In survey data, four potential sources of measurement error can jeopardize the results: the respondents, the interviewers, the questions, and the data collection method. In the past two decades, a shift has occurred in the way survey data are collected; telephone surveys and, to a lesser degree, mail surveys are now more extensively used. This has stimulated empirical research on the influence of the data collection method on data quality. Most of these mode comparisons used univariate criteria. In this study we concentrate on the potential influence of the data collection method on two substantive structural models. A controlled field experiment was conducted in which a mail, a telephone, and a face-to-face survey were compared. Using multigroup comparisons, two substantive structural equation models (one for loneliness and one for general well- being) were compared across the different data collection methods. The different data collection methods turned out to produce significantly different covariance matrices. Subsequent analyses showed that structural models, based on these covariance matrices, also differed.

The Influence of Data Collection Method on Structural Models

A Comparison of a Mail, a Telephone, and a Face-to-Face Survey

EDITH D. DE LEEUW
Vrije Universiteit Amsterdam, The Netherlands
GIDEON J. MELLENBERGH
JOOP J. HOX
University of Amsterdam, The Netherlands

Until the 1970s, the face-to-face interview was the accepted method for conducting social science surveys. Since then there has been dramatic progress in the other principal modes of survey implementation, namely, telephone and mail procedures (Dillman

AUTHORS' NOTE: *The data collection and analysis have been partly funded by the Netherlands Organization for Scientific Research (NWO) under grant 500278008. The first author has also contributed in the context of the NESTOR research program "Living Arrangements and Social Networks of Older Adults," conducted by the Vrije Universiteit Amsterdam and the Netherlands Interdisciplinary Demographic Institute in The Hague. The authors thank Dr. Godfried L. H. van den Wittenboer and two anonymous reviewers for their comments.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 24 No. 4, May 1996 443-472

1978; Groves 1989). This can be attributed to two important factors. First, the fast growing costs of the face-to-face interview and the growing nonresponse rates for this method have led survey researchers to consider "alternative" data collection procedures (Dillman 1978; Dillman and Tarnai 1988; Groves and Kahn 1979; Goyder 1985). Second, research in the past two decades shows that mail and telephone surveys have far greater potential than had previously been thought and that high response rates can be attained with lower survey costs and fewer time constraints (Baumgartner and Heberlein 1984; Dillman 1991; Lyberg and Kasprzyk 1991).

Still, the increased acceptance of alternatives to the face-to-face interview was limited, pending further demonstrations that the data quality would not suffer. As a result, the influence of the data collection method on the quality of the data has received considerable attention in survey research. To systematically summarize the knowledge in this field, a meta-analysis of comparative studies on survey data collection methods was performed. Several indicators of data quality were used in the original studies reviewed, reflecting the complexity of this construct. Indicators used were response validity (record check), item nonresponse, social desirability, and number of statements to open questions. The main conclusions were as follows:

1. Face-to-face interviews result in data with slightly less item nonresponse and slightly more statements to open questions. No differences were found concerning response validity and social desirability.
2. Self-administered questionnaires in comparison to both face-to-face and telephone interviews resulted in less social desirability and more self-disclosure, especially when sensitive questions are asked; however, compared to both interview methods, self-administered questionnaires also resulted in more item nonresponse (De Leeuw 1992).
3. On average, face-to-face interviews achieve the highest response (70%), telephone interviews the next highest (67%), and mail surveys the lowest (61%). However, the average response to face-to-face and telephone interviews is going down over the years, whereas the average response to mail surveys is at least stable (Hox and De Leeuw 1994).

In the studies reviewed, comparisons of data collection methods were found to be mainly restricted to the analysis of univariate

distributions, and the meta-analysis found that different modes indeed produce small differences in the univariate distributions (De Leeuw 1992). However, in the literature reviewed little attention had been given to the potential effect of mode of data collection on the estimates of the relationships between variables. This last topic will be the focus of this article.

Two rival hypotheses can be formulated about the effect of the data collection method on the estimated relationships between variables. The first one states that even if mode effects may exist when univariate statistics are compared, this does not necessarily imply an effect on multivariate statistics, such as covariances. The reasoning is that the observed differences between the marginals of the univariate distributions just reflect a shift of position of a specific variable on the x- or y-axis; however, the shape of the bivariate distribution of any two variables—as reflected in the bivariate scatter plot—will not be altered. This idea is sometimes called the "form-resistant correlation hypothesis" and rarely has been tested empirically (but see Krosnick and Alwin 1988; Schuman and Presser 1981, p. 4). This reasoning leads to the hypothesis that, if mode effects are detected in marginal distributions, multivariate statistics will remain comparatively stable. There is some support for this hypothesis in the survey literature, which finds that the net effects of data collection mode are usually small (for a discussion, see Groves 1989, pp. 501-20). The second hypothesis derives from statistical distribution theory, which states that, in general, higher order moments are less stable than first-order moments. The implication is that a few outliers in a specific data set can cause a dramatic change in statistics based on higher order moments such as covariances and correlations. This reasoning leads to the hypothesis that, if mode effects are detected in marginal distributions, multivariate statistics are expected to show larger effects. There is some support for this hypothesis from the statistical literature about the instability of higher order moment matrices used in distribution-free methods for covariance structure analysis (cf. Browne 1984).

Which hypothesis is more likely to be true remains to be seen. There is no consensus on this subject. For instance, an informal survey conducted by the authors among 115 experts in the fields of data collection methods and/or multivariate analysis revealed some belief in the first hypothesis that states that multivariate mode effects are

smaller (52% favored this hypothesis). However, there were large differences in the expressed opinions; 20% thought both hypotheses equally likely and 28% favored hypothesis 2 (multivariate mode effects are larger).

In this study we investigate the potential influence of data collection method on the parameter estimates of two substantive structural equation models: a model about loneliness and a model about subjective well-being. Two different aspects of structural modeling are investigated: The loneliness model is a causal model of the determinants of loneliness, and the subjective well-being model is a confirmatory factor model of the structure of well-being.

DATA COLLECTION

In the autumn of 1989 a controlled field experiment was conducted in the Netherlands. The modes of data collection investigated were the mail questionnaire, the telephone interview, and the face-to-face interview. In criticizing mode comparisons it often has been noted that mail and telephone surveys are very limited regarding type, format, and number of questions asked and that therefore only very restricted surveys have been compared. For a fair and meaningful comparison a diverse questionnaire should be used; in constructing this questionnaire, we tried to push the mail and telephone survey to their limits. At the same time care was given to the validity of the experiment. To optimize the internal validity of the experiment without jeopardizing the external validity, it is important to implement the survey procedures realistically in terms of general survey practice, while controlling the influence of extraneous variables as best as possible. To fulfill this goal, decisions had to be made concerning the construction of the questionnaire, the sample used, the interviewers, and so on. These decisions will be discussed shortly below.

THE QUESTIONNAIRE

The subject of the questionnaire was well-being. The questionnaire included potentially sensitive questions regarding subjective phenomena such as loneliness and happiness, in combination with more

factual questions on objective attributes such as financial situation, labor force participation, and extension of social network. Also standard biographical information was asked. Different question formats were included: checklists, open questions, and closed questions. The latter differed in number of answer categories, ranging from two to seven categories. The questions also varied in question threat and saliency.

