Nonresponse versus Measurement Error: Are Reluctant Respondents Worth Pursuing?

Joop J. Hox, Edith D. de Leeuw
Department of Methodology and Statistics, Utrecht University, the Netherlands
Hsuan-Tzu Chang
Department of Psychology, National Taiwan University, Taiwan

In: Bulletin de Méthodologie Sociologique (2012), *113*, 5–19.
DOI: 10.1177/0759106311426987

Corresponding author: Joop Hox, Department of Methodology and Statistics, Utrecht University, the Netherlands. P.O.B. 80140, NL-3508 TC, Utrecht, the Netherlands. Email: j.hox@uu.nl

Abstract

To increase response rates survey researchers intensify their efforts to bring sampled persons into the respondent pool. The question is whether ‘reluctant’ survey respondents provide answers of lower quality than ‘eager’ respondents. WE define eager respondents as persons who respond to the first round of a mail survey, and reluctant respondents as persons who respond in later rounds. We used a multitrait-multimethod (MTMM) design, which allows statistical separation of substantive or trait variance, method variance, and error variance. The results show that the measurement structure does not differ between eager and reluctant respondents. There was also no systematic difference in the reliability and validity estimates for both groups.

Keywords

Data Quality, Measurement Error, Multitrait-MultiMethod, MTMM, Nonresponse Bias, Reluctant Respondents, Total Survey Error

INTRODUCTION

Participation in surveys has been declining over time (De Leeuw and De Heer, 2002), and this trend is visible in all sectors of the survey industry (Brehm, 1994; Goyder, 1987) and for all survey modes (Hox and De Leeuw, 1994). In survey methodology a high response rate is commonly viewed as indicating a 'good' survey, and in the last three decades survey researchers have devoted much time and effort to counteract the downward trend in response rates (e.g., Dillman, 1978; De Leeuw, 1999; Goyder, 1987; Groves and Couper, 1998; Groves, Dillman, Eltinghe, and Little, 2002; Morton-Williams, 1993; Singer, 2006; Stoop, 2005). However, in recent years there is a growing concern that achieving a high response rate may not always lead to higher quality data, and although nonresponse should not be taken lightly, survey researchers should regard the totality of survey quality indicators (Biemer, 2010, Biemer and Lyberg, 2003, 95, Groves and Lyberg, 2010, see also Groves, 1989, 133, 147).

The quality of survey data can be threatened by sample composition bias, due to nonresponse and self-selection of respondents and by response bias from several sources. Increasing the response rate diminishes the potential impact of selection bias. For example, research has shown that reminders and increased fieldwork effort, not only bring in more respondents, but also can bring in those respondents that are underrepresented, such as the elderly, lower educated, and lower income groups (e.g., Dillman, 1978; Stoop, 2005). However, this could be purely cosmetic. As nonresponse error is a function of the nonresponse rate and the difference between respondents and nonrespondents on a particular variable of interest (for an overview, see Couper and De Leeuw, 2003), nonresponse error will only be reduced by drawing in those specific respondents that narrow this gap. That this is not always the case is shown by Groves and Peytcheva (2008), who in a meta-analysis of 59 methodological studies, found only a weak relationship between the response rate achieved and the nonresponse bias.

But, even if increasing the response rate does reduce nonresponse errors, by convincing special subpopulations to respond, the question remains whether it decreases the total survey error. Increasing the response rate by increasing the level of effort to bring a sampled person in the respondent pool and drawing in reluctant respondents may actually increase the amount of measurement error (Groves and Couper, 1998); in other words, it is feared that reluctance to respond may be related to data quality. Two theoretical models can be discerned to describe this situation: an *independence* model and a *common cause* model (e.g., Olson, Feng, and Witt, 2008; Tourangeau, Groves, Presser, Toppe, Kennedy, and Yan, 2008). In the independence model nonresponse error and measurement error are uncorrelated and have different sources. Nonresponse is caused by situational (e.g., time, opportunity, at home patterns) and motivational (e.g., altruism, low cost compared to benefits, high saliency) factors. Measurement error, on the other hand, is largely cognitive and related to the question-answer process (e.g., poor comprehension of question, memory and retrieval difficulties).

In the common cause model, respondents that are difficult to persuade also answer less thoroughly, and the underlying cause of survey nonparticipation and measurement error is the same. The common cause model states that there is indeed a relationship between reluctance to respond and data quality. Two different mechanisms can be posed for this common cause relationship between response propensity and measurement error. First, there is evidence that reluctant respondents tend to be older, have a lower education and a lower social economic status (SES), see for instance Couper and Groves (1993), Dillman (1978), Goyder (1987), Groves (1989 and Stoop (2005). To obtain a representative sample of the population, extra survey effort is often exercised to get these reluctant respondents in. As a

consequence, there will be small but replicable socio-demographic differences between eager and reluctant respondents, which in turn can give rise to differences in the amount of measurement error. Respondents with lower education or language problems are expected to produce more measurement errors because they are less capable to go optimally through all phases of the survey question-answer process. Such correlates of measurement error are denoted *extrinsic* error sources, because they derive from a different composition of the response groups, and are not related to the survey process itself. In other words, a relationship between reluctance to respond and data quality is indeed present, but this relationship is spurious and can be explained by differences on background variables (e.g., age, education) between eager and reluctant respondents.

