing scale. In this

[8] The percentage of reported dyads for informational and emotional support was approximately one-third to one-half of the percentage of dyads reported for instrumental support. The percentage of reported dyads varied from 17 to 85 for the recognition data collection technique, and from 7 to 34 for free call (60% of dyads were reported for the social companionship dimension in one classroom where free recall was used).

[9] Complete report tables for meta-analyses are too space-consuming to be reported and can be obtained from the authors.

combination, both the ordinal scale and the line drawing scale were equally good and much better than the binary scale. Recognition data collection was also used. From these findings, it cannot simply be concluded that the five-point ordinal scale is the best choice for measuring the strength of social support relations, and that the line drawing scale is the second best choice. The reason lies in MTMM design: when two of three scales are too similar; the reliability estimates are overestimated (De Wit and Billiet, 1995).

To test the possibility that the mean values for the reliability estimates are too large, the two five-point scales were compared across different experiments. When compared with the two presentations of the five-point scale in one experiment, the single presentation of the five-point scale gave higher reliability estimates, which is quite contrary to our hypothesis. A closer look at the combination of the two five-point scales within the one experiment showed that the ordering of presentations played a substantial role in determining the reliability. The first scale presented to respondents had a lower reliability, while the third had the highest reliability across all experiments.

Therefore, a five-point ordinal scale appears optimal for measuring the strength of social support relations, or possibly some other scale [10] with which respondents are comfortable. Within the Slovene school population, the five-point ordinal scale may also be the best because this is a common grading system within schools. Hence, this recommendation cannot be generalized to other populations without further research.

The data collection method does not have much effect on the quality estimates. It appears that for the type of relationship where respondents know each other very well, free recall functions just as well as the method where the full list of members is presented, when stability of measurement is in question. This is true for social support and also for some other types of relationships (see Hammer, 1984; Sudman, 1985, 1988). Since there appears to be an interaction between the data collection technique and the measurement scale, the data collection methods being used should not be taken as interchangeable. When one uses the recognition data collection technique, more ties and weaker ties are also reported in contrast to reports from the free recall technique.

Interview design is the second most powerful predictor of reliability estimates. If a question is presented alone [11] in a questionnaire, the degree of reliability is substantially lower. This finding supports the hypothesis regarding the increase in the reliability coefficients when two measures [12] of the same trait are presented in the same interview.

[10] In discussions with respondents, they all agreed that the binary scale is quite crude. When asked for the most appropriate scale, they disagreed substantially with one another. Some preferred the line drawing scale because it did not use numbers. Others preferred the five-point ordinal scale with the labels (labels without numbers). Still, others liked the 11-point ordinal scale without labels. It seems that different scales should be used for different populations.

[11] The time interval between the interview with a single presentation and that with two presentations was 1 week.

[12] When the results of the study were discussed with the respondents, some reported that they could remember all their answers. Other respondents claimed that they could not remember their previous answers, and that the number of reported exchange partners may play a substantial role in the memory effect. It is also possible that the memory effect is partially a function of the size of an egocentric network. However, this hypothesis cannot be tested in this design since the unit of analysis is a dyad, not a respondent.

Concomitantly with the reasoning about the effect of response scale on the reliability estimates, one could then expect a compound effect involving interview design and presentation ordering. Since the experimental design did not allow us to include interview design and presentation ordering in the same meta-analysis as the predictors, a separate meta-analysis was done where only these two variables were included. As hypothesized, these two variables interact in such a way that the first presented measure had the lowest or the second lowest reliability when presented alone or together with another measure (0.849 and 0.856 for original questions; 0.857 and 0.846 for reversed questions). The third presented measure had the highest reliability when presented together with another measure (0.931 for original questions and 0.933 for reversed questions), and the third highest values when presented alone (0.858 for original questions and 0.867 for reversed questions). The measure that is presented second, and always presented with another measure, has the second highest reliability (0.896 for original questions and 0.893 for reversed questions).

In most cases, network generators are presented at only one point in the questionnaires. In the presentation of network generators, it is important that the respondents are given prior information on how to complete the questions. Among the interviewed students, only those in one class were familiar with network generators.

Since the effects on reliability estimates of both the social support domain and the response scale are similar in both meta-analyses, only the effects for the order of presentation will be commented upon. It is clear that the first presented measure has the lowest reliability estimates, and that the third has the highest reliability.