To investigate mode influences on multivariate relationships and models, specific questions were included on the basis of two well-documented conceptual models for loneliness and well-being (cf. De Jong-Gierveld 1987; Burt, Wiley, Minor, and Murray 1978). Therefore, the questionnaire contained a multiple-item scale, consisting of 11 questions on loneliness (De Jong-Gierveld and Kamphuis 1985), a Dutch form of Rosenberg's self-evaluation scale, consisting of 8 questions (Rosenberg 1979); and a balanced extension of Bradburn's affect balance scale, consisting of a positive-affect scale and a negative-affect scale. Both the multi-item positive-affect scale and the multi-item negative-affect scale contain 9 questions (Bradburn 1969; Hox 1986).

A first version of the questionnaire was drafted following the basic rules of question writing as formulated by, among others, Converse and Presser (1986), Keenan and Mauch (1986), and Sudman and Bradburn (1982). This draft version was pretested by using cognitive interview methods (cf. Forsyth and Lessler 1991). Special attention was given to the understanding of the questions and terms used. The revised version was then used to develop three equivalent versions, one for each data collection method. An iterative procedure was used in which experts on each data collection method optimized the questionnaire for each method (e.g., layout in mail surveys, interviewer instructions in interview schedule); this optimization was followed by group discussions on changes and comparability of questions. Following the rules of the Delphi method (Sackman 1974), the process was stopped after consensus was reached. A major decision was to use response cards with answer categories in the face-to-face interview for all checklists and closed questions with five or more answer categories. In the telephone version, special interviewer instructions were added about reading and repeating answer categories. The final three equivalent versions were field-tested in a pilot study, in which

the implementation of the total data collection procedure was also tested. No changes in the questionnaire were necessary. The final questionnaire contained 82 questions; the mean completion time (time from first to last question) was 31 minutes for the face-to-face interview and 24 minutes for the telephone interview.[1]

IMPLEMENTATION

The data were collected by using three data collection methods: a mail survey, a (paper-and-pencil) telephone interview, and a (paper-and-pencil) face-to-face interview. Care was taken to implement each data collection method as optimally as possible. In the mail survey condition, Dillman's Total Design Method (TDM; Dillman 1978; De Leeuw and Hox 1988) was followed completely, including a third and last reminder by certified mail. Twenty specially trained interviewers conducted the interviews; besides a standard interview training an additional training was given in telephone techniques. Ten randomly assigned interviewers started with the telephone interviews and then conducted face-to-face interviews; the other 10 started with face-to-face interviews. The training and supervision of the interviewers was successful; when interviewer data were analyzed, only small interviewer effects were found, which did *not* differ between the telephone and the face-to-face mode (Hox, De Leeuw, and Kreft 1991).

A random stratified sample from the telephone directory of the Netherlands was used. Stratification was according to urbanization. The reason for using the telephone directory as a sampling frame is that it is one of the rare lists in the Netherlands that give both name and address and cover the whole country. The telephone directory is fairly efficient as a sample frame; the Netherlands have a high telephone coverage, which is comparable to Japan, Germany, and the United States (cf. Trewin and Lee 1988). Furthermore, according to Dutch Telecom, approximately 92% of all private households have a telephone, and only approximately 8% of all private numbers are unlisted. These unlisted numbers are not recorded, either by intention of the telephone owner or because it is a new number that has not yet been entered in the directory.

On the basis of this sampling frame, a stratified random sample of addresses was taken for each data collection mode. Businesses and institutions were dropped. On each address, a respondent 18 years or older was selected by using the next-birthday method. This method is nonintrusive, does not take much time, and is fairly effective (cf. Oldendick, Bishop, Sorenson, and Tuchfarber 1988). Furthermore, the next-birthday method can be implemented without difficulties in both mail surveys and face-to-face and telephone interviews. All sampled addresses received an advance letter. The interviewers used a flexible script in asking for respondent cooperation. In both the telephone survey and the face-to-face survey condition the request for cooperation was made by telephone; at least seven callbacks were made to reach sample units. In the mail survey after the advance letter, a questionnaire was sent, followed by three reminders. The last reminder was sent by certified mail according to the basic Dillman system (Dillman 1978). No attempts were made to convert definite refusals; refusers were not called back by special interviewers.

RESPONSE

The initial sample sizes were 400 (mail survey), 530 (face-to-face), and 450 (telephone interview). The response rate was calculated as the percentage of completed interviews or questionnaires for all eligible cases (including noncontacts). The mail survey resulted in a final response rate of 68% or 254 completed questionnaires. The face-to-face interview had a response rate of 51% (243 completed questionnaires), and the telephone interview had a response rate of 66% (266 completed questionnaires). The difference in response rate was almost entirely due to more explicit refusals in the face-to-face condition.

In the United States and Canada, in general, higher response rates are found for face-to-face surveys than for telephone surveys (for an overview see Goyder 1987). This is not the case in the Netherlands. In the 1980s, Statistics Netherlands still reported no marked differences between face-to-face and telephone surveys (Kerssemakers 1985); in the 1990s, on the other hand, higher response rates for the telephone survey are generally found (Snijkers 1992). This is in line with the results of a large-scale research project on the willingness of

the Dutch to cooperate in surveys; when asked, people are far more willing to cooperate to a request for a telephone interview than to a request for a comparable face-to-face interview (Louwen 1992).

Nonresponse, especially the relatively large nonresponse in the face-to-face interview, could be a potential source of bias. We could check this, because external information was available on both respondents and nonrespondents. For the complete initial sample (respondents and nonrespondents) detailed background information was available based on the Dutch zip code system (collected by GEO-marktprofiel); this information is mainly based on administrative records. The Dutch zip codes form an extremely fine grid with on average 15 households per zip code. So, aggregated information was available on, for instance, socioeconomic status (SES), income, type of household, and type of community for clusters of, on average, 15 households each.

Using this auxiliary zip code information, we investigated the possibility of selective nonresponse. Respondents and nonrespondents did differ slightly in affluence. Nonrespondents more often lived in the big cities, in rented houses, and had a lower income. Respondents, on the other hand, lived more often in rural areas, owned their homes, and belonged to the middle- and higher-income class. However, no interaction effects with mode of data collection were observed: Although the response rates differ, respondents and nonrespondents do *not* differ across the modes on the available zip code-based information.

In addition, further analyses were done to investigate whether the individuals who did respond differed in important background variables across modes. When the respondents in the three conditions were compared on their answers to the sociodemographic questions in the questionnaire, the only statistically significant differences observed over modes concerned the variables gender and marital status. Slightly more women responded in the face-to-face condition, whereas in the mail condition relatively more respondents were male. In the telephone condition the sexes were equally distributed. Furthermore, more married persons responded to the mail survey and more widowed and divorced people responded to both interview surveys. Comparisons with figures on the general population from Statistics Netherlands

showed that women were overrepresented in the face-to-face survey and that there was a general overrepresentation of unmarried individuals in all modes.[2]

It is important to note that no significant differences were detected across modes for important variables such as age, education, and previous interview experience (testing was done at the 5% level). For more details on data collection procedures and nonresponse, see De Leeuw (1992).

MODELS AND ANALYSIS METHOD

We used two structural equation models to investigate the effect of data collection method on estimated relationships: a causal path model for loneliness and a confirmatory factor model for well-being. For our goal the substance of the models is not of primary importance. However, because we wanted to compare parameter estimates across the three data collection methods, it is important that both models have a satisfactory overall fit. To achieve this, we chose existing models for which earlier publications had established a good fit in moderately large samples.