The second mechanism depends on *intrinsic* error sources, which are related to the survey itself. Intrinsic error sources include respondent motivation, interest in the study, or degree of suspicion if sensitive questions are involved. For instance, highly motivated sample persons or people highly interested in the study will be more prone to respond and will be also more prone to invest effort to go carefully through the question-answer process. Lesser motivated sample members will more easily refuse to cooperate with the survey request or when persuaded will fall back on easy, suboptimal response strategies, such as satisficing, rather than using an optimal response strategy (Krosnick, 1991). Therefore, the lesser motivated will not only refuse more often, but also produce more measurement errors when persuaded to respond. If, indeed intrinsic errors play a major role in the common cause model, there will be a relationship between reluctance to respond and data quality, which cannot be explained away by socio-demographic differences between reluctant and eager respondents.

The purpose of this study is to investigate whether reluctant respondents produce larger measurement errors, and whether this is the result of self-selection or of intrinsic differences. Using confirmatory factor analysis of multitrait-multimethod data from a mail survey, we explore the relationship between reluctance and data quality. Furthermore, we investigate whether differences in sample composition on background variables between eager and reluctant respondents (i.e., extrinsic factors) may explain differences in data quality.

REVIEW OF EARLIER STUDIES

There is some evidence that supports the hypothesis that reluctant sample persons, which are brought into the respondent pool through increasingly persuasive efforts, may provide data with more measurement error than sample persons who respond immediately. As early as 1963, Cannel and Fowler (1963) found that respondents who reacted immediately to a mail survey provided more accurate responses regarding hospital episodes, than more reluctant respondents who reacted to the second mailing, while respondents who had to be prompted a third time by telephone or personal visit gave the least accurate answers. However, Olsen (2006) using record checks for divorce and marriage data finds only few and small associations between response propensity and measurement error, while Olsen and Kennedy (2006), who compare survey results on donations and academic performance with administrative records on university alumni, find no support for the hypothesis that respondents who are more difficult to recruit gave less useful data than more amenable respondents. Muller, Krauter and Trappman (2009) did find a relationship between contactibility and conversion of soft refusals with measurement error on employment data, but the relationships between contactibility and underreporting became nonsignificant when variables related to task difficulty were added to the regression model. Finally Olsen, Feng,

and Witt (2008) summarize seven studies that look into differences in response accuracy between high and low recruitment effort respondents on such diverse topics as medical history, voting, delinquency and academic performance. They report that findings differ dramatically by type of effort and that when follow-up call attempts are made, small but significant effects are found between immediate respondents and reluctant respondents who needed more follow-ups, with less accuracy for the latter. However, hardly any effect was found for refusal conversion.

All studies cited above investigated behavioural data and could use hard validating information (e.g., records) to investigate response accuracy. However, the situation that hard criteria for data quality by means of validating data are available is rare. Furthermore, when subjective phenomena, such as attitudes, are investigated, hard validating data do not exist. In the absence of validation data, a variety of proxy indicators for measurement error are used; a common used proxy is item nonresponse. In their meta-analysis, Olsen et al (2008) summarize the results of 15 studies with a total of 178 questions for which question-level item nonresponse rates were available. They find that respondents recruited with more effort have higher item nonresponse rates than those recruited easily; they also find that this effect is larger for refusal conversion studies than for other studies.

When attitudinal data are investigated, no consistent evidence is found that more effort to get sample persons into the respondent pool leads to worse data. For instance, Yan, Tourangeau and Arens (2004) find hardly any relationships between nonresponse propensity and various indicators of response bias, such as acquiescence, extremeness, and non-differentiation; however, they do find some effect for no-opinion responses with later respondents producing more no-opinion answers. When multiple-item attitude scales are used, it is possible to calculate the psychometric reliability of measurements. Green (1991) found small, nonsignificant, and inconsistent effects of follow-ups on scale reliability, but did find that late respondents score lower on several attitude scales. De Leeuw and Hox (1988) find similar results: small differences between eager and reluctant respondents in four scales indicating attitude towards surveys: faithfulness, apprehension, suspiciousness, and perceived question threat, with reluctant respondents indicating less positive attitudes on all scales, but they find no differences in psychometric quality. Finally, Chen, Wei, and Syme (2003) show that although poor response is associated with biographical background variables, there is no clear association between delayed response and psychographical variables, such as personality traits. A similar result was reported by Hox, De Leeuw, and Vorst (1996).

