Available online at www.sciencedirect.com
SCIENCE DIRECT®
Social Networks 28 (2006) 209–231
SOCIAL NETWORKS
www.elsevier.com/locate/socnet

# Reliability and validity of egocentered network data collected via web A meta-analysis of multilevel multitrait multimethod studies

Lluís Coromina *, Germà Coenders

*Department of Economics, University of Girona, Faculty of Economics, Campus Montilivi, 17071 Girona, Spain*

**Abstract**

Our goal in this article is to assess reliability and validity of egocentered network data collected through web surveys using multilevel confirmatory factor analysis under the multitrait multimethod approach.

In this study, we analyze a questionnaire of social support of Ph.D. students in three European countries. The traits used are the frequency of social contact questions. The methods used are web survey design variants.

We consider egocentered network data as hierarchical; therefore, a multilevel analysis is required. Within and between-ego reliabilities and validities are defined and interpreted.

Afterwards, we proceed to a meta-analysis of the results of the three countries where within and between-ego validities and reliabilities are predicted from survey design variables which have to do with question order (by questions or by alters), response category labels (end labels or all labels) and lay-out of the questionnaire (graphical display or plain text). Results show that question order by questions, all-labeled response categories and a graphical display lay-out with images lead to a better data quality. Our basic approach consisting on multilevel and meta-analysis can be applied to evaluate the quality of any type of egocentered network questionnaire, regardless of the data collection mode.


*Keywords:* Meta-analysis; Multitrait multimethod (MTMM) model; Egocentered network; Multilevel model

* Corresponding author. Tel.: +34 972418815; fax: +34 972418032.
*E-mail address:* lluis.coromina@udg.es (L. Coromina).

0378-8733/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2005.07.006

## 1. Introduction

Our aim in this article is to assess the quality of measurements of egocentered social networks obtained by means of web questionnaires. With this purpose, we do a meta-analysis of reliability and validity estimates obtained with multilevel factor analysis models.

Egocentered networks consist of a single individual (usually called ego) with one or more relations defined between him/her and a number of other individuals, called alters. Another common type of network is the complete network, which consists of a group of individuals with one or more relations defined among all of them.

The composition of egocentered networks is elicited by means of the so-called name generator questions. After that, several characteristics (variables) are usually measured which describe the 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 his/her alter in the network (e.g., partner, boss, co-worker), feelings of closeness or importance, frequency of advice, collaboration, social support and so on. Since the data about the characteristics of ties are used as important explanatory variables in social network research and are, moreover, usually reported only by the ego, it is very important to know to what extent these data are reliable and valid. The fact that questions about ties are frequently reported as being sensitive and placing a high response burden (de Lange et al., 2004) makes this quality assessment even more necessary.

A considerable number of authors have evaluated the methodological characteristics of various methods for collecting ego-centered network data. There are studies comparing the characteristics of the measured networks (e.g., Burt, 1984; Marsden, 1987), and evaluating the characteristics of the measured ties (e.g., Marsden and Campbell, 1984). Several researches emphasize accuracy of social network questions and data collection in order to obtain significant results (Bernard et al., 1990; Bondonio, 1998; Brewer, 2000; Feld and Carter, 2002; Sudman, 1985).

There are also studies that have predominantly focused on the reliability and validity of measured networks and data collection methods used (Hoffmeyer-Zlotnik, 1990; Marsden, 1993; Ferligoj and Hlebec, 1999; Kogovšek et al., 2002; Hlebec and Ferligoj, 2002). Most of these studies used only the face-to-face data collection mode. However, Kogovšek et al. (2002) compared face-to-face and telephone surveys.

Web surveys have already proved to be a valid and reliable survey method for classic survey questionnaires (Couper, 2000; Dillman, 2000; Couper et al., 2001; Vehovar et al., 2002) for populations that are motivated to respond the questionnaire and have internet access. However, they have very rarely been used for collecting data on ego-centered networks. As exceptions, Marin (2002) and Lozar Manfreda et al. (2004) did use a web questionnaire to collect ego-centered network data. However, the first of these studies did not evaluate the quality of data with respect to the web-based administration and the second did only so with respect to name generator questions and using indicators such as the percentage of completed interviews, not reliability and validity.

We do not mean that web surveys are a panacea for collecting network data. For studies of the general population, non-response and coverage errors will likely be high; due to the non-universal internet access. However, many social network studies are focused on members of specific organizations, not on a general population. In these cases, the use of web surveys

may well be considered by researchers. Web surveys also have some networking similarities with the other self-administered data collection modes, so that some of the findings in this article may be of a rather more general applicability.

In this study, we analyze data from a web questionnaire designed by the International Network on Social Capital and Performance (INSOC, http://srcvserv.ugent.be/insoc/insoc.htm) research group to predict the performance of Ph.D. students in different European countries. We designed the survey for Ph.D. students and their supervisors in the universities of Girona (Spain), Ljubljana (Slovenia) and Ghent (Belgium). The traits used in the article are the frequency of scientific advice in work problems, the frequency of collaboration in research, the frequency of asking for crucial information and the frequency of social activities outside the work context. The methods used are two different waves of a panel web survey using different combinations of questionnaire design variables.

The first stage of the analysis is to estimate the reliability and validity of these tie characteristic questions. With this purpose, we will use the multitrait multimethod (MTMM) approach (Campbell and Fiske, 1959). Several other approaches exist to estimate the quality of a measurement instrument (Saris, 1990) like the quasi-simplex approach (Heise, 1969) and the repeated multimethod approach (Saris, 1995) but will not be dealt with in this paper.

Many different MTMM models have been suggested in the literature. Among them are the correlated uniqueness model (Marsh, 1989); the confirmatory factor analysis (CFA) model for MTMM data (Althauser et al., 1971; Alwin, 1974; Werts and Linn, 1970; Andrews, 1984); the direct product model (Browne, 1984); the true score (TS) model for MTMM data (Saris and Andrews, 1991). The MTMM model has rarely been used for measurement quality assessment in social network analysis. Hlebec (1999), Ferligoj and Hlebec (1999) and Kogovšek et al. (2002) used the TS model on network data, the first two in the context of complete networks and the last in the context of egocentered networks. The CFA specification is used in this study, not the TS. However, both models are equivalent (Coenders and Saris, 2000).

We consider egocentered network data as hierarchical; therefore, a multilevel analysis is required. We use this type of analysis because, as shown in Coromina et al. (2004), for egocentered networks, the multilevel approach provides a less biased and much richer view on measurement quality than classic methods (e.g., Härnqvist, 1978) and analyses of the data only at group (ego) level considering averages of all alters within the ego (Kogovšek et al., 2002). Multilevel factor analysis decomposes the total observed scores at the individual level into a between-group (ego) component and a within-group component (Muthén, 1989, 1990, 1994; Hox, 1993). Multilevel MTMM models make it possible to compute among others within ego, between ego and overall reliability and validity estimates. We suggest using this model on a general basis for the analysis of measurement quality of egocentered network data.

The second stage of the analysis is a meta-analysis of these reliability and validity estimates. Meta-analysis can be defined as the statistical analysis of a collection of results from individual studies (in our case done in Girona, Ljubljana and Ghent) with the purpose of integrating the findings (Glass, 1976). In the past, some meta-analyses of MTMM reliability and validity estimates have been done (Andrews, 1984; Sherpenzeel, 1995; Költringer, 1995; Hlebec, 1999). Hlebec (1999) specifically considered social network measurement. All these meta-analyses were concerned by personal, mail and telephone interviews only,

not web surveys. Besides, all these meta-analyses used a standard CFA MTMM model, not a multilevel model.

The structure of this article is as follows. First, we present a CFA MTMM model. Next, we extend the model to the multilevel case and interpret reliability and validity estimates in this context. Then, we present the questionnaire and the data collection in the three countries and, as illustration, we show the detailed results of the Girona sample. Next, we discuss the design of the meta-analysis and finally, we present and discuss its results, which consist on the contribution on reliability and validity of a number of questionnaire variables for network questions regarding frequency of contact. The implications for other self-administered data collection modes are also discussed.