THE LONELINESS MODEL

The first model, a causal structural equation model about the determinants of loneliness, is derived from De Jong-Gierveld (1987). This model has four exogenous variables (living alone, extension of social network, self-evaluation, and age) and two endogenous variables (evaluation of social network and loneliness).

The exogenous variable "living alone" indicates whether respondents live with important others. This variable is empirically scaled on the basis of responses to questions about the living arrangements of the respondents. The scale values range from 1 (*living with more than one important other*) to 3 (*living completely alone*). "Extension of social network" is measured by asking respondents to state the number of persons who are very important to them. This variable has a minimum value of 0. "Self-evaluation" is measured as the sum score on a multi-item scale consisting of eight questions. The minimum

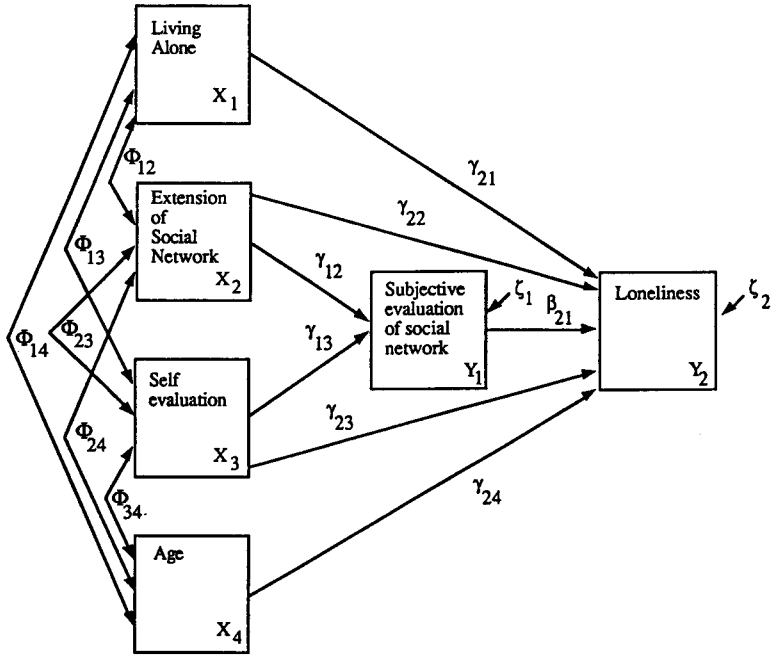

Figure 1: Loneliness Model

score is 0, the maximum score (*very positive self-evaluation*) is 8. "Age" is measured in years.

The endogenous variable "evaluation of social network" was measured with a 5-point closed question about the degree of satisfaction with social relationships; the value 1 indicates that the respondent is very dissatisfied, the value 5 means very satisfied. "Loneliness" is measured as the sum score on an 11-item scale; the minimum score is 0, the maximum score (*extreme loneliness*) is 11.

In our model, loneliness is negatively determined by the extension of the social network (number of important relationships), the amount of satisfaction with the social network, and a positive self-evaluation. Loneliness is (positively) determined by living alone and age (see also Figure 1). The loneliness model is a path model with observed variables only.

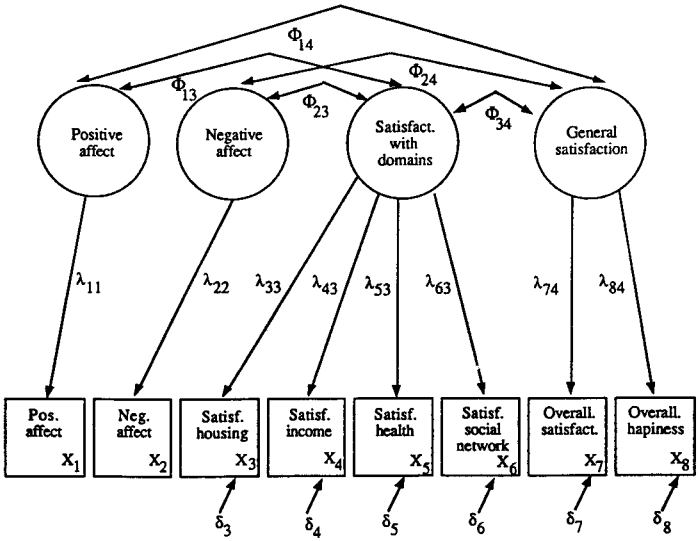

Figure 2: Well-Being Model

THE WELL-BEING MODEL

The second model is a confirmatory factor analysis model for the structure of well-being, derived from Burt et al. (1978) and Burt, Fischer, and Christman (1979). Burt et al. (1978) discuss a series of models for the structure of well-being. Our model is derived from their final model, which showed a satisfactory fit. Four dimensions are distinguished: general satisfaction, satisfaction with specific domains, positive affect, and negative affect (see also Figure 2).

The general satisfaction dimension is measured by two global variables: general happiness as indicated on a seven-step ladder, ranging from 1 (*worst that could happen*) to 7 (*best that could happen*); and satisfaction with life in general as indicated by the response on a 5-point closed question (1 = *very dissatisfied*, 5 = *very satisfied*). The satisfaction with the specific-domains dimension is measured by four variables on satisfaction with certain domains of life (i.e., housing, income, health, and social network). Again, closed questions were used with five response categories, ranging from 1 (*very dissatisfied*) to 5 (*very satisfied*). The positive-affect dimension is measured as the

sum score on a multi-item scale consisting of nine questions; the minimum score is 0, indicating the absence of any feelings of positive affect (happiness), the maximum score 9 (*extremely happy*). The negative-affect dimension is also measured as the sum score on a nine-item scale; the minimum score is 0, indicating the absence of any feelings of negative affect; the maximum score is 9. The positive- and negative-affect dimensions are uncorrelated (cf. Bradburn 1969; Hox 1986).

The original well-being model, as published by Burt et al. (1978), is not identified. For a discussion of possible restrictions to make the model identifiable, see Burt et al. (1979). In our model, we used the following restrictions: The variance of the factors was fixed at 1.00, and the measurement error variances of the observed variables positive and negative affect were fixed at zero.

ANALYSIS METHOD

For each model, the following analysis strategy was used. To begin, we tested whether the covariance and correlation matrices differed across the three data collection methods. Two distinct approaches were used to assess this. In the first one, we performed a multigroup covariance structure analysis with equality constraints between all groups (Bollen 1989, chap. 8; Jöreskog and Sörbom 1989, chap. 9). If this model is rejected by the usual chi-square test, we may conclude that there are systematic differences between the three groups. This statistical test is not without problems. Basically, it tests the appropriateness of between-groups equality constraints for a saturated model. Because the model is not very specific, the power of the statistical test may be low (cf. Byrne 1989). Also, the maximum likelihood test used assumes multivariate normal distributions. As the equality of the covariance structure across data collection modes is the central problem in our study, we decided to test it also by a second, more direct approach. This second approach is a permutation test using Monte Carlo simulation.[3] The question is whether the differences between the three data collection modes are larger than the expected sampling variability. We investigated this by dividing the total observed sample at random into three groups, with group sizes equal to the observed group sizes. Next, a covariance structure analysis was performed to

test the equality of the covariance and correlation matrices across these groups. This process was repeated 1,000 times. The resulting 1,000 chi-square values form an empirical null distribution to which the chi-square of the model test for the actual groups can be referred.[4] This approach corresponds to the randomization tests proposed by Edgington (1987) for univariate group comparisons; for a comparison of permutation tests with other computer intensive methods such as bootstrap sampling, see Noreen (1989) and Good (1994). An advantage of the permutation test is that it is not limited to the chi-square as an indication of the goodness of fit. We also included the Normed Fit Index (NFI) by Bentler and Bonett (1980; cf. Bollen 1989, p. 270) in our simulation.