4. Conclusions

The results from four meta-analyses show that the domain of social support, as measured by the binary scale, is the least reliable when the recognition data collection technique is used. In contrast, when free recall is used, a smaller number of dyads is reported and all scales are equally reliable. It seems that when a full list of membership is available, it should be used in any measurement procedure to simplify the reporting task for respondents, and to increase the number of reported ties. It appears that a measurement scale which measures the strength of ties at the same time as network membership will produce the largest number of ties of different strength. When a researcher is interested only in strong ties, the free recall data collection technique may also be used. In this case, the number of ties reported by the binary scale is equal to the number of ties reported by other scales. Since other scales simultaneously provide information about the strength of ties, a researcher should decide whether to measure the strength of ties with a network generator or with characteristics of reported ties, such as intimacy, duration or frequency of contacts (Marsden and Campbell, 1984). In any quest for the best measurement scale, all characteristics of the group being analyzed should be considered. The most reliable results were obtained by the ordinal scales. For Slovenian high school students, a five-point ordinal scale seems to be the best. Additional experiments need to be done for different populations and cultures before these findings can be generalized.

Reliability is lower when a single measure is used in interviews. Since repeated measurements are rare in survey measurement, the quality of measurement can be improved by introducing an informative example prior to the most important measures. A follow-up measurement can also be used to complete answers for the most important relations.

In a comparison of the original to the reversed questions, the same characteristics of measurement instruments were found to influence reliability, though less variance was explained in the case of the inverted questions. It appears that perceptions of received and given support are equally stable. The results also suggest that the original questions and the reversed questions should not be used interchangeably. In this study, the original and reversed questions are treated as separate groups and no direct comparison is performed.

Different dimensions of social support were not equally reliable. The stability of reported support networks seems to depend on the characteristics of ties which provide specific types of social support. Stronger ties are more stable than weaker ties. In our case, stronger ties provided informational and emotional support, whereas weaker ties provided instrumental support and companionship. Similar effects can be expected for ties that provide several different types of support, as opposed to ties that are specialized. Since the contents network generators were adjusted to a specific population, further experiments for other network generators should be done to generalize these findings.

Additional experiments on different populations, different social network domains, and in different countries should be performed to further test the quality of the measurement instruments used. Special attention should be paid to the characteristics of ties, and features such as duration and frequency of contacts, which can influence the stability of measurement and which vary in the general population.

An important drawback of our approach is that analysis is done on the level of networks. The unit of analysis is a dyad and not a respondent. A combined approach should be found to attach the characteristics of respondent and characteristics of its network to the measurement quality of reported ties in complete networks.

Similar studies can and should be done also to estimate the quality of survey measurement instruments for egocentric networks.

Acknowledgements

The authors are grateful to the members of IRMCS for their critical comments and especially to Willem Saris and Brendan Bunting for their useful suggestions on the text.

Appendix A. Wording of the network generators

Altogether, there were eight different forms of network generators with varying scales and data collection techniques. The questionnaire had four sections.

(1) Network generator measuring instrumental support with a binary scale and with the recognition data collection technique (original question).

You have known your classmates for some time now. It sometimes happens that you cannot take courses for various reasons. From which of your classmates would you borrow study materials? Indicate your answers on the list below in the following way. Mark 1 in the box next to a person's name if you would borrow study material from her/him. Mark 0 in the box next to a person's name if you would not borrow study materials from her/him.

Reversed question: Which of your classmates would ask you to lend your study materials? (Instructions for respondents were the same as for the original question.)

(2) Network generator measuring informational support with an ordinal scale without labels and with the recognition data collection technique (original question).

Suppose you were ill at the beginning of May, and you had to stay in hospital for a month. Which of your classmates would you ask to obtain information about important study assignments? Indicate your answers on the list below in the following way. Select a number from 0 to 4 to indicate how likely you would be to ask your classmates for help. Mark 4 in the box next to a person's name if you would certainly ask for help from her/him. Mark 0 in the box next to a person's name if you would not ask for help from her/him. The more likely it is that you would ask for help from a person, the larger the number should be.

Reversed question: Which of your classmates would ask you to obtain study information in the case of a long absence? (Instructions for respondents were the same as for the original question.)

(3) Network generator measuring companionship with a line production scale and with the free recall data collection technique (original question).

Suppose your birthday falls next week, and you want to give a birthday party. Which of your classmates would you invite? Indicate your answers on the list below in the following way. List the names of any classmates that you would invite to your birthday party; for each listed person, indicate by the length of the line how likely you would be to invite her/him. The longer the line, the more likely you would be to invite that person.

Reversed question: Which of your classmates would invite you to her/his birthday party? (Instructions for respondents were the same as for the original question.)

(4) Network generator measuring emotional help with an ordinal scale with labels and with the free recall data collection technique (original question).