## 2. Reliability and validity assessment

### 2.1. Reliability and validity defined

Reliability can be defined as the extent to which any questionnaire, test or measure produces the same results on repeated experiments. Due to random error, the repeated measures will not be exactly the same, but will be consistent to a certain degree. The more consistent the results given by repeated measurements, the higher the reliability. Validity is defined as the extent to which any measure measures what is intended to measure (Carmines and Zeller, 1979, p. 12). Validity is affected by the error called systematic, which has a biasing effect on measurement instruments.

Within construct validity we consider nomological, convergent and discriminant validity. Nomological validity implies that the relationships between measures of different concepts must be consistent with theoretically derived hypotheses concerning these concepts. Convergent validity refers to common trait variance and is inferred from large and statistically significant correlations between measures of the same trait using different methods. Discriminant validity refers to the distinctiveness of the different traits; it is inferred when correlations among different traits are less than one.

The amount of both random and systematic error present in a measurement can depend on any characteristic of the design of the study, such as data collection mode, questionnaire wording, response scale, type of training of the interviewer (Groves, 1989), all of which can be broadly considered as methods.

### 2.2. MTMM model

In this study, the main concerns are convergent and discriminant validity and reliability. Convergent and discriminant validity of different methods was first assessed in a systematic way by the design that we use, the MTMM design. In this design, three or more traits (variables of interest) are each measured with usually three or more methods.

Reliability assessment is based on the classical test theory (Lord and Novick, 1968) whose main equation is:

$$Y_{ij} = S_{ij} + e_{ij} \tag{1}$$

where $Y_{ij}$ is the response of variable $i$ measured by method $j$, $S_{ij}$ the part of the response that would be stable across identical independent repetitions of the measurement process and is called true score (Saris and Andrews, 1991) and $e_{ij}$ is the random error, related to lack of reliability.

In coherence with the MTMM approach, the stable part is assumed to be the combined result of trait and method:

$$S_{ij} = m_{ij}M_j + t_{ij}T_i \tag{2}$$

where $M_j$ is the variation in the scores due to the method, related to invalidity, $T_i$ the unobserved variable of interest (trait), related to validity, and $m_{ij}$ and $t_{ij}$ are factor loadings on the method and trait factors, respectively.

Eqs. (1) and (2) constitute the specification of the true score MTMM model of Saris and Andrews (1991). By substitution, we obtain Eq. (3) which corresponds to the confirmatory factor analysis specification of the MTMM model (Andrews, 1984):

$$Y_{ij} = m_{ij}M_j + t_{ij}T_i + e_{ij} \tag{3}$$

Coenders and Saris (2000) show both models to be equivalent and we select the CFA specification for its simplicity. In either of the models it is necessary to make some additional assumptions (Andrews, 1984):

$$\begin{aligned} &\mathrm{Cov}(T_i, e_{ij}) = 0, \quad \forall ij \\ &\mathrm{Cov}(M_j, e_{ij}) = 0, \quad \forall ij \\ &\mathrm{Cov}(T_i, M_j) = 0, \quad \forall ij \end{aligned} \tag{4}$$

These assumptions imply:

- There is no correlation between errors and latent variables, both traits and methods.
- There is no correlation between traits and methods.

These assumptions make it possible to decompose the variance of $Y_{ij}$ into trait variance $[t_{ij}^2 \mathrm{Var}(T_i)]$, method variance $[m_{ij}^2 \mathrm{Var}(M_j)]$ and random error variance $[\mathrm{Var}(e_{ij})]$ to assess measurement quality (Schmitt and Stults, 1986). Another set of assumptions is not always necessary but strongly recommended at least in some cases:

$$\begin{aligned} &\mathrm{Cov}(M_j, M_{j'}) = 0, \quad \forall j \neq j' \\ &m_{ij} = 1, \quad \forall ij \\ &t_{ij} = t_{i'j}, \quad \forall i \neq i' \end{aligned} \tag{5}$$

which imply:

- There is no correlation between methods.
- Method effects are equal within methods.
- The $t_{ij}$ coefficients are constant within method, which implies that the relationship between the units of measurement of methods is constant across traits.

The first two assumptions in Eq. (5) were suggested by Andrews (1984) and Sherpenzeel (1995) as a means to increase the rate of convergence of the estimation procedures and reduce the rate of appearance of inadmissible solutions (e.g., negative variances).

The last constraint in Eq. (5) was suggested by Coromina et al. (2004) for models with two methods, like the one in this article. If only two methods are used, the model without this constraint is still identified but standard errors can get very large. The constraint is reasonable if the response scales do not vary across methods (this will be the case in our article) or they vary in the same way for all traits. Coromina et al. (2004) reported standard errors to get lower by about 30% after introducing this constraint. In the case of four traits and two methods, the path diagram of the model with all these assumptions is displayed in Fig. 1.

The definitions of reliability and validity from classical test theory used in Saris and Andrews (1991) for the TS model can also be implemented in the CFA formulation of the model as follows. Reliability is the proportion of variance in $Y_{ij}$ that is stable across repeated measures with the same method:

$$\text{Reliability} = \frac{\text{Var}(S_{ij})}{\text{Var}(Y_{ij})} = \frac{m_{ij}^2 \text{Var}(M_j) + t_{ij}^2 \text{Var}(T_i)}{\text{Var}(Y_{ij})} \tag{6}$$

and the reliability coefficient is the square root of reliability. Thus, reliability increases not only with true or trait variance, but also with method variance, which also belongs to the stable or repeatable part of the measurements.

Validity, assuming that method is the only source of invalidity, is:

$$\text{Validity} = \frac{t_{ij}^2 \text{Var}(T_i)}{\text{Var}(S_{ij})} = \frac{t_{ij}^2 \text{Var}(T_i)}{m_{ij}^2 \text{Var}(M_j) + t_{ij}^2 \text{Var}(T_i)} \tag{7}$$

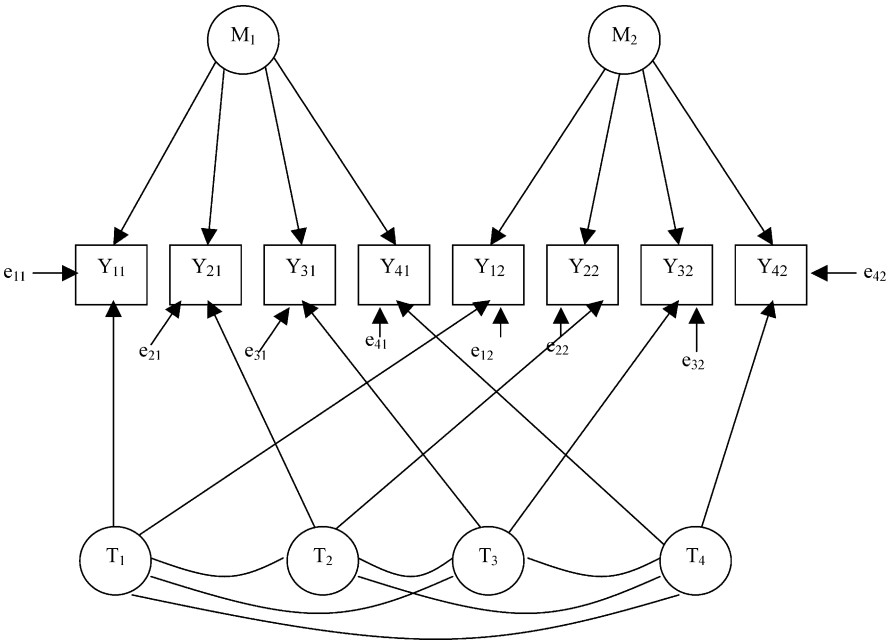



Fig. 1. Path diagram of a CFA MTMM model for two methods and four traits.

and the validity coefficient is the square root of validity. Validity is thus the percentage of variance of the true score explained by the trait. As explained before, the true score is the trait effect plus the method effect. Then, we can assess invalidity as 1 minus validity.

Another definition of validity uses the total variance in the denominator of Eq. (7), thus making reliability be the upper bound of validity. The advantage of the definition used in Saris and Andrews (1991) and presented here is that it makes the range of validity independent of the value of reliability, as validity can be equal to 1 even for unreliable measures.

