Measurement error in comparative surveys

Oberski, D.L.

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Measurement error in comparative surveys

Daniel L. Oberski

Measurement error in comparative surveys

Proefschrift ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 28 januari 2011 om 14:15 uur door Daniel Leonard Oberski, geboren op 25 augustus 1981 te Amsterdam.

Promotores:

Prof. Jacques A. P. Hagenaars
Prof. Willem E. Saris
Prof. Albert Satorra

Overige commissieleden:

Prof. J. W. Marcel Das
Prof. Gideon J. Mellenbergh
Dr. Guy B. D. Moors
Prof. Jeroen K. Vermunt

Contents

Preface

Writing the articles in this dissertation was literally a trip. I traveled from Amsterdam to Barcelona and from applied research to methodology and statistics, meeting many wonderful people, colleagues and friends.

The trip began when, still in Amsterdam, I met Willem Saris. Willem was the coordinator of the specialization in research methods at the University of Amsterdam, and I worked for him as an assistant to the 2005 European Survey Research Association conference and programmer on the SQP software. Just about to start a new group for his work on the European Social Survey, he invited me to work for him in Barcelona. Although I was unsure what I would do there, the concreteness and attractiveness of this proposal obliterated my vague plans to do "something" in the United States. With the decisiveness of someone who has no idea what he is getting into, I packed my bags and moved to Spain, where I had never been before in my life.

Before I got there, Laura Guillén had braved the gruelling Barcelona housing market for me and had found an apartment for me to live in. This incredible act of kindness from Laura, who I had only met once before, immediately made me feel welcome. By an unknown fortune, many subsequent acts of kindness from colleagues and other people would only strengthen that feeling.

I started working at the ESADE business school in Barcelona where I was received with open arms by Willem and his wife Irmtraud Gallhofer, Laura, Lluís Coromina, Desiree Knoppen, and the department's director Joan Manuel Batista. Lluís immediately became my Catalan teacher and my friend. I'm afraid I did not study much Catalan in those days, but I did learn how to drink from a *porró* (sort of). Joan Manuel needed just one look at me to proclaim that I was "still landing", and did everything he could to help me "taxi to the gate".

Willem became my patient mentor and tour guide. He showed me many important sights of the methods and statistics landscape and gave me just the right amount of freedom to explore on my own without getting lost. His innate empathy and uncanny ability to motivate inspired me to accomplish tasks that would have otherwise been insurmountable.

At the same time I started a PhD at the department of methods and statistics at Tilburg University and the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS). I preferred this to a PhD in business administration or political science; by now I was firmly hooked on research methods and statistics.

This was possible thanks to Jacques Hagenaars, who became my promotor together with Willem. On trips to Tilburg I came to know Jacques as a warm, friendly man with a resounding laugh that he used often and effectively to dispel any possible discomfort one might have had on meeting such an impressive figure. He also provided me with his

personal perspectives on research, and enthusiastically explained the many applications of latent class analysis. He was always critical and constructive, and could help me a great deal by making one small remark. Every one of our meetings produced a collection of scribbled-on slips of paper which I hoard to this day.

On the same trips I was extremely fortunate to meet Jeroen Vermunt. On one occasion especially, Jeroen spent quite a bit of his time instructing me and answering my cloddish questions. This private session aided me enormously in writing the second chapter of this book. Later he would also comment on the manuscript and was always supportive and helpful.

Together with Willem and Jacques I had been working on a paper about differences in quality between countries. This paper would later become the first chapter of this dissertation. Ever-capable IOPS secretary Susañña Verdel arranged for me to present this paper at my first IOPS conference, where I received useful feedback from the IOPS members. Eventually, thanks to the editing efforts of Tim P. Johnson, Michael Braun, and Janet Harkness, a revised version was also published as a book chapter by Wiley.

Around the same time I met and started working with another essential person for this thesis, who would become my third promotor: Albert Satorra. I had an idea for a paper, and wanted to talk to Albert about it. He had an even better idea: I could get a temporary contract teaching the practicum of his multivariate statistics course at the economics department of Pompeu Fabra University, and in the meantime we would work on the paper.