Petchev and Petcheva (2007) are among the first who go beyond proxy indicators and apply a more complicated model-based definition of measurement error when no validating data are available. Based on mean-variance models they show that although older and less educated respondents do provide more measurement error, but there is no association between response propensity and measurement error. Finally, Kamiska, McCutcheon, and Billiet (2010) use latent class analysis and structural equation modelling to explore satisficing among reluctant respondents in a cross-national context. Their findings suggest a relationship between reluctance and response quality, but this relationship could be explained away by differences in cognitive ability. For the present study we reanalyzed a dataset which included the rare combination of information on the nonresponse process and a multitrait-multimethod matrix to investigate response quality. This enabled us to use a very strong model (cf. Saris and Gallhofer, 2007) to explore the relationship between reluctance and data quality.

MEASUREMENT ERROR

Measurement error is operationalised in different ways in different studies. Ideally, the true

value is known and this true value is then compared with the reported value. Some studies do have access to a validation criterion and therefore can carry out a record check. But, when subjective phenomena are studied, hard validation data are per definition not available, and researchers have to rely on various proxy indicators of data quality. Biemer (2001) points out that these often rely on assumptions on the direction of the biases (e.g., underreporting of sensitive information) and argues in favour of a model-based approach instead. Petchev and Petcheva (2007) also plead for a model-based approach.

A direct model-based approach to the analysis of measurement error in surveys on subjective phenomena is the multitrait-multimethod (MTMM) design that allows separation of substantive or trait variance, method variance, and error variance (Campbell and Fiske, 1959; Alwin, 1974; Saris and Andrews, 1991).

The most common approach to evaluating the measurement model in MTMM designs is confirmatory factor analysis, which defines both the substantive traits and the measurement methods as latent factors. See Hox (1995) for an application using different software packages. To investigate if eager and reluctant respondents produce different measurement errors, we have to compare the measurement model for both groups.

Three questions can be addressed when groups are compared using confirmatory factor analysis. First, the question is whether the eager and reluctant respondents share the same factor structure. This is the weakest form of measurement invariance, in comparative research this is often denoted as factorial or *functional equivalence* (Vandenberg and Lance, 2000). If functional equivalence holds, the *same constructs* are measured in both groups. The second question is whether the constructs are measured equivalently in both groups. If the factor loadings for eager and reluctant respondents are identical, we have a form of equivalence that is referred to as *metric equivalence* (Vandenberg and Lance, 2000): the same constructs are measured in the *same way* for both groups. The third question is if the intercepts for the observed variables are identical for both groups. If these intercepts can be considered invariant across groups, this is called *scalar* equivalence (Vandenberg and Lance, 2000) and the actual scores can be compared across groups. When scalar equivalence holds, the same constructs are measured in the same way and on the *same scale* for both eager and reluctant respondents. Finally, a fourth question is whether both groups contribute equal amounts of measurement error, as indicated by the error variances of the responses.

If differences are found between eager and reluctant respondents, this may be because the reluctant respondents, who did not respond initially, are different on background characteristics due to initial selective nonresponse (extrinsic factors). It may also be because reluctant respondents produce more measurement error for intrinsic reasons, for instance, because they are less motivated and tend to satisfice more. Thus, the final step is to analyze which part of the difference is due to differences between the two groups in socio-demographic characteristics because of selective nonresponse in the reluctant group, and which part is due to cognitive aspects in the question-answer process. This question can be addressed by comparing measurement models with and without propensity score adjustment for differences in socio-demographic background variables (cf. Rubin and Thomas, 1996).

In this study we use this model-based MTMM approach to investigate whether reluctant respondents who are pressured to comply with the survey request produce data of lower quality than eager respondents who immediately respond to the survey request. Furthermore, we investigate if differences are caused by extrinsic causes, that is, by mere differences in background variables, or by intrinsic causes, which are related to the survey and question-answer process itself.

METHOD

Sample and Survey Procedure

A secondary analysis was performed on data collected for a survey on well-being in The Netherlands (Hox, 1986). These data have the advantage that both a careful record of the (non) response was kept and that an MTMM approach was used, giving us the rare opportunity to use a model-based approach to investigate the relationship between nonresponse and measurement error.

The data were collected with a mail survey using Dillman's (1978) TDM approach (see also Dillman et al, 2008), including two reminders with a replacement questionnaire. The questionnaire was mailed to a sample of 1000 addresses from the telephone directory of the Netherlands, which at the time of data collection (1984) constituted a good sampling frame for the general population. According to Dutch telecom approximately 90% of the private households at that time had a listed landline telephone (see also Trewin and Lee, 1988). The response rate is 53% (AAPOR, 2011, standard definitions RR3). Three returned questionnaires contained a large fraction of missing data and were discarded, leaving 498 cases for the analysis.