### 2.3. Multilevel confirmatory factor analysis

Egocentered network data can be considered as hierarchical. In the hierarchical structure the egos are at the top of the hierarchy (i.e., at the group level), and all their alters at the bottom (i.e., at the individual level). Thus, alters are nested into the egos in what constitutes a nested data structure. Responses to the tie characteristic questions for each alter-ego combination constitute the data.

To analyze egocentered network data, Coromina et al. (2004) suggest using two-level factor analysis, or, more particularly, two-level MTMM models. Ignoring the inherent hierarchical structure of egocentered network data leads both to a poorer view or measurement quality and to biased standard errors and test statistics of the MTMM model. Therefore, two-level models must be recommended on a general basis for this type of network data.

According to two-level factor analysis models, the total population covariance matrix $\boldsymbol{\Sigma}_{\mathrm{T}}$ can be decomposed into a between-group covariance matrix $\boldsymbol{\Sigma}_{\mathrm{B}}$ and a within-group covariance matrix $\boldsymbol{\Sigma}_{\mathrm{W}}$:

$$\boldsymbol{\Sigma}_{\mathrm{T}} = \boldsymbol{\Sigma}_{\mathrm{B}} + \boldsymbol{\Sigma}_{\mathrm{W}} \tag{8}$$

This decomposition makes it possible to analyze each component separately and can be also applied to our MTMM model (Eq. (3)). We are thus able to decompose the model in two parts using the "W" and "B" subscripts for the within and between part, respectively:

$$Y_{ij} = \underbrace{m_{\mathrm{B}ij} M_{\mathrm{B}j} + t_{\mathrm{B}ij} T_{\mathrm{B}i} + e_{\mathrm{B}ij}}_{Y_{\mathrm{B}ij}} + \underbrace{m_{\mathrm{W}ij} M_{\mathrm{W}j} + t_{\mathrm{W}ij} T_{\mathrm{W}i} + e_{\mathrm{W}ij}}_{Y_{\mathrm{W}ij}} \tag{9}$$

An additional constraint is introduced:

$$t_{\mathrm{B}ij} = t_{\mathrm{W}ij}, \quad \forall ij \tag{10}$$

This constraint is reasonable because $t_{\mathrm{B}}$ coefficients depend on the units of measurement of traits and methods which must be the same at the within and between levels.

To estimate the model, Härnqvist (1978) proposes to do separate analyses on the within and between sample covariance matrices. Muthén (1989, 1990) shows that this can lead to biased estimates and suggests the Muthén's partial maximum likelihood (MUML) estimator which is formulated as a two-sample problem in which the within and between covariance matrices are treated as belonging to two different samples. The same model is fitted to both samples with certain constraints. If all group sizes are identical (balanced case), this estimator has maximum likelihood properties. In the unbalanced case, maximum likelihood

estimation implies specifying a separate sample for each distinct group size, which is what the full information maximum likelihood (FIML) approach does (Muthén, 1994). Thus, only in the balanced case (each ego has the same number of alters) are MUML and FIML equivalent (Hox, 1993). In the more common unbalanced case, MUML produces results close to FIML (Hox and Mass, 2001). In this paper, we use FIML whenever its requirement of number of groups larger than the number of model parameters is fulfilled. Otherwise, we use the MUML estimator.

Estimation can be done with the Mplus3 program (Muthén and Muthén, 2004). Yuan and Bentler's (2002) $\chi^2$ statistic is computed instead of the standard likelihood ratio statistic to increase robustness to non-normality.[1]

### 2.4. Interpretation

In a multilevel or hierarchical context, the evaluation of measurement quality can be much enriched. Quite trivially, we can obtain two reliabilities and two validities for each trait–method combination, that is, between and within groups. The fact that groups are respondents and individuals are stimuli evaluated by them makes these reliabilities and validities interpretable in a somewhat different way from standard multilevel factor analysis.

The between-group reliabilities and validities can be computed from the parameters of the between part of the model and can be interpreted with respect to the quality of the measurement of the egocentered network as a whole (average values of the answers for each ego computed over all his/her alters). In many social network studies, the ego is the focus of interest and these averages are used as data instead of the raw responses regarding individual alters. If this is the case, between-group measurement quality is the relevant one to look at.

The within-group reliabilities and validities can be computed from the parameters of the within part of the model and can be interpreted in a classic psychometric sense in which each subject is a separate unit of analysis and thus variance is defined across stimuli presented to the same subject, not across subjects (e.g., Lord, 1980). In many social network studies, the relationship is the focus of interest and if this is the case, within-group measurement quality is the relevant one to look at. A one-level analysis would ignore this distinction.

Hox (2002) suggests that percentages of variance cannot only be computed in each part of the model separately. The fact that the between and within scores add to a total score (Eq. (9)) makes it possible to compute percentages of variance in other attractive ways. In our case, if we decompose the variance according to Eq. (9) we have:

$$\begin{aligned}\mathrm{Var}(Y_{ij}) = {} & m_{\mathrm{W}ij}^2\,\mathrm{Var}(M_{\mathrm{W}j}) + m_{\mathrm{B}ij}^2\,\mathrm{Var}(M_{\mathrm{B}j}) + t_{\mathrm{W}ij}^2\,\mathrm{Var}(T_{\mathrm{W}i}) + t_{\mathrm{B}ij}^2\,\mathrm{Var}(T_{\mathrm{B}i}) \\ & + \mathrm{Var}(e_{\mathrm{W}ij}) + \mathrm{Var}(e_{\mathrm{B}ij}) \end{aligned} \tag{11}$$

Coromina et al. (2004) suggest that each of the six components in Eq. (11) can have its own interpretation:

[1] MLR option in Mplus3.

- Within method variance corresponds to differences in the use of methods among alters evaluated by the same ego. In a panel design like ours, a time specific component in the alters' evaluations would emerge here.
- Between method variance corresponds to differences among respondents (egos) in the use of methods. Thus, it is in complete agreement with the usual definition of method effect (e.g., Andrews, 1984).
- Within trait variance is the error-free variance corresponding to differences in the alter evaluations made by the same ego.
- Between trait variance is the error-free variance corresponding to differences in the average levels of the egos.
- Within error variance is not systematic in any way and thus truly corresponds to the definition of pure random measurement error.
- Between error variance is the error variance associated to measurements of average levels of the egos. Thus, it is somehow systematic as it is constant for all alters within the ego (otherwise it would average to zero).

Thus, from the decomposition in Eq. (11), percentages like the following could be of interest and can easily be computed. One can:

- Compute overall reliabilities and validities by aggregating all trait, method and error components.
- Compute overall percentages of within and between variance by aggregating all within components and all between components.
- Do the former only with respect of error-free variance, that is compute the percentage of between and within trait variance over the total trait variance. A higher within percentage shows a higher alter variability in the tie characteristics within a network and a lower variability of average values of the tie characteristics between networks.
- Compute a percentage of pure random error variance (i.e., within error variance) over the total variance of the observed variables (grand total, i.e., including all six components). The percentage of total variance explained by any of the other five components can be computed in a similar way.

## 3. Multitrait multimethod data

### *3.1. Survey design*

For this research web surveys were used. A specific characteristic of web questionnaire instruments is that they do not only consist of verbal features (words and numbers), but can make use of rich visual features (Couper, 2001; Best and Krueger, 2004). These features could include the use of multiple colors, special navigational features (e.g., indexes, tables of contents, progress indicators), animations, etc. These can be added to traditionally presented survey questions in order to illustrate them or simply to motivate the respondents. Other advantages of using web questionnaire are lower costs (though some research done in Couper et al., 1998, has shown that this is not always the case), faster data collection and data analysis process and that questionnaires can quickly be modified (Watt, 1997; Sheehan

and McMillan, 1999; Brennan et al., 1999; Crawford et al., 2002) and possibility to respond at the time the respondent finds convenient (Totten, 2003). An important advantage in the context of network questionnaires is that some software can remember the alter. This feature is called "piping" (assigning questionnaire items based on earlier answers from the respondent). The software remembers the text entry of the alter's name in name generator questions and uses it in tie characteristics questions regarding this alter (Lozar Manfreda et al., 2004). Many of these features are shared with computer assisted personal interviews, some others are shared with paper and pencil self-administered interviews.