It turned out Albert's idea was the best possible one, as I learned much about structural equation modeling from patient explanations in his office over the course materials. Without his teachings and guidance on SEM I would have been unable to write the third chapter of this dissertation. We worked on the paper, which later became the last chapter of this dissertation. Albert treated me to many lunches in a certain Basque restaurant with an even larger number of stimulating conversations about all kinds of topics. He showed me that it is possible to combine an unforgivingly serious demeanor when it comes to science with kindness and generosity.

The trip continued as our research group grew, with the addition of my wonderful colleagues Wiebke Weber and Mélanie Révilla. We moved to a new institution, the Pompeu Fabra University in Barcelona. There we were welcomed by a new group of colleagues: Aina Gallego, Maria José Hierro, Gerardo Maldonado, Clara Riba, and the subdirector of our newly-founded center Mariano Torcal. Later on we would be joined by Paolo Moncagatta, Diana Zavala, André Pirralha, and Tom Gruner. I immensely enjoyed all of their company and support. Mélanie as well as Guy Moors from Tilburg helped me with their comments on earlier versions of chapter two. Collaborating with Mélanie, Tom, Aina, Wiebke, and Paolo is a joy – and a productive one!

Becoming more involved in the European Social Survey, I attended ESS meetings. There I had the privilege of collaborating with members of the Central Coordinating Team, in particular Jaak Billiet, Annelies Blom, Michael Braun, Brita Dorer, Gillian Eva, Rory Fitzgerald, Matthias Ganninger, Eric Harrison, Roger Jowell, Knut Kalgraff Skjåk, Joost Kappelhof, Achim Koch, Kirstine Kolsrud, Geert Loosveldt, Brina Malnar, Hideko Matsuo, Lorna Ryan, Angelika Scheuer, Ineke Stoop, and Sally Widdop. Even though these meetings did not contribute directly to my dissertation, they certainly provided a very stimulating experience in the world of survey research, which I am all too conscious of being extremely fortunate to have had.

On a course on SEM in Ljubljana, I was course assistant to Peter Schmidt, who turned

out also to be an expert on Ljubljana nightlife. In Berlin I met Frauke Kreuter. We talked about latent class analysis and lost spectacularly in a Wii karaoke competition because – according to the obviously faulty software – I did not get a single note right. When even such abominable singing cannot destroy a friendship, it is clearly something to hold onto, which I do gratefully.

As I neared completion of my dissertation I returned temporarily to the Netherlands to finish up. There I found that even though I had left my friends and family behind, they had not abandoned me. My parents, Arnan and Helga, Sacha and Micha, Lucas, Lea, Simon, and all my friends and family members; without their love and friendship I would be a different person. My friend Joost Heetman made the design for the cover of this dissertation.

At first after arriving in Amsterdam I worked on my own, until one day Heike Schröder put me in touch with Harry Ganzeboom who extended great hospitality by offering me a desk to work at in the VU University Amsterdam. This helped me prepare for several courses and talks I was giving and gave me a more scientific environment to work in.

Carla is the only secret I have kept from this account so far. Through the past years she gave me her love and support. And when we dance together I cannot help but go from flustered to excited about life.

Now I write this preface in Amsterdam it might appear as though the trip has come full circle. In reality I'm still on it, and look forward to the rest of it. But this is a good moment, from the bottom of my heart, to thank all the people who have in smaller or larger ways joined my trip.

Amsterdam, December 2010

Introduction

Comparative surveys nowadays provide a wealth of survey data on a diverse range of topics covering most countries in the world. The online companion[1] to the "SAGE handbook of public opinion research", for example, (Donsbach & Traugott, 2007) lists some 65 cross-national comparative social surveys that have been conducted around the world since 1948. Besides these general social surveys, there are also many surveys with specific topics such as education, old age and retirement, health, working conditions, and literacy, to name just a few.

Comparative surveys have several goals. One the one hand, they may serve to estimate and compare population means, totals, and marginal distributions, while on the other hand relationships between variables can be estimated. Van de Vijver and Leung (1997) called studies with these goals respectively "level" and "structure" oriented.

Comparative surveys are clearly popular, but not necessarily completely successful: errors due to various sources may interfere with the attainment of the two goals. Many categorizations exist for the sources of such survey errors (Groves, 1989; Weisberg, 2005). A relatively simple division can be drawn between *errors due to the selection of sample units*, and *errors due to the measurement instrument*. The error sources can have an effect in the form of both bias and variance, which together influence the root mean square error of the estimator. There are thus several different possible sources of error, which can have an effect in the form of bias and variance on the two different goals.