If the global test rejected the hypothesis of equal covariance or correlation matrices across groups, it was followed by a series of multigroup analyses to investigate the differences between the mail, the telephone, and the face-to-face survey. We started with the strictest model in which each parameter in the model is assumed to be invariant across all three groups (i.e., the mail, the telephone, and the face-to-face survey). In this model, the measurement error variances were all fixed at zero.

In the well-being model, multiple observed variables were available for the latent variables "general satisfaction" and "satisfaction with specific domains." This made it possible to test a model that allows the estimated variances of the measurement errors for these variables to differ across groups.

The next model includes information about the reliability of the multiple item scales in the model. Preliminary analyses indicated that the reliability of multiple item scales showed small differences across data collection methods: The mail survey had the most reliable results, whereas the telephone survey was the least satisfactory (De Leeuw 1991). Therefore, in this step, we allowed differences in variances of measurement errors between the groups, using reliability estimates under the congeneric test model for the multiple item scales (Bollen 1989, p. 168).

The most consistent finding in earlier studies comparing responses in face-to-face and telephone interviews is the lack of differences in (univariate) results obtained through these two methods (Groves 1989, p. 551). The main (univariate) differences detected are between mail

surveys, on one hand, and interview surveys (both telephone and face-to-face) on the other hand (De Leeuw 1992, p. 77). Therefore, in the next step, invariance restrictions between groups were imposed only on parameter estimates for the two interview modes (face-to-face and telephone). The model for the self-administered mail survey group was only restricted to have the same pattern of free parameters and fixed elements as the corresponding model for the two interview groups; however, the free parameters in the mail survey group were allowed to differ from the values in the other groups (cf. Jöreskog and Sörbom 1989, pp. 229-38). Subsequently, in the well-being model, it was investigated if allowing for different measurement errors in the two interview modes improved the fit further. Finally, for all three groups the only restrictions concerned the form (same dimensions and patterns); all parameter estimates were allowed to differ in the three groups.

To compare subsequent models, the overall chi-square and the NFI were calculated (Bentler and Bonett 1980). Furthermore, in most cases, the subsequent models are nested within each other. For two nested models the difference in chi-squares is again chi-square distributed with a *df* equal to the difference in *df* for the two models. This makes it possible to test if the improvement of fit is statistically significant.

RESULTS

For each model, the covariance matrix of the observed variables was computed for each data collection method (mail, telephone, and face-to-face survey). Table 1 presents the asymptotic results in conjunction with the results of the Monte Carlo permutation test.

The asymptotic chi-square test rejects the hypothesis of equal covariance or correlation matrices in all cases. The *p* values based on the permutation test are the proportion of simulated chi-squares that were larger than the actual chi-square and the proportion of NFI indexes that were lower than the actual NFI. The largest *p* value is .02 for the permutation test performed on the NFI for equal correlation matrices for the happiness data. This is still significant at the .05 level, and the value of .89 for the NFI of the equal correlations model is

TABLE 1: Degrees of Freedom (*df*), Chi-Square (χ^2), Normed Fit Index (NFI), Asymptotic *p* Value for χ^2, Permutation *p* Values for χ^2 and NFI for Models Specifying Equal Covariance (Σ) and Correlation (R) Matrices for the Loneliness and Happiness Data

Model	df	χ^2	*NFI*	*Asymptotic* p *for* χ^2	*Permutation* p *for* χ^2	*Permutation* p *for* *NFI*
Lonely equal Σ	56	115	.81	.00	.00	.00
Lonely equal R	42	84	.86	.00	.00	.00
Happy equal Σ	72	166	.82	.00	.00	.00
Happy equal R	56	102	.89	.00	.01	.02

below the value of .90 that is usually suggested as a criterion for goodness of fit (Bollen 1989). We conclude that the permutation tests also reject the hypothesis of equal covariance or correlation matrices in all cases.

THE LONELINESS MODEL

Because the global hypothesis of equal covariance matrices is rejected, it is not surprising that the strictest model did not fit. This model (model 1) constrains all parameter estimates (path coefficients and residual variances) to be equal across all three groups.

In model 1, the measurement error variances were all fixed at zero. In the next model (model 2), estimates of the measurement error variance of the multiple item scales (loneliness and self-evaluation) were set in the error-variance matrices; on the basis of the reliability estimates under the congeneric test model, different values were used for each data collection group. This did not improve the fit of the model, and the subsequent models do not include these estimates of the measurement errors.

In the next step (model 3), all parameters are constrained to be invariant for the face-to-face and the telephone interview group. In the mail survey group, the parameter matrices are constrained only to have the same dimensions and patterns of fixed and free elements as in the two interview groups. This model has a reasonable fit (see Table 2). Because model 3 is nested in model 1, the difference in chi-squares

TABLE 2: Chi-Square (χ^2), Degrees of Freedom (*df*), *p* Value, and Normed Fit Index (NFI) of Three-Group Path Model for Loneliness

Model	χ^2	df	p *Value*	*NFI*
(1) Mail = FtF[a] = Tel[b]	39.8	24	.02	.93
(2) Mail = FtF = Tel/α	39.4	24	.02	.93
(3) Mail $\approx$ FtF = Tel	24.3	15	.06	.96
(4) Mail $\approx$ FtF $\approx$ Tel	6.4	6	.38	.99

NOTE: = indicates that all parameters in this model are invariant over groups, $\approx$ indicates the weaker same-pattern restriction, and /α indicates that in this model the measurement error variance for the variables loneliness and self-evaluation is set according to their reliability.
a. FtF = face-to-face.
b. Tel = telephone.

can be used to test whether the increase in fit is statistically significant. The difference in chi-squares between model 1 and model 3 turns out to be not significant ($p = .08$).

In the final step (model 4), the restrictions are freed even further. In model 4, the only constraints are on the pattern of the parameter matrices. The same dimension and pattern are demanded, without restricting any of the nonfixed parameters to have the same *value* across groups. Model 4 shows an almost perfect fit. Compared to model 1, the fit is significantly better ($p = .02$). Also, compared to model 3, the fit of model 4 is better ($p = .04$). For an overview of the model fit, see Table 2.

From the results in Table 2 we conclude that the model that constrains both interview methods to be invariant and constrains the mail method to have only the same pattern is acceptable; however, the model that allows the two interview methods to vary as well fits the data better. If we examine the fit of the models for each survey condition under all four models, the conclusion is that model fit problems are generally somewhat larger in the face-to-face condition.

When comparing parameter estimates across groups, the unstandardized parameter estimates are preferred (Bollen 1989, p. 126). The unstandardized parameter estimates for the least restrictive model (model 4) are given in Table 3.