Respondents who responded to the initial mailing were classified as eager respondents; respondents who responded to the reminders were classified as reluctant respondents. There are 237 eager and 239 reluctant respondents. Compared to the eager respondents the reluctant respondents differed on important background characteristics; they were older, lower educated, more often rented their house and were more often without a paid job. Furthermore, there were small differences in marital status and gender with more females and more married persons in the reluctant group.

The questionnaire included a MultiTrait-MultiMethod (MTMM) design consisting of three different aspects of well-being (traits), each measured by five question formats (methods). The three traits measuring well-being are 'satisfaction with housing', 'satisfaction with income', and 'satisfaction with health'. As methods both verbal and graphical question formats were used. The verbal question formats are a 'direct question' and a 'social comparison question'; the graphical question formats are 'Cantril's ladder', 'faces (smileys)', and 'circles' (Andrews and Withey, 1978). An example of each question format is presented in the Appendix.

Analysis

The most common model for MTMM data is a confirmatory factor model with a factor for each trait and a factor for each method, with trait and method factors mutually uncorrelated (Alwin, 1974; Eid, Lischetzke and Nussbeck, 2006). In this model, the trait factors are allowed to correlate, and the method factors are usually uncorrelated. Saris and Gallhofer (2007) present a confirmatory factor model for MTMM data that is formally equivalent to the classic MTMM model, but that also allows separate estimation of the reliability and the validity of each question. This model has been used by Scherpenzeel and Saris (1993) to compare the quality of measurement in 10 different European countries. Figure 1 depicts this model for three traits and three methods. The three trait factors and the three methods define 3×3=9 survey questions labelled v1 to v9. The latent variables t1 to t9 represent the true scores for v1 to v9, and the standardised loading of v1 to v9 on the true scores t1 to t9 represent the reliabilities of the survey questions. The loadings of t1 to t9 on the corresponding trait factors (labelled tr1 to tr3) are their validities. The loadings on the method factors (labelled m1 to m3) represent systematic method effects unrelated to the traits. The

latent scales for the factors are identified by constraining the variances of all trait and method factors equal to one.

--- Figure 1 about here ---

Our MTMM design includes three traits (satisfaction with housing, income, and health) and five methods (direct question, social comparison, ladder, faces, and circles), which leads to 15 different questions. We estimated the MTMM model described above for the eager and reluctant respondents simultaneously using a two-group structural equation model. A series of equality constraints across the two groups is used to test for measurement equivalence. All analyses were carried out in the program Amos (Arbucle, 2007).

To assess how much of the difference is due to socio-demographic differences (extrinsic errors), a propensity score method is used (Rubin and Thomas, 1996). In the propensity score method, the propensity to be in the eager group was estimated using a logistic regression model with the background variables age, gender, marital status, education, having a job, and house ownership, variables on which the eager and reluctant respondents differed. Next, this propensity score was included in the MTMM model as a covariate, thereby statistically controlling for differences between the eager and reluctant respondents in age, gender, marital status, education, having a job, and house ownership.

RESULTS

The MTMM model was first estimated on the entire sample. The fit of this model is good (χ^2=123.1, *df*=72, *TLI*=0.99, *CFI*=0.99, *RMSEA*=0.04). Subsequently, a series of nested multi-group models were fitted comparing eager and reluctant respondents. The first model imposes functional equivalence (identical factor structure) and is the model with the least restrictions which assesses the weakest form of measurement invariance. The second model adds metric equivalence (identical factor loadings); this model assesses if the constructs are measured equivalently in both groups. The third model adds scalar equivalence (identical intercepts); when this holds it means both groups can be compared on their factor means. Finally, we test if the error variances are equal across the two groups. This tests whether there are differences in measurement error between the eager and the reluctant respondents. Table 1 shows the fit indices (chi-square, *df*, *p*, *TLI*, *CFI* and *RMSEA*) for these models. In addition, Table 1 shows the results of a chi-square difference test, testing the model under consideration against the previous model in the table.

--- Table 1 about here ---

As Table 1 shows, the model that poses functional equivalence (no constraints) fits well. The models that pose metric and scalar equivalence also fit well, and do not differ significantly from the previous model or the functional equivalence model. This means that the strongest form of equivalence, scalar equivalence, holds across the eager and reluctant respondents.

Table 1 also shows the results for a model that poses equal *error variances*. This model has a significantly worse fit than the scalar equivalence model. Thus, the same constructs are measured in both groups of respondents with the same factor structure and identical factor loadings and intercepts, but with different amounts of error variance. This means that eager and reluctant respondents indeed differ in the amount of random measurement error. This is also expressed in the estimates of the reliability of the nine

questions reported in Table 2.