This book consists of four chapters that deal with a particular subset of these effects: the effect of measurement error on inference for and comparison of relationships. Before giving an outline of the chapters, however, this topic is placed in the more general framework of survey errors, discussing the various combinations of errors, goals, and effects. Figure 1 shows these combinations schematically in the form of arrows. The eight numbered arrows stand for effects of selection procedure and measurement instrument on two aspects (bias and variance) of the two goals (means and relationships) of surveys. Although it is impossible to review all of the literature that has been written about survey errors and their effects on estimators, the next few paragraphs are intended to give a short general overview, discussing each of the arrows in the figure in turn.

Arrows 1–4 in figure 1 stand for the effects of the selection procedure. The selection procedure comprises sampling, unit and item nonresponse, and coverage issues. The effect of sampling on variance (arrow 1 in figure 1) and bias (arrow 3) in the estimation of means is perhaps the most well-known topic in survey research (Neyman, 1934; Cochran,

[1] http://www.gesis.org/en/services/data/portals-links/comparative-survey-projects/

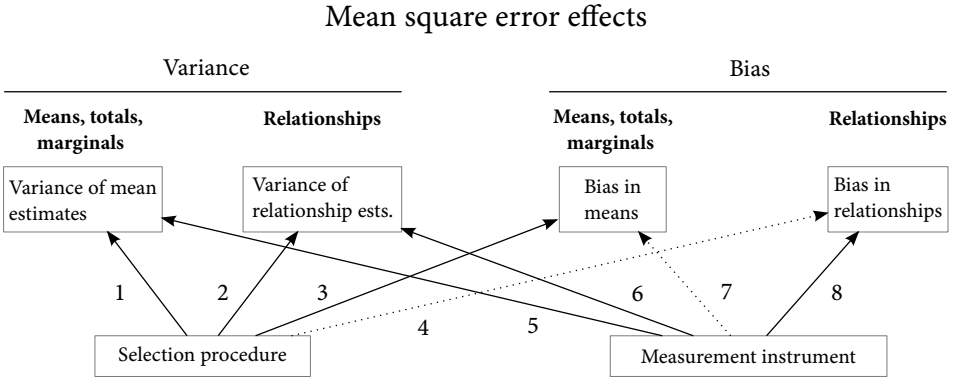


Figure 1: Effects of the measurement instrument and selection procedure on the estimation of means and relationships.

1977). The effect of complex sampling designs on the estimation of linear regression coefficients (arrows 2 and 4) was discussed by Scott and Holt (1982) and extended to structural equation models by Muthén and Satorra (1995).

More recently, growing nonresponse rates in household surveys have driven an increasing interest for the effect of nonrandom selection processes such as coverage, unit nonreponse, item missingness on bias in means (Little & Rubin, 2002; Groves & Couper, 1998; Groves, 2002). It remains an open question for any given variable whether a bias due to nonresponse can be expected, although many examples of bias in means have been encountered in empirical studies (e.g. Stoop, 2005; Stoop et al., 2010). On the other hand, the few studies that have examined bias in relationships (arrow 4 in the figure) were unable to find bias in relationships due to nonresponse (Goudy, 1976; Voogt, 2004).

A possible explanation for these findings is that, while a relationship between participation and the target variable is sufficient to cause nonresponse bias in means, nonresponse bias in relationships requires an interaction between one of the target variables and participation (Groves & Couper, 1998, chapter 2). This of course does not rule out the possibility that such a bias might exist (Groves & Peytcheva, 2008, 182), but does suggest that nonresponse can be expected to be play a smaller role for bias in relationships than it does for the goal of estimating means.

The effect of nonresponse on the variance of means and relationships (arrows 1 and 2) has been treated theoretically by Rubin (1987) as the so-called "proportion of missing information". The effect of nonresponse on variance of estimators without assuming equal variances for respondents and nonrespondents was discussed by Tångdahl (2005). The literature on variance increase due to nonresponse weighting and adjustments can also be placed in this category (Little & Rubin, 2002; Little & Vartivarian, 2006).

The effect of the measurement instrument is symbolized in figure 1 by arrows 5–8. It comprises interviewer effects and systematic and random measurement errors.