To interpret the relative importance of the parameter estimates correctly, it is essential to keep in mind the scale on which the variables are measured. For loneliness, the minimum score is 0 and the maximum

TABLE 3: Unstandardized ML[a] Estimates Three-Group Path Model (Mail ≈ FtF[b] ≈ Tel[c]) and Squared Multiple Correlations Endogenous Variables for Loneliness Model

	Method		
Parameter	*Mail*	*Face-to-Face*	*Telephone*
β_{21}	–2.11 (0.17)	–1.29 (0.16)	–1.37 (0.19)
γ_{21}	0.55 (0.33)	0.51 (0.30)	0.76 (0.30)
γ_{22}	–0.29 (0.10)	–0.30 (0.11)	–0.23 (0.12)
γ_{12}	0.08 (0.04)	0.15 (0.04)	0.05 (0.04)
γ_{13}	0.09 (0.03)	0.10 (0.03)	0.05 (0.03)
γ_{23}	–0.18 (0.07)	–0.28 (0.07)	–0.37 (0.07)
γ_{24}	0.00 (0.01)	0.03 (0.01)	–0.00 (0.01)
Ψ_{11}	0.75 (0.07)	0.83 (0.08)	0.62 (0.06)
Ψ_{22}	4.58 (0.44)	4.58 (0.43)	5.33 (0.48)
	Squared Multiple Correlation		
Evaluation of:			
Social network	.08	.11	.02
Loneliness	.52	.41	.29

NOTE: The standard errors of the parameter estimates are given in parentheses.
a. ML = maximum likelihood.
b. FtF = face-to-face.
c. Tel = telephone.

score is 11; the self-evaluation score ranges from 0 to 8. The variable "living alone" ranges from 1 to 3. Extension of the social network is a count of the number of important relations with a minimum of 0. Age is measured in years. Satisfaction with social network is measured on a single 5-point scale.

Most of the parameter estimates in Table 3 are similar in direction and significance (from zero). The most important difference is the explained variance for the dependent variable "loneliness": .52 in the mail condition, .41 in the face-to-face condition, and .29 in the telephone condition. The overall test rejects the null hypothesis that all parameter estimates may be considered equal across groups. To interpret the differences between specific parameter estimates, the following (conservative) rule was used: A difference in parameter estimates between modes is evaluated as substantial if that difference is larger than twice the largest standard error for that specific parameter.

Inspection of Table 3 shows that the major differences between data collection methods occur for the parameters γ_{23} (effect of self-evaluation on loneliness), γ_{24} (effect of age on loneliness), γ_{12} (effect of extension of social network on the subjective evaluation of social network), and β_{21} (effect of subjective evaluation of social network on loneliness).

These differences are large enough to influence the interpretation of social science results. In substantive research, the parameter estimates are often standardized to facilitate interpretation. Figure 3 contains the graphic presentation and parameter estimates for model 4 (same pattern for each data collection method), with the observed variables standardized to a common metric for the three groups. (This standardization preserves the comparability across groups; cf. Jöreskog and Sörbom 1989, p. 238.) Again, there are some apparent differences, the most important being the size of the path coefficient from "self evaluation" to "loneliness" across all methods, the size of the path coefficient of "age" on "loneliness" in the face-to-face interview, the size of the path coefficient of "extension of social network" on "evaluation of that network" in the face-to-face interview, and the size of the path coefficient of "evaluation of network" to "loneliness" in the mail survey.

THE WELL-BEING MODEL

The results for the well-being model closely follow the results for the loneliness model. Using both the asymptotic chi-square test and the permutation test (Table 1), the covariance and correlation matrices were significantly different at the 5% level in the three groups. The strictest model (model 1), which constrains all parameter estimates to be equal across the three data collection groups, did not fit.

For the two latent variables "domain satisfaction" and "general satisfaction," more than one observed variable was available. In the next model (model 2), the measurement error variances of the observed variables for these factors were estimated separately in the three groups. This resulted in a model that fits much better than the first model ($p = .00$), although the overall fit is still not good (see also Table 4).

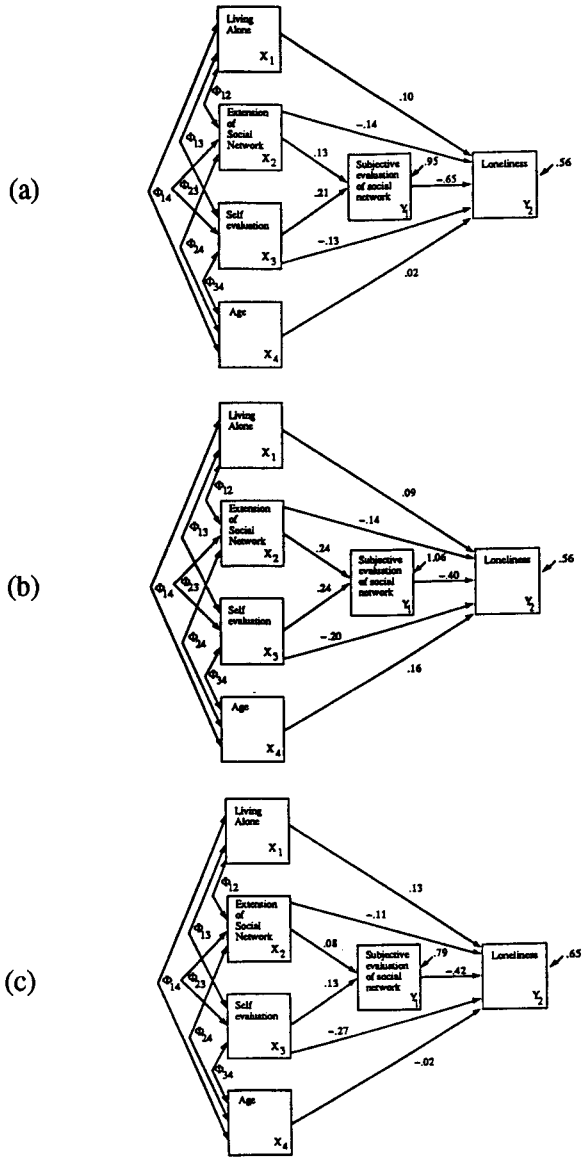

Figure 3: Standardized Parameter Estimates Loneliness Model (Model 4) for (a) Mail Survey, (b) Face-to-Face Interview, and (c) Telephone Interview

TABLE 4: Chi-Square (χ^2), Degrees of Freedom (*df*), *p* Value, and Normed Fit Index (NFI) of Three-Group Factor Model for Well-Being

Model	χ^2	df	*p Value*	*NFI*
(1) Mail = FtF[a] = Tel[b]	220.1	89	.00	0.76
(2) Mail = FtF = Tel/δ	149.1	77	.00	0.84
(3) Mail = FtF = Tel/δ + α	148.6	77	.00	0.84
(4) Mail ≈ FtF = Tel	131.1	70	.00	0.86
(5) Mail ≈ FtF = Tel/δ	117.6	64	.00	0.87
(6) Mail ≈ FtF = Tel/δ + α	117.2	64	.00	0.87
(7) Mail ≈ FtF ≈ Tel	93.0	51	.00	0.90

NOTE: = indicates that the parameters in this model are invariant over groups, ≈ indicates the weaker same-pattern restriction, /δ indicates that in this model measurement error variances are estimated separately in the three groups. δ + α indicates that in addition the measurement error variance for the variables positive and negative affect is set according to their reliability.
a. FtF = face-to-face.
b. Tel = telephone.

Including the measurement error variances for the observed scores on the multi-item scales "positive affect" and "negative affect," by setting the error variances according to the reliability estimates (model 3), did not improve the fit.

In model 4, all parameters are constrained to be invariant for the face-to-face and the telephone interview group only. In the mail survey group, the parameter matrices are constrained only to have the same dimensions and patterns as in the two interview groups. This model fits better than model 2 and 3, which constrain the factor loadings and correlations but allow the measurement errors to differ across all groups (see Table 4).