--- Table 2 about here ---

Table 2 shows the parameter estimates for the model with equal loadings and intercepts. Using the model proposed by Saris and Gallhofer (2007), we obtain for each question in the MTMM an estimate of its validity (the trait loading), its reliability (the proportion systematic variance) and the error variance. The error variances are allowed to differ across the two groups. As Table 2 shows, the error variances and the reliabilities do not consistently differentiate between the two groups. So, although the groups differ in the error variances, there is no systematic tendency for the reluctant respondents to respond with more error and hence lower reliability. The validities also do not differ across the two groups, since the trait loadings can be constrained equal across the eager and reluctant respondents.

Table 2 shows one clear difference: in general the graphical question formats (faces, ladder, circles) perform very well compared to the verbal question formats (direct question and social comparison). This is clearly shown in Table 3, which reports the average reliability estimates for the verbal and graphical questions, for both eager and reluctant respondents.

--- Table 3 about here ---

The same sequence of models was fitted including propensity score adjustment to account for differences between eager and reluctant respondents on age, gender, marital status, education, having a job, and house ownership. The propensity score adjustment was carried out by regressing each of the observed variables on the propensity score, effectively estimating the MTMM model on the residuals of these regressions. After propensity score adjustment, the results were essentially the same as found without propensity score adjustment: scalar equivalence holds, and the error variances are not equal. Thus, the propensity score adjustment results in very small differences, indicating that differences in background characteristics do not explain away the difference in error variances. Reliabilities differed between the eager and reluctant respondents, but not systematically. Again, the graphical question formats showed a much higher reliability than the verbal question formats for both groups.

DISCUSSION

The main outcome from this study is that eager and reluctant respondents differ very little in the quality of their responses to our well-being questions. Overall, the graphical scales are performing very well; both validity and reliability tend to be high for these question formats. Andrews and Withey find similar results, in their analyses the graphical scales also performed better than the verbal scales, with especially the social comparison question performing very poorly (Andrews and Withey, 1976, p204), a finding that is also replicated in our analysis.

A positive outcome is that all differences between eager and reluctant respondents reside in the error structure of the MTMM data, and that these differences are not systematic. On average, the reliability of the responses of the eager and the reluctant group do not differ. We did *not* find differences in the measurement model proper, which represents the construct validity. In other words, the same constructs are measured in the same way for both the eager and the reluctant respondents. We may therefore conclude that a second round of data collection brings in more and demographically slightly different respondents, without affecting the quality of the data.

The generally small effect of propensity score adjustment on the results indicates that the existing differences between eager and reluctant respondents in socio-demographic characteristics do not affect the quality of the answers.

We should note that the results are based on a paper-and-pen mail survey. In such a survey, the respondents can view all questions, and page back and forward in the questionnaire at will. Hence, some amount of correlated error based on memory effects can be expected. These correlated errors are not included in the MTMM models, because including correlated errors led to estimation problems (nonconvergence). Given that the results are very similar between the eager and reluctant respondents, we assume that these effects do not bias our results. When using modern computer assisted interviewing methods or Internet surveys, it is preferable to present the MTMM questions randomly, or to pose different subsets of questions to different respondents, thereby reducing the redundancy in the question set. We refer to Scherpenzeel and Saris (1997) for a discussion.

REFERENCES

American Association for Public Opinion Research (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* http://www.aapor.org/For_Researchers/4228.htm (accessed May 2, 2011).

Alwin, DF (1974). Approaches to the interpretations of relationships and the multitrait-multimethod matrix. In Costner HL (ed) *Sociological Methodology 1973–74*. San Francisco: Jossey-Bass, 79–105.

Andrews FM and Withey SB (1978). *Social Indicators of Well-being.* New York: Plenum.

Arbucle J (2007). *Amos 16 User's Guide*. Chicago, IL: SPSS, Inc.

Biemer PP (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly, 74* (5): 817-848.

Biemer PP (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics, 17* (2): 295–320.

Biemer PP and Lyberg LE (2003). *Introduction to Survey Quality.* New York: John Wiley.

Brehm J (1994). *The Phantom Respondent: Opinion Surveys and Political Representation*. Ann Arbor, University of Michigan Press.

Campbell DT and Fiske DW (1959). Convergent and discriminant validation by the multimethod-multitrait matrix. *Psychological Bulletin,* 56: 833–853.

Cannell CF, and Fowler FJ (1963). Comparison of a self-enumerative procedure and a personal interview: a validity study. *Public Opinion Quarterly*, *27*: 250–264.

Chen R, Wei L and Syme PD. (2003). Comparison of early and delayed respondents to a postal health survey: a questionnaire study of personality traits and neuropsychological symptoms. *European Journal of Epidemiology, 18*: 195–202.

Couper MP and De Leeuw ED (2003). Nonresponse in cross-cultural and cross-national surveys. In Harkness JA, Van de Vijver FJR, and Mohler PP (eds). *Cross-Cultural Survey Methods*. New York: Wiley.