However, web surveys have been restricted to populations with nearly universal internet access; otherwise, an important coverage error can be made (Schaefer and Dillman, 1998; Schonlau et al., 2004). A solution proposed for the coverage error in those papers is to use web surveys as a part of a mixed mode design. Moreover, they suggest taking the greatest care in making the questions easy to answer and not using too many media effects which can increase the questionnaire burden with the consequence of a pre-mature abandon by the respondent. In order that all respondents receive an uniform questionnaire (Best and Krueger, 2004), the design must also be robust to variations of the software that respondents may have.

Web questionnaires are especially appropriate for our target population of Ph.D. students who use the computer daily for their job and have fast internet connection. Thus, coverage of the population is not a problem in our case. Besides, the population members can be assumed to be highly motivated to respond out of solidarity, as many researchers involved in the project were Ph.D. students at the time of the data collection. The proximity between respondents and researchers also made recalls easier. All non-respondents were offered the possibility of a personal or self-administered paper and pencil interview but this offer did not produce any additional refusal conversion.

The web questionnaire design has been a complex and long process led by Daniëlle de Lange and involving 2 years of discussion within the INSOC research group, several international meetings, several focus groups and pre-tests (de Lange et al., 2004). The fact that we had to produce comparable versions that could work in three languages (Spanish, Slovenian and Dutch) and university systems lengthened the process even further (Behling and Law, 2000) and involved two independent translations, a pre-test of the translated questionnaires and further discussions and modifications. Our questionnaire uses respondent-friendly design principles (Dillman et al., 1998a), which implies constructing the web questionnaire in a manner that increases the likelihood that sampled individuals will respond to the survey request, and that they will do so accurately. The web questionnaire administration was centralized at the Ghent University, in Belgium, using the SNAP software in its Version 7 (Mercator Research Group, 2003).

### *3.2. Population, sample and data collection*

The population we analyze for this paper are Ph.D. students who began their doctoral studies at the Universities of Girona, Ljubljana and Ghent in the academic years 1999/2000 and 2000/2001. These Ph.D. students must have a strong tie with their university; in other words, these students must have grants, be assistants or be researchers hired for research projects. This choice has been made because these people have more frequent contact with

other researchers, and they can spend more time doing research, which is their main job. Therefore, these students are likely to have more need for advice, cooperation or information than those who are only linked to the university only by their doctoral studies and who may even not belong to a research group.

The procedure used to define the research group members was the following. During April 2003, Ph.D. students in Spain and Belgium were phoned in order to know who their promoter was. In Slovenia, the names of promoters were obtained from a list provided by the Slovenian Ministry of Science. Afterwards, promoters from all three countries were interviewed in person and were given the name generator questions in order to obtain a list of their research groups in connection with the topic of the dissertation of the Ph.D. student. This group does not have to correspond to any official or formal research group recognized by the university because we are interested in getting the research groups that are relevant to each Ph.D. student.

During November 2003, e-mail invitations to participate in web questionnaires were sent to all Ph.D. students fulfilling the above mentioned eligibility criteria, (86 in Girona, 191 in Ljubljana and 233 in Ghent). After 1, 4 and 8 weeks follow-up e-mails were sent to remind non-respondents, Belgium used one additional follow-up. After the last follow-up the total was 67 respondent Ph.D. students in Girona, 118 in Ljubljana and 198 in Ghent, which represents response rates of 78, 62 and 85%, respectively.

Two weeks (in Belgium this period was much longer, up to 6 months) after answering the first questionnaire (which we refer to as method 1), students were sent an invitation to answer a much shorter follow-up questionnaire containing only the tie characteristic questions considered in this article (see Section 3.3) using a different web questionnaire design with different question order, different style of response category labels and different graphical display and lay-out of the questions (method 2). The differences between both methods are explained in Section 5 in greater detail. We got 61 complete responses (91% second wave response rate) in Girona, 81 complete responses (69% second wave response rate) in Ljubljana and 55 complete responses (60% second wave response rate) in Ghent for the second method. For the second method, we only used a part of the Ph.D. student sample in Ghent because the other part was used for another type of experiment.

### 3.3. Traits used

In our case, egos are Ph.D. students and alters are the members of their research group as defined in Section 3.2. The questionnaire was personalized for each and every Ph.D. student as it contained the list of his or her social network group members. This questionnaire contained, among others, questions used to describe the relationships between Ph.D. students and members of their networks, which will constitute the traits of our study (frequency of scientific advice in work problems, frequency of collaboration in research, frequency of asking for crucial information and frequency of social activities outside work context):

- *Trait 1*: Consider all the work-related problems you have had in the past year (namely since 1 November 2002) and that you were unable to solve yourself. How often did you ask each of your colleagues on the following list for *scientific advice*?

- *Trait 2*: Consider all situations in the past year (namely since 1 November 2002) in which you *collaborated* with your colleagues concerning research, e.g., working on the same project, solving problems together, etc. The occasional piece of advice does not belong to this type of collaboration. How often have you collaborated with each of your colleagues concerning research in the past year?
- *Trait 3*: Consider all situations in the past year (namely since 1 November 2002) in which you needed crucial information, data, software, etc., for your work but did not have it in your possession. How often did you ask each of your colleagues for *information/data/software*, etc., in the course of the past year?
- *Trait 4*: Sometimes colleagues do *social activities outside the work* context, such as sport activities or attending social or cultural events. [Attention: lunching together on a working day and activities organized by the university itself, such as courses, formal dinners or conferences do not belong to the target group of activities!] How often did you engage in social activities outside of work with your colleagues in the past year (namely since 1 November 2002)?

## 4. MTMM results in the Girona sample

As an illustration of the use and interpretation of the multilevel MTMM model, we show the results in the Girona sample. The data of the other samples were treated in a similar way and used in the meta-analysis of Section 5.

The goodness of fit assessment was done with the same standard procedures used for confirmatory factor analysis models (e.g., Batista-Foguet and Coenders, 2000; Bollen and Long, 1993; Coromina et al., 2004) and was on the border of being acceptable after constraining to zero two insignificant negative variances ($M_{W1}$ and $e_{B12}$): Yuan-Bentler $\chi^2 = 103.1$ with 33 d.f.; Tucker and Lewis Non-Normed Fit Index = 0.946.

Table 1 shows the variance decomposition according to Eq. (11) for the eight variables (trait–method combinations) obtained from this analysis. From this table, within, between and total reliabilities and validities and all other results described in Section 2.4 can be obtained. Table 2 shows reliability and validity coefficients at the within, between, and overall levels. Boldfaced values correspond to variances fixed to zero to estimate the model.

From these results on Table 2, overall, method 1 is the most valid and method 2 the most reliable. This overall result also holds both at the between level (the relevant one if the focus of analysis are ego averages) and at the within level (the relevant focus of analysis are individual contacts). Trait 1 (frequency of asking for advice) seems to be the most valid and reliable for both methods and trait 4 (frequency of socializing) the least. The overall relative measurement quality of traits, only emerges at the within level.

As suggested in Section 2.4, the results of the analysis can be used to decompose variance in many interesting ways by combining interesting sets of the six variance components in Eq. (11) and Table 1. Table 3 shows the percentage of within trait variance over all trait variance. The results show that most of the error-free variance corresponds to the within level. This means that egos really discriminate between the different alters.

Table 1
Decomposition into six variance components[a]

| | $T_1$ (%) | $T_2$ (%) | $T_3$ (%) | $T_4$ (%) |
|---|---|---|---|---|
| Within trait variance | | | | |
| $M_1$ | 67.5 | 65.3 | 55.6 | 67.0 |
| $M_2$ | 66.1 | 59.3 | 47.7 | 42.4 |
| Within method variance | | | | |
| $M_1$ | **0.0** | **0.0** | **0.0** | **0.0** |
| $M_2$ | 9.6 | 9.2 | 12.3 | 13.1 |
| Within error variance | | | | |
| $M_1$ | 13.3 | 17.2 | 17.3 | 14.5 |
| $M_2$ | 7.0 | 12.3 | 16.3 | 25.2 |
| Between trait variance | | | | |
| $M_1$ | 12.2 | 13.1 | 18.5 | 11.4 |
| $M_2$ | 12.0 | 11.9 | 15.9 | 7.2 |
| Between method variance | | | | |
| $M_1$ | 0.4 | 0.4 | 0.6 | 0.9 |
| $M_2$ | 5.2 | 5.0 | 6.7 | 7.1 |
| Between error variance | | | | |
| $M_1$ | 6.6 | 3.9 | 8.0 | 6.3 |
| $M_2$ | **0.0** | 2.3 | 1.2 | 4.9 |

[a] Boldfaced for variances constrained to zero.