In the next two steps, we again allowed differences in measurement errors. In model 5, we allowed differences in the variances of the measurement errors δ of the observed variables for domain satisfaction and general satisfaction. This resulted in a significantly better fit than model 4 ($p = .04$). Model 5 can also be compared statistically with model 2, which also allows for different measurement errors across groups but constrains all other parameter estimates to be equal. Model 5 fits significantly better than model 2 ($p = .00$). Using reliability estimates, model 6 also estimates the fixed error variances of observed positive and negative affect. This did not improve the fit.

In the final step (model 7), the only constraints are on the pattern of the parameter matrices. The same dimension and patterns are assumed, without equality restrictions on any of the parameter estimates. Compared to model 2 (identical loadings and correlations, different measurement errors), the fit is significantly better ($p = .00$). Compared to model 4 (restrictions across face-to-face and telephone conditions only) and model 5 (restrictions across face-to-face and telephone conditions, but different measurement errors), the fit of model 7 is also better (p values .00 and .03). Even in model 7 the overall fit is still not quite satisfactory. However, the relative size of the chi-square and the degrees of freedom ($\chi^2/df = 1.82$) and the value of .90 for the NFI suggest that this model could be considered acceptable. Post hoc model exploration to achieve a better fit would lead to different models for the three data collection methods. We will return to this point in our discussion.

To compare the parameter estimates across the three data collection methods, the unstandardized parameter estimates for model 7 are given in Table 5. To interpret the relative importance of the parameter estimates, it is important to know the scale on which the variables are measured. Positive and negative affect are measured by two nine-item scales, with a range from 0 (*lowest score*) to 9 (*highest score*). The domain satisfactions and global satisfaction variables are measured by single 5-point questions. Global happiness is measured with one closed question with seven response categories.

Most of the parameter estimates in Table 5 are similar in sign and direction. To evaluate the differences between the parameter estimates, we again adopted the decision rule that a difference in parameter estimates was evaluated as substantial if that difference was larger than twice the largest standard error for that specific parameter. Relatively large differences between the groups are found for the loadings of the observed variables housing satisfaction and social network satisfaction (λ_{33} and λ_{63}) on the domain satisfaction factor. Smaller, but still substantial differences (more than twice the largest standard error) are found for the loadings of the positive-affect scale on the positive-affect factor (λ_{11}) and for the variable overall satisfaction on the global satisfaction factor (λ_{74}). Furthermore, it should be noted that the correlations of the satisfaction factor (factor 3) with the

TABLE 5: Unstandardized ML[a] Estimates for Three-Group Factor Model (Mail ≈ FtF[b] ≈ Tel[c])

	Method		
Parameter	*Mail*	*Face-to-Face*	*Telephone*
λ_{11}	2.29 (0.11)	2.01 (0.10)	1.81 (0.09)
λ_{22}	2.14 (0.10)	2.25 (0.11)	2.07 (0.10)
λ_{33}	0.33 (0.07)	0.23 (0.07)	0.09 (0.07)
λ_{43}	0.42 (0.07)	0.28 (0.08)	0.34 (0.09)
λ_{53}	0.27 (0.06)	0.27 (0.08)	0.25 (0.08)
λ_{63}	0.41 (0.06)	0.65 (0.10)	0.21 (0.07)
λ_{74}	0.60 (0.04)	0.54 (0.06)	0.47 (0.05)
ϕ_{13}	0.56 (0.09)	0.39 (0.09)	0.35 (0.15)
ϕ_{23}	–0.62 (0.09)	–0.41 (0.09)	–0.40 (0.15)
ϕ_{14}	0.45 (0.05)	0.39 (0.07)	0.42 (0.07)
ϕ_{24}	–0.46 (0.05)	–0.52 (0.07)	–0.40 (0.08)
ϕ_{34}	1.13 (0.09)	0.68 (0.11)	1.21 (0.25)
$\theta\delta_3$	0.92 (0.09)	0.69 (0.07)	0.95 (0.09)
$\theta\delta_4$	0.88 (0.09)	0.92 (0.09)	0.83 (0.09)
$\theta\delta_5$	0.69 (0.06)	0.78 (0.08)	0.91 (0.09)
$\theta\delta_6$	0.64 (0.06)	0.54 (0.11)	0.54 (0.05)
$\theta\delta_7$	0.12 (0.02)	0.23 (0.05)	0.28 (0.04)
$\theta\delta_8$	0.53 (0.08)	1.23 (0.15)	1.22 (0.16)

NOTE: The standard errors of the parameter estimates are given in parentheses.
a. ML = maximum likelihood.
b. FtF = face-to-face.
c. Tel = telephone.

other factors show some differences over the groups (ϕ_{13}, ϕ_{23}, ϕ_{34}). The latter even shows two values outside the permitted range, which again indicates that there are still problems with model 7.[5]

Figure 4 contains the graphic representation of model 7 and presents the completely standardized solution (standardized to a common metric) for the purpose of interpretation. Again, there are some sizable differences.

SUMMARY AND DISCUSSION

To investigate the potential influence of data collection method on the estimates of relationships between variables, we compared two

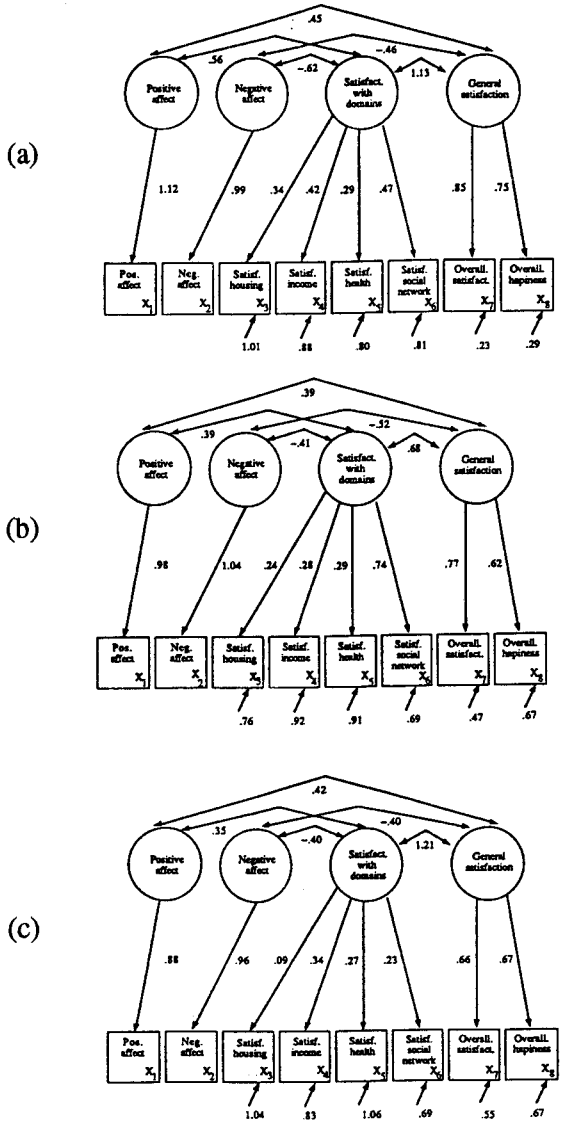

Figure 4: Standardized Parameter Estimates Well-Being Model (Model 7) for (a) Mail Survey, (b) Face-to-Face Interview, and (c) Telephone Interview

substantive structural-equation models across different data collection methods: a loneliness model and a well-being model. The loneliness model analyzed in this study is a path model with four exogenous variables (living alone, extension of social network, self-evaluation, and age) and two endogenous variables (evaluation of social network and loneliness). The well-being model analyzed is a confirmatory factor analysis model with four latent factors (positive affect, negative affect, domain satisfaction, and general satisfaction) measured by eight observed variables.