De Leeuw ED (1999). Preface: special issue on survey nonresponse. *Journal of Official Statistics*, 15: 2: 127–128.

De Leeuw ED and De Heer W (2002). Trends in household survey nonresponse: a longitudinal and international comparison." Groves RM, Dillman DA, Eltinge JL and Little RJA (eds) *Survey Nonresponse*. New York: John Wiley, 41–54.

De Leeuw ED and Hox JJ (1987). Artifacts in mail surveys. the influence of Dillman's total design method on the quality of the responses. In Saris WE and Gallhofer IN (eds) *Sociometric Research. Vol. II: Data Analysis,* London: MacMillan, 61–73.

Dillman DA (1978). *Mail and Telephone Surveys; The Total Design Method*. New York: Wiley.
Dillman DA, Smyth JD and Christian LM (2008). I*nternet, Mail, and Mixed Mode surveys; The Tailored Design Method.* New York: Wiley.
Eid M, Lischetzke T and Nussbeck FW (2006). Structural equation models for multitrait-multimethod data. In Eid M and Diener E (eds.), *Handbook of Multimethod Measurement in Psychology*. Washington, DC: American Psychological Association, 283–299.
Goyder J (1987). *The Silent Minority: Nonrespondents on Sample Surveys.* Boulder: Westview Press.
Green KE (1991). Reluctant respondents: differences between early, late, and nonresponders to a mail survey. *Journal of Experimental Education, 59* (3): 268–276.
Groves RM (1989). *Survey errors and Survey Costs*, New York: John Wiley.
Groves, RM, Dillman, DA, Eltinge, JL, and Little, RJA (1992). *Survey Nonresponse*. New York: John Wiley.
Groves RM and Couper, MP (1998). *Nonresponse in Household Interview surveys.* New York: John Wiley.
Groves RM and Peytcheva E (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, *72*: 167–189.
Groves RM and Lyberg L (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74* (5): 849-879.
Hox JJ (1986). *Het Gebruik van Hulptheorieën bij de Operationalisering. Een studie Rondom het Begrip 'Subjectief Welbevinden'.* [Using Auxiliary Theories for Operationalization] Amsterdam, the Netherlands: University of Amsterdam (Unpublished Ph.D. Thesis).
Hox JJ (1995). Covariance structure modeling in windows: a multitrait-multimethod analysis using Amos, Eqs, and Lisrel. *Bulletin de Méthodologie Sociologique*, *46*, 71-87.
Hox JJ and De Leeuw ED (1994). A comparison of nonresponse in mail, telephone, and face to face surveys: applying multilevel modeling to meta-analysis. *Quality & Quantity, 28*: 329–344.
Hox JJ, De Leeuw ED and Vorst H (1996). A reasoned action explanation for survey nonresponse. In Laaksonen S (ed) *International Perspectives on Nonresponse*. Helsinki: Statistics Finland, 101–110.
Kaminska O, McCutcheon AL, Billiet J (2010) Satisficing among reluctant respondents in across-national context. *Public Opinion Quarterly, 74,* 5, 956-984.
Krosnick JA (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*: 213–236.
Muller G, Kreuter F and Trappmann M (2009). *Nonresponse and Measurement Error in Employment Research.* Paper presented at the International Total Survey Error Workshop, Tallberg, Sweden 2009. (Retrieved May 2011 http://niss.org/node/703 also at http://niss.org/content/nonresponse-and-measurement-error-employment-research)
Morton-Williams J (1993). *Interviewer Approaches.* Aldershot: Darthmouth.
Olsen K, Feng C and Witt L (2008). *When do nonresponse follow-ups improve or reduce data quality?: A meta-analysis and review of the existing literature.* Paper presented at the International Total Survey Error Workshop, Research Triangle Park, NC June 2008. (Retrieved May 2011 from http://niss.org/event/itsew-2008-multiple-sources-error-and-their-interaction)
Olsen K (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, *70*: 737–758.
Olsen K and Kennedy C (2006). Examination of the relationship between nonresponse and

measurement error in a validation study of alumni. In *Proceedings of the Annual Meetings of the American Statistical Association*, Toronto 2006, 4181– 4188. (Retrieved ay 2011from http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000229.pdf).

Petchev A and Petcheva E (2007). *Relationship between measurement error and unit nonresponse in household surveys: an approach in the absence of validation data.* Paper presented at the International Workshop on Household Survey Nonresponse, Southampton, September 2007.

Rubin DB and Thomas N (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics, 52*: 254–268.

Saris WE and Andrews FM (1991). Evaluation of measurement instruments using a structural modelling approach. In Biemer PP, Groves RM, Lyberg LE, Mathiowetz N and Sudman S (eds) *Measurement Errors in Surveys*. New York: John Wiley, 575–599.

Saris WE and Gallhofer IN (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: Wiley.