Table 2
Multilevel reliabilities and validities[a]

| | Within level | | | | Between level | | | | Overall level | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ |
| Reliability coefficients | | | | | | | | | | | | |
| $M_1$ | 0.91 | 0.89 | 0.87 | 0.91 | 0.81 | 0.88 | 0.84 | 0.81 | 0.90 | 0.89 | 0.86 | 0.89 |
| $M_2$ | 0.96 | 0.92 | 0.89 | 0.83 | **1.00** | 0.94 | 0.98 | 0.86 | 0.96 | 0.92 | 0.91 | 0.84 |
| Validity coefficients | | | | | | | | | | | | |
| $M_1$ | **1.00** | **1.00** | **1.00** | **1.00** | 0.98 | 0.98 | 0.98 | 0.96 | 1.00 | 1.00 | 1.00 | 0.99 |
| $M_2$ | 0.93 | 0.93 | 0.89 | 0.87 | 0.83 | 0.84 | 0.84 | 0.71 | 0.92 | 0.91 | 0.88 | 0.84 |

[a] Boldfaced for variances constrained to zero which lead to reliabilities or validities equal to one.

Table 3
Percentage of trait variance operating at the within level

| | $t^2_{\mathrm{W}ij}\,\mathrm{Var}(T_{\mathrm{W}i})/[t^2_{\mathrm{W}ij}\,\mathrm{Var}(T_{\mathrm{W}i}) + t^2_{\mathrm{B}ij}\,\mathrm{Var}(T_{\mathrm{B}i})]$ |
|---|---|
| $T_1$ | 84.6 |
| $T_2$ | 83.3 |
| $T_3$ | 75.0 |
| $T_4$ | 85.4 |

## 5. Meta-analysis

Most often a meta-analysis is done with the aim of integrating statistical analysis in literature reviews (e.g., de Leeuw and van der Zouwen, 1988). The main disadvantage of such meta-analysis is that the researcher has no control over the design of the individual studies (Sherpenzeel, 1995).

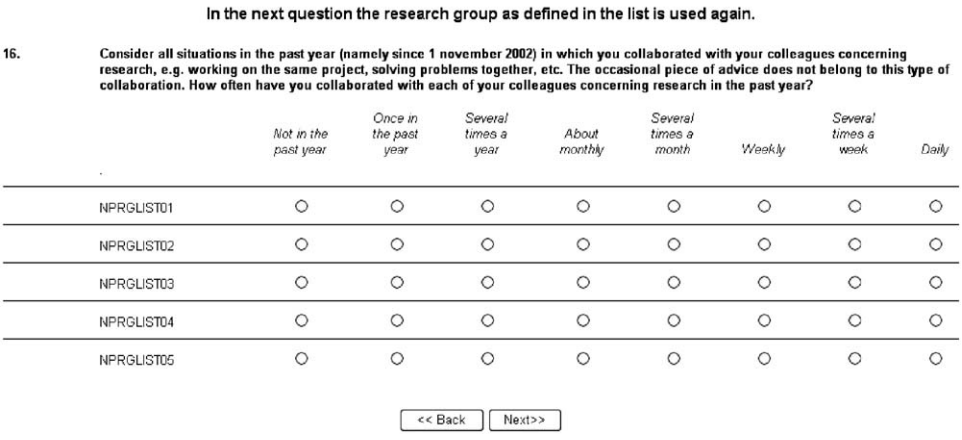
Another type of meta-analysis is done to summarize the results of studies carried out by the same research team (e.g., Wolf et al., 1984). In this paper, we are concerned by this latter type because the data collections in the three universities of three countries (Girona in Spain, Ljubljana in Slovenia and Ghent in Belgium) were centrally coordinated in Ghent, which made it possible to control and vary a series of factors that we believed to have an influence on the quality of the data, that is, on reliability and validity on network data collection. The aim of meta-analysis is then to estimate the contribution of each of these factors on reliability and validity.

We have considered three factors along which measurement methods can differ in the context of web questionnaires focused in social network data collection and which can affect reliability and validity.

### 5.1. Question order: by alters or by questions

After we obtain the list of alters with name generators, we can ask tie characteristic questions in two ways. One way ("by questions", see Fig. 2) is to take the question and ask this question for all alters on the list, and then go to the next question. The other way ("by alters", see Fig. 3) is to take each alter individually and to ask all questions about him/her, and then go to the next alter.

The quality of egocentered network measurements, when the network questions are organized by alters or by questions has been studied by Kogovšek et al. (2002) using



In the next question the research group as defined in the list is used again.

16. Consider all situations in the past year (namely since 1 november 2002) in which you collaborated with your colleagues concerning research, e.g. working on the same project, solving problems together, etc. The occasional piece of advice does not belong to this type of collaboration. How often have you collaborated with each of your colleagues concerning research in the past year?

| | Not in the past year | Once in the past year | Several times a year | About monthly | Several times a month | Weekly | Several times a week | Daily |
|---|---|---|---|---|---|---|---|---|
| NPRGLIST01 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| NPRGLIST02 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| NPRGLIST03 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| NPRGLIST04 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| NPRGLIST05 | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

<< Back Next>>

Fig. 2. Formulation by questions, with all labels and with a plain lay-out. NPRGLIST01 to NPRGLIST05 show where alters' names are inserted.

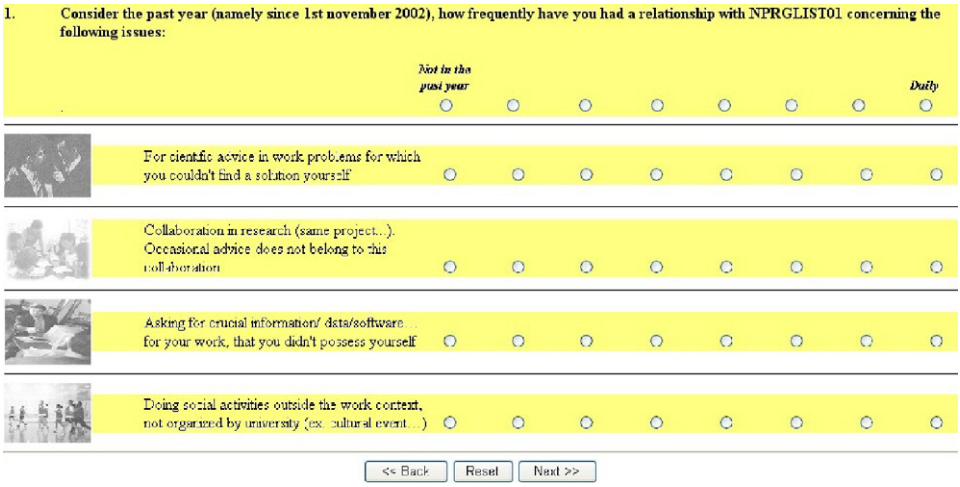



Fig. 3. Formulation by alters, with end labels and with a graphical lay-out. NPRGLIST01 shows where the alters' name is inserted.

telephone interviews. According to the findings of the authors the formula "by alters" seems to be more reliable for telephone interviews. The explanation given by the authors is that when the respondent answers all questions by alters, the reference frame is the current alter. When the respondent answers by questions for all alters, the reference frame is the current question. By questions, 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. Therefore, when using the telephone method, context effects would be more present in the case of network data collection by questions. We will see if this phenomenon replicates for web surveys, in which the respondent can simultaneously view on the screen the complete list of all alters in his/her network or the list of all questions and can respond in any order.

### *5.2. Response category labels: for all categories or for the end points of the response scale*

The response categories used for all questions were:

1. not during the past year;
2. once in the past year;
3. several times a year;
4. about monthly;
5. several times a month;
6. weekly;
7. several times a week;
8. daily.