Two rival hypotheses were investigated. The first hypothesis states that although small mode effects are in general found on marginal distributions of variables, the multivariate estimates will remain stable (form resistant correlation hypothesis). The second hypothesis states that if (small) mode effects are found in marginal distributions, multivariate statistics will show even larger effects (instability of higher order moments hypothesis). Our data supported the second hypothesis.

For both the loneliness model and the well-being model the strictest statistical model was rejected; this model assumes invariance of all parameters over the three groups (i.e., the mail, the telephone, and the face-to-face survey). For the loneliness model, a model specifying equal parameter estimates for the two interview methods leads to an adequate fit, but a model that allows different estimates for all three methods fits better. For the well-being model, a model that allows different estimates for all three models has a barely acceptable fit, judging from the NFI. For both loneliness and well-being we adopted the "same pattern-different estimates" model as the final model for interpretation.

The difference between the estimates for the three data collection methods could follow from differences in the marginal distributions of the variables involved, especially concerning the higher moments (kurtosis and skewness). However, these differences are small and do not explain the different estimates across the three data collection methods.[6]

The principal issue is, of course, whether using different data collection methods might lead to substantively different conclusions about loneliness or well-being. In other words, suppose that three hypothetical investigators each did a study on loneliness or well-being, starting by taking the same model from the literature and each using

a different data collection method. Would they reach the same conclusions? The most "form-resistant" model in our study was the loneliness path model (cf. the parameter estimates in Table 3 and Figure 3). Two important aspects of this model are the amount of loneliness variance explained and the importance of the various path coefficients. In both the mail survey and the face-to-face interview group the proportion variance explained was relatively high (.52 and .41); in the telephone condition this figure was only .29 (cf. Table 3). Thus the same variables explain far less variance in the telephone survey condition. Also, the importance of the individual predictors varies considerably across data collection method (cf. Figure 3). In all three methods, evaluation of social network is the most important determinant of feelings of loneliness, but a striking difference is found when age is considered. In the face-to-face condition, age is significant and a relatively important determinant of loneliness; in the other two conditions, the path coefficient for age is virtually zero and not significant.[7] Similar differences between the three methods arise with the path coefficients for living alone on loneliness and extension of the social network on subjective evaluation of the network. Social scientists who are speculating about the importance of explanatory variables or about the importance of indirect paths would reach different conclusions in the three data collection methods.

In the case of the well-being model, three different investigators would reach three different conclusions about the importance of the different domains for the factor domain satisfaction: The most important domains are "social network" and "income" in the mail survey, "social network" in the face-to-face interview, and "income" in the telephone interview. "Housing" in the telephone interview is not significant and may be dropped from the model. As mentioned above, the statistical fit for even the least restrictive (same pattern) model was not quite satisfactory. To achieve an acceptable fit, our three hypothetical investigators would probably perform an exploratory specification search. Exploratory analyses in which restrictions between groups were freed (based on the modification indexes) until a fitting model was found resulted, in fact, in different models, with a different pattern of factor loadings for each of the three data collection methods (De Leeuw & Hox 1993).

In sum: There was a clear influence of data collection method on estimated relationships between variables. In our case, using a different data collection method would not lead to drastically different models. For loneliness, the same model fits in each method, but the relative importance of some variables varied considerably across methods. For well-being, the relative importance of different domains varies, and investigators would probably resort to some exploratory model specification, leading to models that actually differ across methods.

The effects of data collection mode are especially important when so-called mixed-mode designs are used, that is, survey designs that combine telephone, mail, and/or face-to-face procedures (cf. Dillman and Tarnai 1988). For instance, in a panel survey, a face-to-face interview is often used in the first recruitment wave, whereas in the following waves the less expensive telephone or mail surveys are used. In mixed-mode cross-sectional surveys, the usual strategy is to employ one method (often the relatively inexpensive mail survey) in the first round(s) of data collection, followed by a switch to another method to address the reluctant respondents (e.g., having experienced interviewers call refusers by telephone). Such designs confound data collection method variance with systematic wave or respondent variance. To counteract this, we recommend incorporating the method of data collection as a control variable in the data analyses.

In the social science literature, it is frequently stated that empirical research studies must be replicated before their results are accepted as facts (Fruchtgott 1984; Schuman and Presser 1981, Appendix A) or even before they can be published (Lubin 1957). Nevertheless, it is not obvious which type of study should be considered as a replication. Schweizer (1989) argued that the function of a replication study is to prove a scientific fact and that, therefore, a replication study must be as similar as possible to the original study. Our study showed that data collection method can have a substantial influence on research conclusions. Consequently, a new study using another data collection method is not a replication in this strict sense.

The concept of external validity applies to the generalization of research conclusions. Cook and Campbell (1979, chap. 2) considered the generalization of conclusions across persons, settings, and times as an aspect of external validity. The generalization across data

collection methods must also be considered as an aspect of external validity. This implies that for demonstration of external validity of research findings data collection method is a factor that must be used systematically in the design of series of research studies. The results of different studies with different data collection methods can be combined by using statistical meta-analysis techniques; see, for example, Hedges and Olkin (1985).

NOTES

1. For a description in English of the equivalent versions of the questionnaire, including question format and question wording, see De Leeuw (1992, Appendix B). The complete Dutch text of the questionnaires is available from the first author.

2. In all subsequent analyses potential effects of the differences between the modes in gender and marital status were checked by weighing. Because the results from the weighted analyses were virtually identical to the unweighted results, we present only the unweighted results here.

3. This approach was inspired by an analogous suggestion by one of the reviewers.

4. To prevent confounding of the covariances with mean differences between groups, the variables were first centered around the group means. Results from analyses in which the algorithm had not converged after the default number of iterations were omitted; this happened most frequently in the equal covariances model for the happiness data (27 times). Inclusion of nonconverged results does not change the results in Table 1.

5. The values outside the permitted range may also be caused by the standardization to a common metric instead of standardization to a group-specific metric.

6. Differences in variances are taken into account in the comparison of the correlation matrices. To assess the similarity of the marginal distributions, we computed kurtosis and skewness values for all variables for each data collection method. There is some evidence for nonnormality of the variables "extension of social network" and "subjective evaluation of social network," but the degree of nonnormality is similar in all three data collection methods. The correlations between the kurtosis values of the variables across the three methods were all larger than .90, and the correlations between the skewness values were all larger than .97. In summary: The pattern of kurtosis and skewness values of the variables was highly similar across methods.

7. This is unlikely to be the result of differences in statistical power, because the final sample sizes for the three methods are almost equal: 254 respondents in the mail survey, 243 in the face-to-face interview, and 266 in the telephone interview.

REFERENCES

Baumgartner, Robert M. and Thomas A. Heberlein. 1984. "Recent Research on Mailed Questionnaire Response Rate." Pp. 65-76 in *Making Effective Use of Mailed Questionnaires*, edited by D. C. Lockhardt. San Francisco: Jossey-Bass.

Bentler, Peter M. and Douglas G. Bonett. 1980. "Significance Tests and Goodness-of-Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88:588-600.

Bollen, Kenneth E. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Bradburn, Norman M. 1969. *The Structure of Well-being*. Chicago: Aldine.