With which of your classmates would you discuss important things? Indicate your answers on the list below in the following way. List the names of any classmates with whom you would discuss important matters; for each listed person, use a number from 0 to 4 to indicate how likely you would be to discuss your important personal matters with her/him. Mark 4 if it is certain that you would discuss personal matters with her/him. Mark 3 if it is very likely that you would discuss personal matters with her/him. Mark 2 if it is likely that you would discuss personal matters with her/him. Mark 1 if it is not likely that you would discuss personal matters with her/him. Mark 0 if it is certain that you would not discuss personal matters with her/him.

Reversed question: Which of your classmates would discuss important personal matters with you? (Instructions for respondents were the same as for the original question.)

References

Andrews, F.M., 1984. Construct validity and error components of survey measures: a structural modeling approach. Public Opinion Quarterly 48, 409–422.

Bernard, H.R., Shelley, G.A., Killworth, P.D., 1987. How much of a network does the GSS and RSW dredge up?. Social Networks 12, 179–215.

Bernard, H.R., Johnsen, E.C., Killworth, P.D., McCarty, C., Shelley, G.A., Robinson, S., 1990. Comparing four different methods for measuring personal social networks. Social Networks 12, 179–215.

Bondonio, D., 1998. Predictors of accuracy in perceiving informal social network. Social Networks 20, 301–330.

Brewer, D.D., Webster, C.M., 1997. Patterns in the recall and recognition of friends by residents of a university dormitory. Paper presented at the International Sunbelt Social Networks Conference, February, San Diego, USA.

Burt, R.S., 1984. Network items and the general social survey. Social Networks 6, 149–174.

Burt, R.S., 1986. A note on sociometric order in the general social survey network data. Social Networks 8, 149–174.

Campbell, K.E., Lee, B.A., 1991. Name generators in surveys of personal networks. Social Networks 12, 203–221.

Casciaro, T., 1998. Seeing things clearly: social structure, personality, and accuracy in social network perception. Social Networks 20, 331–351.

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

Ferligoj, A., Hlebec, V., 1995. Reliability of network measurements. In: Ferligoj, A., Kramberger, T. (Eds.), Contributions to Methodology and Statistics. Metodološki zvezki, Vol. 10. 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., Leskosek, K., Kogovsek, T., 1995. Zanesljivost in Veljavnost Merjenja. FDV, Ljubljana (in Slovene).

Hammer, M., 1984. Explorations into the meaning of social network interview data. Social Networks 6, 341–371.

Hlebec, V., 1993. Recall versus recognition: comparison of two alternative procedures for collecting social network data. In: Ferligoj, A., Kramberger, T. (Eds.), Developments in Statistics and Methodology. Metodološki zvezki, Vol. 9. FDV, Ljubljana, pp. 121–128.

Jöreskog, K.G., Sörbom, D., 1986. LISREL VI: analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods. University of Uppsala, Uppsala.

Killworth, P.D., Bernard, R.H., 1976. Informant accuracy in social network data. Human Organization 35, 269–286.

Marsden, P.V., 1987. Core discussion networks of Americans. American Sociological Review 52, 122–131.

Marsden, P.V., Campbell, K.E., 1984. Measuring tie strength. Social Forces 63, 482–501.

Sarason, B.R., Sarason, I.G., Pierce, G.R., 1990. Social Support: An Interactional View. Wiley, New York.

Saris, W.E., Andrews, F.M., 1991. Evaluation of measurement instruments using a structural modeling approach. In: Biemer, P.P., et al. (Eds.), Measurement Errors in Surveys. Wiley, New York, pp. 575–599.

Saris, W.E., Münnich, A., 1995. The Multitrait–Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press, Budapest.

Saris, W.E., van Meurs, A., 1990. Memory effects in MTMM studies. In: Saris, W.E., van Meurs, A. (Eds.), Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies. North-Holland, Amsterdam, pp. 134–147.

Scherpenzeel, A., 1995. A Question of Quality: Evaluating Survey Questions by Multitrait–Multimethod Studies. Royal PTT Netherlands, Amsterdam.

Sudman, S., 1985. Experiments in the measurement of the size of social networks. Social Networks 7, 127–151.

Sudman, S., 1988. Experiments in measuring neighbour and relative social networks. Social Networks 10, 93–108.

Wasserman, S., Faust, K., 1994. Social Network Analysis — Methods and Applications. Cambridge Univ. Press, Cambridge.

Wellman, B., Wortley, S., 1990. Different strokes from different folks: community ties and social support. American Journal of Sociology 3, 558–588.