We can label all of the categories (Fig. 2) or only a subset of them (Fig. 3), for instance the two extreme ones. Evidence of similar meta-analysis studies regarding the virtues of each alternative is mixed. Andrews (1984) reports partial labeling to be better both in terms of reliability and validity. Költringer (1995) reports no effect of the way of labeling on either reliability or validity. Both studies referred to personal and telephone interviews and mostly related to attitudinal (i.e., not network) variables using vague category labels of the type "rather satisfied", "completely agree" and the like.

On the one hand, the way the scale is presented in the computer screen by the web questionnaire when only the end categories are labeled (Fig. 3) is analogous to a line production scale, which is advocated to produce higher quality data (Saris, 1987; Lodge, 1981; van Doorn et al., 1983), and for which only extreme labels are reported to be necessary (Saris, 1988). On the other hand, since the questions used are dealing with frequency of contact in a network, the category labels we use in this study are not vague quantifiers but precise actual frequencies of behavior and thus additional labels may help respondents give precise answers about the frequency of contact with their social network.

### *5.3. Lay-out of the questions and the web page: plain or graphical display*

In a web survey we have the choice between a plain questionnaire without color, images and html tables, which requires low transmission times and has a conventional questionnaire format, and a questionnaire with graphical display web design options. Some results for a similar experiment can be found in Dillman et al. (1998b), who suggest that using a plain questionnaire provided better results than a graphical display version. In their study, the plain questionnaire obtained a higher response rate, was more likely to be fully completed and took respondents less time to complete. Thus, the utilization of graphical display and page lay-out design features available to create web questionnaires does not seem to improve data quality.

Dillman's study was carried out 6 years ago. From then to now transmission times are likely to have significantly dropped while the power of most people's browsers has improved. This means that it is not so sure that nowadays the plain questionnaire still offers advantages over a graphical display design. In a recent study, Deutskens et al. (2004) found that visual effects actually increase response quality.

In our case, the plain design included only text (Fig. 2) and the graphical display design a background color and pictures related to the topic asked in each question (Fig. 3). These pictures can provide hint about the type of network relationship that is being asked for.

In all participant universities network data were first collected by questions, with all labels and a plain design (method 1). For the follow-up questionnaire (method 2), the different universities used different combinations of factors in what can be considered to be a fractional factorial experimental design. Ljubljana had the largest sample size which made it possible to split it into two for the follow-up questionnaire. Table 4 shows this experimental design. Although the design is not orthogonal, it performs quite well in terms of collinearity: all tolerances for the meta-analysis model are between 0.4 and 0.7 when including the main effects of country, trait and all three factors along which methods differ.

In the meta-analysis, dependent variables were as computed in Tables 1–3 but variances constrained to zero in the MTMM model were treated as missing:

Table 4
Experimental design

| University and country | Sample | Repetition | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|---|---|
| Girona (Spain) | Only one | Main questionnaire | By questions | All labels | Plain |
| Girona (Spain) | Only one | Follow-up | By alters | End labels | Plain |
| Ljubljana (Slovenia) | 1 and 2 | Main questionnaire | By questions | All labels | Plain |
| Ljubljana (Slovenia) | 1 | Follow-up | By questions | End labels | Graphical |
| Ljubljana (Slovenia) | 2 | Follow-up | By alters | All labels | Graphical |
| Ghent (Belgium) | Only one | Main questionnaire | By questions | All labels | Plain |
| Ghent (Belgium) | Only one | Follow-up | By alters | End labels | Plain |

- percentage of between trait variance over total variance;
- percentage of within trait variance over total variance;
- percentage of between method variance over total variance;
- percentage of within method variance over total variance;
- percentage of between error variance over total variance;
- percentage of within error variance over total variance;
- between reliability coefficient;
- between validity coefficient;
- within reliability coefficient;
- within validity coefficient;
- overall reliability coefficient;
- overall validity coefficient;
- percentage of trait variance that operates at the within level.

As all predictors are categorical (country, trait and factors 1–3), a multiple classification analysis (MCA) was used (Andrews et al., 1973). This is a variant of analysis of variance that presents estimates in a way specially suited for non-orthogonal designs and which has been used by most of the meta-analyses cited in this article.

## 6. Meta-analysis results

The $\beta$ statistics of each factor are in Table 5, boldfaced if significant ($\alpha = 0.05$). These statistics can be interpreted as standardized regression coefficients for categorical predictors. The means of each level of all factors, corrected by the levels of other factors, are displayed in Table 6.

If we concentrate on overall reliabilities and validities we first see that their means are always around or above 0.85, thus showing valid and reliable tie characteristic data. By traits, the most validly and reliably measured is trait 2 (collaboration), followed by trait 1 (advice), trait 3 (asking for crucial information) and the less reliable and valid is trait 4 (socializing). This pattern is roughly consistent at the within, between and overall levels. As regards the questionnaire design factors, overall reliability and validity is higher when the social network questions in a survey are organized by questions. This pattern mostly emerges both at the within and the between levels. Between and overall validities are

Table 5
$\beta$ statistics for the different factors, country and trait[a]

| | Country | Trait | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|---|---|
| Percentage of between trait variance | 0.449 | 0.415 | **0.484** | 0.140 | 0.113 |
| Percentage of within trait variance | 0.199 | **0.562** | **0.560** | 0.021 | 0.053 |
| Percentage of between method variance | **0.373** | **0.281** | **0.442** | **0.847** | **0.396** |
| Percentage of within method variance | 0.170 | **0.493** | **0.781** | 0.038 | **0.310** |
| Percentage of between error variance | 0.206 | 0.459 | 0.299 | 0.107 | 0.018 |
| Percentage of within error variance | 0.166 | 0.450 | **0.532** | 0.207 | 0.106 |
| Between reliability | 0.500 | **0.498** | 0.093 | 0.187 | 0.062 |
| Between validity | 0.164 | **0.299** | **0.440** | **0.678** | **0.325** |
| Within reliability | 0.174 | **0.491** | **0.501** | 0.169 | 0.089 |
| Within validity | 0.259 | **0.549** | **0.750** | 0.088 | 0.297 |
| Overall reliability | 0.130 | **0.514** | **0.431** | 0.202 | 0.159 |
| Overall validity | 0.242 | **0.597** | **0.651** | **0.291** | **0.314** |
| Percentage of trait variance at within level | 0.300 | **0.524** | 0.157 | 0.154 | 0.019 |

[a] Boldfaced if significant ($\alpha = 5\%$).

higher for measurements with all labeled categories and a graphical display questionnaire design, but not within validities. Thus, this last result is mostly relevant for researchers interested in measuring network averages. Country has no significant effect on reliability or validity, thus arguing for the generalizability of the findings to different cultural and linguistic communities of respondents.

As regards percentages of trait variance at the within level, they are highest for trait 1 (advice) and lowest for trait 4 (socializing), thus showing that scientific advice tends to be asked rather often to some group members and not often to other group members by the same ego, but all egos have a rather similar frequency average, while this occurs to a lesser extent for socializing.

## 7. Discussion

In this article we have used a meta-analysis of MTMM estimates to study the quality of egocentered network data. As egocentered network data are hierarchical, multilevel analyses are always recommended.

From this multilevel analysis, we can immediately obtain two reliabilities and validities for each trait–method combination, namely between and within egos. Each of them has a different interpretation. It is also possible to compute overall reliabilities and validities by aggregating all trait, method and error components in order to obtain similar results to a classic (not multilevel) analysis (Coromina et al., 2004). As is usually done, we can also asses which percentage of variance is due to within and between differences. However, other useful variance percentages can be obtained by combining different within and between components in a meaningful way (Hox, 2002) depending on the results one is interested in for a particular research problem.