Browne, Michael W. 1984. "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures." *The British Journal of Mathematical and Statistical Psychology* 71:62-83.

Burt, Ronald S., Michael G. Fischer, and Kenneth P. Christman. 1979. "Structures of Well-Being: Sufficient Conditions for Identification as Restricted Covariance Models." *Sociological Methods & Research* 8:111-20.

Burt, Ronald S., James A. Wiley, Michael J. Minor, and James R. Murray. 1978. "Structure of Well-being: Form, Content, and Stability Over Time." *Sociological Methods & Research* 6:365-407.

Byrne, Barbara M. 1989. *A Primer of Lisrel*. New York: Springer.

Converse, Jean M. and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: Sage.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues*. Chicago: Rand McNally.

De Jong-Gierveld, Jenny 1987. "Developing and Testing a Model of Loneliness." *Journal of Personality and Social Psychology* 53:119-28.

De Jong-Gierveld, Jenny and Frans Kamphuis. 1985. "The Development of a Rasch-Type Loneliness Scale." *Applied Psychological Measurement* 9:289-99.

De Leeuw, Edith D. 1991. *The Influence of Data Collection Procedure on Psychometric Reliability and Scaling Properties*. Response Effects in Surveys, Technical Report no. 5. Amsterdam: Vrije Universiteit, Department of Social Research Methodology.

———. 1992. *Data Quality in Mail, Telephone, and Face-to-Face Surveys*. Amsterdam: TT-Publikaties.

De Leeuw, Edith D. and Joop J. Hox. 1988. "The Effects of Response-Stimulating Factors on Response Rates and Data Quality in Mail Surveys: A Test of Dillman's Total Design Method." *Journal of Official Statistics* 4:241-50.

———. 1993. "Mode Effects in Structural Modeling; A Lisrel Multi-Group Comparison of Mail, Telephone, and Face-to-Face Survey Data." Pp. 119-44 in *Methodische Grundlagen und Anwendungen von Strukturgleichungsmodellen*, edited by J. Reinecke and G. Krekeler. Mannheim, Germany: FRG e.V.

Dillman, Don A. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley.

———. 1991. "The Design and Administration of Mail Surveys." *Annual Review of Sociology* 17:225-49.

Dillman, Don A. and John Tarnai. 1988. "Administrative Issues in Mixed Mode Surveys." Pp. 509-28 in *Telephone Survey Methodology*, edited by Robert M. Groves, Paul P. Biemer, Lars E. Lyberg, James T. Massey, William L. Nicholls II, and Joseph Waksberg. New York: Wiley.

Edgington, Eugene S. 1987. *Randomization Tests*. New York: Marcel Dekker.

Forsyth, Barbara H. and Judith T. Lessler. 1991. "Cognitive Laboratory Methods: A Taxonomy." Pp. 393-418 in *Measurement Errors in Surveys*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. New York: Wiley.

Fruchtgott, Ernest. 1984. "Replicate, Again and Again." *American Psychologist* 33:1315-16.

Good, Philip. 1994. *Permutation Tests*. New York: Springer.

Goyder, John. 1985. "Face-to-Face Interviews and Mailed Questionnaires: The Net Difference in Response Rate." *Public Opinion Quarterly* 49:234-52.

———. 1987. *The Silent Minority: Nonrespondents on Sample Surveys*. Cambridge: Polity Press.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

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

Hedges, Larry V. and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. London: Academic Press.

Hox, Joop J. 1986. *Het Gegruik van Hulptheoriën bij Operationaliseren* (Using auxiliary theories for operationalization: A study of the construct of subjective well-being). Unpublished doctoral dissertation, University of Amsterdam, Amsterdam, the Netherlands.

Hox, Joop J. and Edith D. De Leeuw. 1994. "A Comparison of Nonresponse in Mail, Telephone, and Face-to-Face Surveys; Applying Multilevel Modeling to Meta-Analysis." *Quality and Quantity* 28:329-44.

Hox, Joop J., Edith D. De Leeuw, and Ita G. G. Kreft. 1991. "The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model." Pp. 439-61 in *Measurement Errors in Surveys*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. New York: Wiley.

Jöreskog, Karl G. and Dag Sörbom. 1989. *Lisrel 7: A Guide to the Program and Applications*. 2nd ed. Chicago: SPSS.

Keenan, Brian and Marilyn Mauch. 1986. *Developing and Using Questionnaires*. Washington, DC: U.S. General Accounting Office, Program Evaluation and Methodology Division.

Kerssemakers, Frans A. M. 1985. "Telefonisch Enquêteren" (Telephone interviewing). Pp. 211-30 in *CBS-Select 3*. Voorburg/Heerlen, the Netherlands: Statistics Netherlands.

Krosnick, Jon A., and Duane F. Alwin. 1988. "The Form-Resistant Correlation Hypothesis." *Public Opinion Quarterly* 52:526-38.

Louwen, Frank. 1992. "Bereidwillig maar niet bereikbaar of bereikbaar maar niet bereidwillig" (Willing but unreachable, or reachable but unwilling). *Onderzoek* 10:5-9.

Lubin, Ardie. 1957. "Replicability as a Publication Criterion." *American Psychologist* 12:519-20.

Lyberg, Lars and Daniel Kasprzyk. 1991. "Data Collection Methods and Measurement Error: An Overview." Pp. 237-57 in *Measurement Errors in Surveys*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. New York: Wiley.

Noreen, Eric W. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York: Wiley.

Oldendick, Robert W., George F. Bishop, Susan B. Sorenson, and Alfred J. Tuchfarber. 1988. "A Comparison of the Kish and Birthday Methods for Respondent Selection in Telephone Surveys." *Journal of Official Statistics* 4:307-18.

Rosenberg, Morris. 1979. *Conceiving the Self*. New York: Basic Books.

Sackman, Harold. 1974. *Delphi Assessment: Expert Opinion, Forecasting and Group Process*. Santa Monica, CA: RAND.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Schweizer, Karl. 1989. "Eine Analyse der Konzepte, Bedingungen und Zielsetzungen von Replikationen" (An analysis of the concepts, conditions, and aims of replications). *Archiv für Psychologie* 141:85-97.

Snijkers, Ger J.M.E. 1992. "Computergestuurd Enquêteren: Telefonisch of Persoonlijk?" (Computer-assisted interviewing: By telephone or in person?). *Kwantitatieve Methoden* 39:53-69.

Sudman, Seymour and Norman M. Bradburn. 1982. *Asking Questions: A Practical Guide to Questionnaire Design.* San Francisco: Jossey-Bass.

Trewin, Dennis and Geoff Lee. 1988. "International Comparison of Telephone Coverage." Pp. 9-24 in *Telephone Survey Methodology*, edited by R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, II., and J. Waksberg. New York: Wiley.

Edith D. de Leeuw is a senior researcher at the Vrije Universiteit in Amsterdam. Her main research interests are nonresponse in surveys and comparisons of data collection techniques. At present she is engaged in a methodological study on measurement error in the interrogation of elder adults.

Gideon J. Mellenbergh is a professor of methodology in the Department of Methodology at the University of Amsterdam. He teaches courses on psychometrics and the methodology of empirical psychology. His main research interest is in psychometric theory and its applications in the construction and use of psychological and educational tests.

Joop J. Hox is an associate professor of methods and statistics in the Department of Education at the University of Amsterdam. His main research interests are multilevel modeling, survey research methodology, and problems of operationalization and measurement.