Scherpenzeel AC and Saris WE (1993). The evaluation of measurement instruments by meta-analysis of MTMM studies. *Bulletin de Méthodologie Sociologique*, *39*, 3-19.

Scherpenzeel AC and Saris WE (1997). The validity and reliability of survey questions: a meta-analysis of MTMM studies. *Sociological Methods and Research*, *25*: 341–383.

Singer E (2006). Special issue: nonresponse bias in household surveys. *Public Opinion Quarterly, 70*: 5.

Stoop IAL (2005). *Nonresponse in Sample Surveys: The Hunt for the Last Respondent.* The Hague: Social and Cultural Planning office.

Trewin D and Lee G (1988). International comparison of telephone coverage. In Groves RM, Biemer PP, Lyberg LE, Massey JT, Nicholls WL and Waksberg J (eds) *Telephone Survey Methodology*. New York: Wiley, 9–24.

Tourangeau R, Groves RM, Presser S, Toppe C, Kennedy C and Yan T (2008). *Experiments Exploring the Relationship (or Lack Thereof) Between Nonresponse Error and Measurement Error*. Paper presented at the International Total Survey Error Workshop, Research Triangle Park, NC June 2008. (Retrieved May 2011 from http://niss.org/event/itsew-2008-multiple-sources-error-and-their-interaction)

Vandenberg RJ and Lance CE (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 2: 4–69.

Yan T, Tourangeau R and Arens Z (2004). When less is more: are reluctant respondents poor reporters? In *Proceedings of the Annual Meetings of the American Statistical Association*, Toronto. 2004, 4632–4651. (Retrieved May 2011 from http://www.amstat.org/sections/srms/proceedings/y2004/Files/Jsm2004-000169.pdf).

APPENDIX

Five question formats were used: (1) a standard self–report question (direct question), (2) a social comparison question, (3) a graphical ladder scale, (4) a faces (smileys) scale, and (5) a circle scale. Three traits were measured: (1) satisfaction with house, (2) satisfaction with income, and (3) satisfaction with health. An example of each question format is given below for the domain 'satisfaction with house'.

1. DIRECT QUESTION

- How satisfied or dissatisfied are you with the house you live in?
    1. Very dissatisfied
    2. Dissatisfied
    3. Somewhat dissatisfied
    4. About equally dissatisfies as satisfied
    5. Somewhat satisfied
    6. Satisfied
    7. Very satisfied

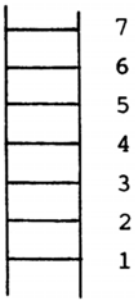
2. SOCIAL COMPARISON QUESTION

- When you compare yourself to the people around you, would you say that you are more satisfied with the house you live in, about equally satisfied, or less satisfied than most people?
    1. Much less satisfied
    2. Less satisfied
    3. A bit less satisfied
    4. About equally satisfied
    5. A bit more satisfied
    6. More satisfied
    7. Much more satisfied

3. GRAPHICAL LADDER QUESTION

Below is a drawing of a ladder. The top of the ladder represents the best that you could reasonably expect in life. The bottom represents the worst that you could expect in life.

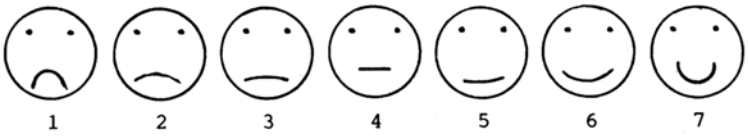
- If you were asked to use the ladder to illustrate how satisfied you are with the house you live in, where are you on the ladder?



4. GRAPHICAL FACES QUESTION

Below are a series of faces that express different feelings. Below each face is a number.

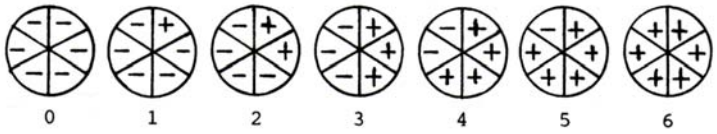
- Which face represents the best how satisfied you are with the house you live in?



5. GRAPHICAL CIRCLES QUESTION

Below are some circles that could represent the lives of different people. Circle 1 has only minuses; this represents persons who have only bad things in their lives. Circle 7 has only plusses; this represents persons who have only good things in their lives. The other circles are in between.

- Which circle represents the best how satisfied you are about the house you live in?