After the meta-analysis results, we can reach some conclusions. Regarding factor 1 (social network question order by alters or by questions), in previous studies, Kogovšek

Table 6
Corrected means for all factor levels[a]

| | Country | | | Trait | | | |
|---|---|---|---|---|---|---|---|
| | Girona | Ljubljana | Ghent | Trait 1 | Trait 2 | Trait 3 | Trait 4 |
| Percentage of between trait variance | 0.133 | 0.088 | 0.117 | 0.079 | 0.125 | 0.118 | 0.111 |
| Percentage of within trait variance | 0.610 | 0.612 | 0.555 | **0.674** | **0.654** | **0.552** | **0.508** |
| Percentage of between method variance | **0.024** | **0.025** | **0.002** | **0.017** | **0.015** | **0.022** | **0.030** |
| Percentage of within method variance | 0.049 | 0.074 | 0.077 | **0.052** | **0.047** | **0.072** | **0.114** |
| Percentage of between error variance | 0.049 | 0.064 | 0.057 | 0.053 | 0.039 | 0.063 | 0.078 |
| Percentage of within error variance | 0.152 | 0.166 | 0.189 | 0.137 | 0.132 | 0.186 | 0.219 |
| Between reliability | 0.874 | 0.735 | 0.839 | **0.794** | **0.875** | **0.804** | **0.698** |
| Between validity | 0.921 | 0.907 | 0.943 | **0.916** | **0.945** | **0.919** | **0.876** |
| Within reliability | 0.899 | 0.895 | 0.875 | **0.916** | **0.917** | **0.877** | **0.856** |
| Within validity | 0.974 | 0.932 | 0.931 | **0.963** | **0.965** | **0.936** | **0.887** |
| Overall reliability | 0.885 | 0.882 | 0.869 | **0.892** | **0.913** | **0.867** | **0.849** |
| Overall validity | 0.963 | 0.923 | 0.935 | **0.955** | **0.960** | **0.932** | **0.867** |
| Percentage of trait variance at within level | 0.823 | 0.864 | 0.826 | **0.895** | **0.840** | **0.823** | **0.804** |

| | Factor 1 | | Factor 2 | | Factor 3 | |
|---|---|---|---|---|---|---|
| | By questions | By alters | All labels | End labels | Plain lay-out | Graphical lay-out |
| Percentage of between trait variance | **0.125** | **0.082** | 0.103 | 0.116 | 0.105 | 0.117 |
| Percentage of within trait variance | **0.650** | **0.508** | 0.595 | 0.600 | 0.593 | 0.608 |
| Percentage of between method variance | **0.013** | **0.031** | **0.006** | **0.042** | **0.026** | **0.008** |
| Percentage of within method variance | **0.035** | **0.120** | 0.070 | 0.074 | **0.082** | **0.045** |
| Percentage of between error variance | 0.065 | 0.046 | 0.061 | 0.054 | 0.059 | 0.058 |
| Percentage of within error variance | **0.136** | **0.215** | 0.181 | 0.147 | 0.163 | 0.183 |
| Between reliability | 0.785 | 0.809 | 0.777 | 0.825 | 0.789 | 0.806 |
| Between validity | **0.948** | **0.880** | **0.961** | **0.856** | **0.902** | **0.958** |
| Within reliability | **0.912** | **0.857** | 0.884 | 0.903 | 0.894 | 0.883 |
| Within validity | **0.975** | **0.888** | 0.942 | 0.932 | 0.927 | 0.965 |
| Overall reliability | **0.895** | **0.852** | 0.873 | 0.893 | 0.884 | 0.868 |
| Overall validity | **0.967** | **0.893** | **0.947** | **0.914** | **0.922** | **0.962** |
| Percentage of trait variance at within level | 0.835 | 0.855 | 0.850 | 0.830 | 0.844 | 0.841 |

[a] Boldfaced if significant ($\alpha = 5\%$).

et al. (2002) suggested that the question order "by alters" was better, but that study was carried out using the telephone method. According to our findings, the most valid and reliable method of social network question order for web questionnaires is "by questions". Regarding validity, the explanation may be that by alters all questions can simultaneously be seen on the screen, which may increase the likelihood of committing common errors for all questions, which is what method effects are about. This result is very relevant for network questionnaires, and is likely to be replicate for any self-administered data collection mode.

Our results for factor 2 (all categories or only end points of the scale labeled) show that a higher validity is obtained when all labels are used. The reason why extra labels are helpful may be that in our questionnaire labels indicate precise social contact frequencies and not vague or unclear quantifiers like "agree", "not much agree", "undecided" and so on. This is typical of any frequency of contact question (a most common type of question in

social network research), and we hypothesize that it may generalize to any data collection mode.

Results for factor 3 (lay-out of the questions and the web page: plain or graphical display) show that a graphical display design improves validity. We obtained same results than Deutskens et al. (2004) when they found that visual effects increase response quality. In our experiment, only background color and some pictures related to the topic asked in each question were incorporated in the graphical display design, not sophisticated multimedia. This may have helped speed up download times together with the generally powerful computers a doctoral student is expected to have. On the other hand, the questionnaire resided in a central server in Belgium, so that Slovene respondents (the ones actually responding to the graphical display design) did not benefit from the quicker than usual connection that would result for an intranet questionnaire. This result is likely to replicate for any self-administered computed assisted data collection mode.

Another issue to take in consideration is the delay up to 6 months in Ghent sample between the first questionnaires (method 1) and second (method 2). Some changes in the social network surrounding Ph.D. students could occur during that period. If this would occur, the reliability would be lower in Ghent. This is not the case because the country variable was non-significant.

As a summary, we suggest that egocentered network data quality collected via web is higher when ordering the questionnaire by questions, labeling all response categories and using graphical display design. Data quality obtained via web seems to compare well with that of traditional modes provided that the characteristics of the population prevent the occurrence of the high coverage and non-response errors often encountered in web surveys.

A possible treat to the external validity of our experimental findings is that, in order to produce a comparative data set for substantive research purposes, the first wave used exactly the same method combination in all countries. Thus, order by alters, graphical display design and end labels can be to some extent confounded with the fact of conducting a separate shorter social network questionnaire in a second wave. The fact that for some factors and some types of validity and reliability the second wave method was no better or even worse than the first, reassures us that this confounding effect cannot have been large.

Our basic approach consisting of multilevel MTMM models and meta-analysis could be extended to study any other design factors that are relevant to any type of question or questionnaire administration mode for egocentered networks.

## Acknowledgements

This work was partly supported by the University of Girona Grants GRHCS66, BR00/UdG and 4E200304 and is a partial result of a wider project carried out at the International Network on Social Capital and Performance (INSOC, http://srcvserv.ugent.be/insoc/insoc.htm). Acknowledgements are due to all INSOC members who contributed to the proposal, the questionnaire design and the data collection: Hans Waege, Daniëlle de Lange, Filip Agneessens, Anuška Ferligoj, Tina Kogovšek, Uroš Matelič, Dagmar Krebs, Jurgen Hoffmeyer-Zlotnik, Brendan Bunting, Valentina Hlebec, Bettina Langfeldt and Joanne Innes. Many of them read earlier version of this article and made crucial suggestions.

## References


Althauser, R.P., Heberlein, T.A., Scott, R.A., 1971. A causal assessment of validity: the augmented multitrait–multimethod matrix. In: Blalock Jr., H.M. (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.

Andrews, F.M., 1984. Construct validity and error components of survey measures: a structural modeling approach. Public Opinion Quarterly 48, 409–442.

Andrews, F.M., Morgan, J.N., Sonquist, J.A., Klem, L., 1973. Multiple Classification Analysis. Institute for Social Research, Ann Arbor, MI.

Batista-Foguet, J.M., Coenders, G., 2000. Modelos de Ecuaciones Estructurales. La Muralla, Madrid.

Behling, O., Law, K.S., 2000. Translating Questionnaires and Other Research Instruments: Problems and Solutions. Sage, Thousand Oaks, CA.

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

Best, S.J., Krueger, B.S., 2004. Internet Data Collection. Sage, Thousand Oaks, CA.

Bollen, K.A., Long, J.S., 1993. Testing Structural Equation Models. Sage, Newbury Park, CA.

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

Brennan, M., Rae, N., Parackal, M., 1999. Survey-based experimental research via the web: some observations. Marketing Bulletin 10, 83–92.

Brewer, D., 2000. Forgetting in recall-based elicitation of personal and social networks. Social Networks 22, 29–43.

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

Burt, R.S., 1984. Network items and the general social survey. Social Networks 6, 293–339.

Campbell, 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, Newbury Park, CA.

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

Coromina, Ll., Coenders, G., Kogovsek, T., 2004. Multilevel multitrait multimethod model. Application to the measurement of egocentered social networks. Metodološki Zvezki. Advances in Methodology and Statistics 1, 323–349.

Couper, M.P., 2000. Web surveys: a review of issues and approaches. Public Opinion Quarterly 4, 464–494.

Couper, M.P., 2001. The promises and perils of web surveys. In: Westlake, A. (Ed.), The Challenge of the Internet. Association for Survey Computing, London.

Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nichols, W., O'Reilly, J., 1998. Computer Assisted Survey Information Collection. Wiley, New York.

Couper, M.P., Traugott, M., Lamias, M., 2001. Web survey design and administration. Public Opinion Quarterly 65, 230–253.

Crawford, S., McCabe, S., Couper, M.P., Boyd, C., 2002. From mail to web. Improving response rates and data collection efficiencies. In: Paper Presented at International Conference on Improving Surveys, Denmark, Copenhagen.

Deutskens, E., Ruyter, K., Wetzels, M., Oosterveld, P., 2004. Response rate and response quality of internet-based surveys: an experimental study. Marketing Letters 15, 21–36.

Dillman, D.A., 2000. Mail and Internet Surveys. The Tailored Design Method. Wiley, New York.

Dillman, D.A., Tortora, R.D., Bowker, D., 1998a. Principles for Constructing Web Surveys. SESRC Technical Report, vol. 98-50, Pullman, Washington.

Dillman, D.A., Tortora, R.D., Bowker, D., 1998b. Influence of plain vs. fancy design on response rates for web surveys. In: The 1998 Proceedings of Section on Survey Research Methods. American Statistical Association, Dallas, TX.

de Lange, D., Agneessens, F., Waege, H., 2004. Asking social network questions: a quality assessment of different measures. Metodološki Zvezki. Advances in Methodology and Statistics 1, 351–378.

de Leeuw, E.D., van der Zouwen, J., 1988. Data quality in telephone and face to face surveys: a comparative meta-analysis. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L., Waksberg, J. (Eds.), Telephone Survey Methodology. Wiley, New York, pp. 283–299.

Feld, S.L., Carter, W.C., 2002. Detecting measurement bias in respondent reports of personal networks. Social Networks 24, 365–383.

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

Glass, G., 1976. Primary, secondary and meta-analysis of research. Educational Researchers 5, 3–8.

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

Härnqvist, K., 1978. Primary mental abilities at collective and individual levels. Journal of Educational Psychology 70, 706–716.

Heise, D.R., 1969. Separating reliability and stability in test–retest correlations. American Sociological Review 34, 93–101.

Hlebec, V., 1999. Evaluation of survey measurement instruments for measuring social networks. Doctoral Dissertation. University of Ljubljana, Slovenia.

Hlebec, V., Ferligoj, A., 2002. Reliability of social network measurement instruments. Field Methods 14, 288–306.

Hoffmeyer-Zlotnik, J.H.P., 1990. The Mannheim comparative network research. In: Weesie, J., Flap, H. (Eds.), Social Networks Through Time. ISOR, Utrecht, pp. 265–279.

Hox, J.J., 1993. Factor analysis of multilevel data: gauging the Muthén method. In: Oud, J.H.L., van Blokland-Vogelesang, R.A.W. (Eds.), Advances in Longitudinal and Multivariate Analysis in the Behavioral Sciences. ITS, Nijmegen, The Netherlands, pp. 141–156.

Hox, J.J., 2002. Multilevel Analysis. Techniques and Applications. Lawrence Erlbaum, Mahwah, NJ.

Hox, J.J., Mass, C.J.M., 2001. The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. Structural Equation Modeling 8, 157–174.

Kogovšek, T., Ferligoj, A., Coenders, G., Saris, W.E., 2002. Estimating the reliability and validity of personal support measures: full information ML estimation with planned incomplete data. Social Networks 24, 1–20.

Költringer, R., 1995. Measurement quality in Austrian personal interview surveys. In: Saris, W.E., Münnich, Á. (Eds.), The Multitrait Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press, Budapest, pp. 207–224.

Lodge, M., 1981. Magnitude Scaling. Quantitative Measurement of Opinions. Sage, Beverly Hills, CA.

Lord, F.M., 1980. Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum, Hillsdale, NJ.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Lozar Manfreda, K., Vehovar, V., Hlebec, V., 2004. Collecting ego-centered network data via the web. Metodološki Zvezki. Advances in Methodology and Statistics 1, 231–295.

Marin, A., 2002. Are respondents more likely to list alters with certain characteristics? Implications for name generator data. In: Presented at International Sunbelt Social Networks Conference, New Orleans.

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

Marsden, P.V., 1993. The reliability of network density and composition measures. Social Networks 15, 399–421.

Marsden, P.V., Campbell, K.E., 1984. Measuring ties 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.

Mercator Research Group, 2003. Snap Survey Software: Version 7. Mercator Research Group, http://www.snapsurveys.com/.

Muthén, B., 1989. Latent variable modelling in heterogeneous populations. Psychometrika 54, 557–585.

Muthén, B., 1990. Mean and Covariance Structure Analysis of Hierarchical Data. Statistics Series, vol. 62. UCLA, Los Angeles.

Muthén, B., 1994. Multilevel covariance structure analysis. Sociological Methods & Research 22, 376–398.

Muthén, L.K., Muthén, B., 2004. Mplus, Statistical Analysis with Latent Variables. User's Guide, third ed. Muthén & Muthén, Los Angeles, CA.

Saris, W.E., 1987. Continuous Scales in the Social Sciences. An Attractive Possibility. Sociometric Research Foundation, Amsterdam.

Saris, W.E., 1988. Variation in Response Functions: A Source of Measurement Error in Attitude Research. Sociometric Research Foundation, Amsterdam.

Saris, W.E., 1990. Models for evaluation of measurement instruments. In: Saris, W.E., van Meurs, A. (Eds.), Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies. North Holland, Amsterdam, pp. 52–80.

Saris, W.E., 1995. Designs of models for quality assessment of survey measures. In: Saris, W.E., Münnich, Á. (Eds.), The Multitrait Multimethod Approach to Evaluate Measurement Instruments. Eötvös University Press, Budapest, pp. 9–37.

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.

Schaefer, D.R., Dillman, D., 1998. Development of a standard e-mail methodology: results of an experiment. The Public Opinion Quarterly 62, 378–397.

Schmitt, N., Stults, D.N., 1986. Methodology review. Analysis of multitrait–multimethod matrices. Applied Psychological Measurement 10, 1–22.

Schonlau, M., Zapert, K., Simon, L.P., Sanstad, K.H., Marcus, S.M., Adams, J., Spranca, M., Kan, H., Turner, R., Berry, S.H., 2004. A comparison between responses from a propensity-weighted web survey and an identical RDD survey. Social Science Computer Review 22, 128–138.

Sheehan, K.B., McMillan, S.J., 1999. Response variation in e-mail surveys: an exploration. Journal of Advertising Research 39, 45–54.

Sherpenzeel, A., 1995. A question of quality. Evaluation of survey questions by multitrait–multimethod studies. Doctoral dissertation. University of Amsterdam, Leidschendam, The Netherlands.

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

Totten, J., 2003. Use of e-mail and internet surveys by research companies. www.ijor.org IMRO Journal of Online Research. Available on: 14 April 2003 (http://www.ijor.org/ijor_archives/articles/use_of_email_and_internet_surveys.pdf).

van Doorn, L., Saris, W.E., Lodge, M., 1983. Discrete or continuous measurement. What difference does it make? Kwantitatieve Methoden 10, 105–120.

Vehovar, V., Batagelj, Z., Lozar Manfreda, K., Zaletel, M., 2002. Nonresponse in web surveys. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), Survey Nonresponse. Wiley, New York, pp. 229–242.

Watt, J.H., 1997. Using the internet for quantitative survey research. Quirk's Marketing Research Review, Available on: June 1997 (http://www.quirks.com/articles/article.asp?arg_ArticleId=248).

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

Wolf, F.M., Savickas, M.L., Saltzman, G.A., Wlker, M.L., 1984. Meta-analytic evaluation of an interpersonal skills curriculum for medical students: synthesizing evidence over successive occasions. Journal of Counseling Psychology 31, 253–257.

Yuan, K.H., Bentler, P.M., 2000. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In: Sobel, M.E., Becker, M.P. (Eds.), Sociological Methodology 2000. American Sociological Association, Washington, DC, pp. 165–200.