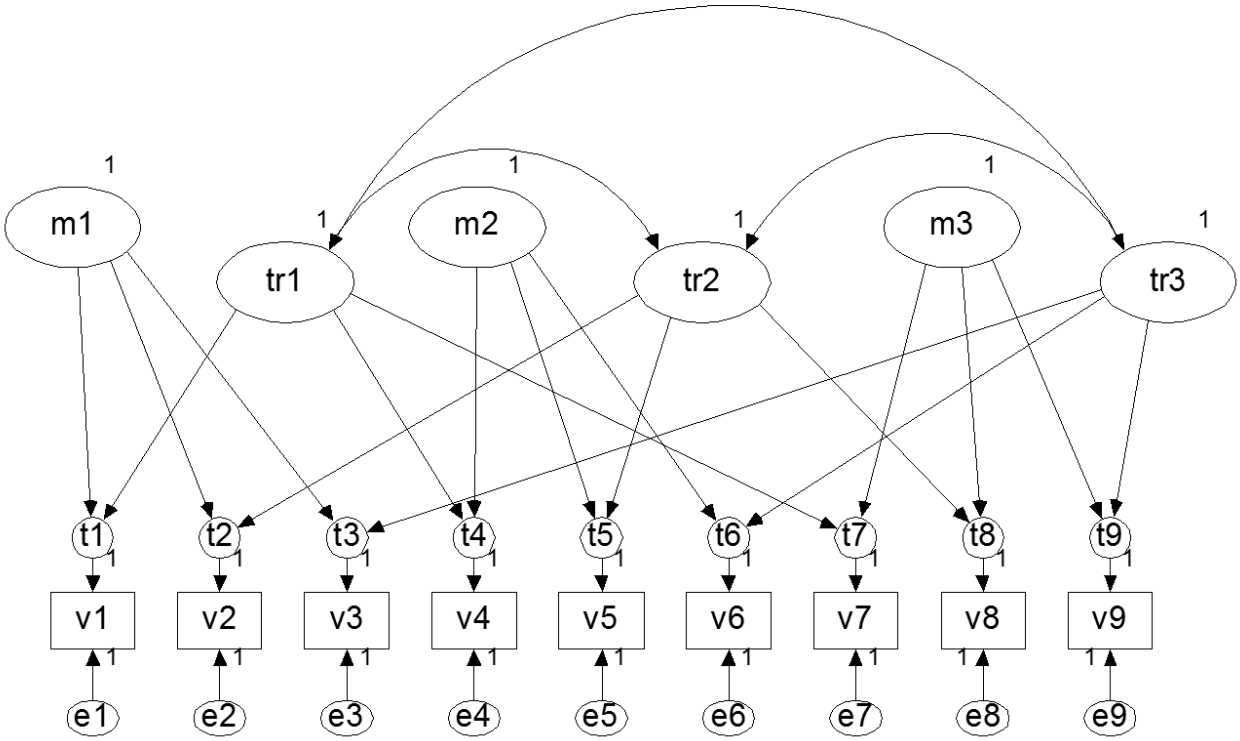


Figure 1. MTMM Confirmatory Factor Model According to Saris & Gallhofer

Table 1. Fit Indices for Models with Equivalence Constraints between Eager and Reluctant Respondents

Model	χ^2 (*d.f.*)	*p*	*TLI*	*CFI*	*RMSEA*	$\Delta\chi^2$ (Δ*d.f.*) [a]	*p*
No constraints (functional equivalence)	211.0 (144)	<.01	.98	.99	.03	-	-
Equal loadings (metric equivalence)	240.0 (166)	<.01	.99	1.00	.03	29.0 (22)	.14
Equal intercepts (scalar equivalence)	258.2 (181)	<.01	.99	.99	.03	18.2 (15)	.25
Equal error variances (equal reliability)	285.0 (196)	<.01	.98	.99	.03	26.8 (15)	.03

[a] $\Delta\chi^2$ reference is previous model

Table 2. Parameter Estimates from MultiTrait-MultiMethod Model.

Question	Trait (Validity)	Method	Reliability [a]	Error variance [a]
House-Direct	1.07	.28	.67 / .71	.60 / .44
House-Social	.69	.49	.55 / .50	.35 / .40
House-Ladder	1.17	.20	.90 / .87	.52 / .38
House-Faces	1.21	.39	.91 / .90	.58 / .65
House-Circles	1.06	.31	.88 / .90	.49 / .52
Income- Direct	1.08	.43	.79 / .80	.80 / .66
Income-Social	.85	.40	.64 / .66	.14 / .13
Income-Ladder	1.17	.16	.92 / .91	.16 / .09
Income-Faces	1.17	.46	.90 / .95	.19 / .11
Income-Circles	1.07	.41	.94 / .90	.12 / .20
Health-Direct	1.07	.19	.74 / .81	.15 / .17
Health-Social	.82	.39	.51 / .58	.16 / .12
Health-Ladder	1.27	.14	.92 / .94	.09 / .11
Health-Faces	1.30	.37	.89 / .94	.24 / .16
Health-Circles	1.07	.21	.84 / .85	.58 / .23

[a] Reliability and error variance separate for eager/reluctant respondents

Table 3. Average Reliability for Verbal and Graphical Questions Across Groups.

Group	Verbal	Graphical	Total
Eager	.65	.90	.80
Reluctant	.68	.91	.81
Total	.66	.90	