Interviewer effects on the variance of mean and relationship estimates have been studied in the past in the same way as clustering effects in sampling (Kish, 1962). Specific research designs, first introduced by Mahalanobis (1946), are generally necessary to estimate such effects so as not to confound interviewer effects with other factors (Hagenaars &

Heinen, 1982; R. Schnell & Kreuter, 2003). Interviewer effects on bias in means has been studied by Cannell et al. (1981); Fowler and Mangione (1990); Smit et al. (1997); Dijkstra and Van der Zouwen (1982), while bias in relationships due to the interviewer is, to my knowledge, a topic open for empirical study.

The systematic components of measurement error may also cause bias in means (arrow 7). For example, socially desirable behavior or yea-saying may cause respondents to provide an answer close to the perceived norm independently of their true opinion (Tourangeau et al., 2000). Solutions to reduce such effects that have been suggested in the literature are random response (Warner, 1965), item count techniques (Droitcour et al., 1991), web surveys, and anonymity (Dillman, 2007; Tourangeau & Smith, 1996).

Measurement error increases the variance of variables and thus the sampling variance of means and marginals (arrow 5). Biemer et al. (2004) gave a design-type of effect of measurement error on the variance of means. Such effects of measurement error on the variance of means do not require a special correction as they are subsumed by regular sampling theory.

Finally, measurement error may both cause bias and variance increase in relationships. This is the general theme in which the four chapters of this book may be placed.

The biasing effect of random and systematic measurement error on parameters of linear models is well-known (Fuller, 1987; Andrews, 1984). Some common nonlinear models are discussed in Carroll et al. (2006). In general one can say that random measurement error will decrease correlations, while stochastic systematic error will tend to increase correlations (Saris & Gallhofer, 2007a). This does not mean, however, that the bias in multiple regression coefficients will also be in these directions: the parameters of linear models depend on the observed correlations in a nonlinear way (Fuller, 1987, chapters 1 and 4). For this reason it is essential to estimate and correct for measurement error whenever relationships are to be studied.

In comparative surveys, the amount of measurement error may differ across countries. This causes the bias in relationships to differ also, rendering comparisons of relationships across countries invalid. **Chapter one** attempts to provide an explanation for cross-country differences in the quality of measures in the European Social Survey, where large cross-country variation was found (Oberski et al., 2007). It is shown how the discrete and non-interval nature of measurements may provide an explanation for this variation. In addition the role of systematic measurement error in the form of method variance is highlighted.

The analyses of chapter one suggest that estimation of the quality of survey measures should, when appropriate, take into account that some variables are discrete and do not have an interval measurement level. In practice in such cases often the ordinal confirmatory factor analysis (CFA) model (Muthén & Christoffersson, 1981) is used, meaning the factor analysis of so-called polychoric correlations (Pearson, 1900; Jöreskog, 1994). This model is equivalent to the two parameter normal ogive model of Lord (1952) in Item Response Theory (Christoffersson, 1975; Muthén, 1978).

The ordinal CFA model is highly restrictive in form since a normality assumption is made on the latent response variables. Equivalently, one may say that the response probabilities of categories in the ordinal CFA model are cumulative and additive and are sums of standard normal cumulative density curves with equal steepness. An alternative and less restrictive model exists in the form of the latent class factor model (Vermunt & Magidson, 2004a). The latent class factor model is a special case of the latent class model (Lazarsfeld

& Henry, 1968), in which instead of one nominal latent variable several latent variables are specified which are discrete but have interval level measurement (Heinen, 1996). The observed variables are then treated as discrete with nominal or ordinal level measurement (Hagenaars & McCutcheon, 2002; Vermunt & Magidson, 2005b).

Chapter two applies the latent class factor model to an existing design for the estimation of measurement error, the multitrait-multimethod (MTMM) design (Campbell & Fiske, 1959). By combining the latent class factor model with the MTMM design a new model, the so-called "latent class MTMM model", is developed. This model can be employed to estimate measurement error in discrete and noninterval-level survey questions under fewer assumptions than made in the ordinal CFA model. The use of the model is demonstrated by application to an MTMM experiment in the European Social Survey, and its utility for cross-national analysis is demonstrated by comparing the results for two countries.

The first part of the book thus deals with the issue of discrete and non-interval level measurement error models in comparative surveys. Chapters three and four in the second part of the book both treat the influence of measurement error on the sampling variance of relationships (arrow 6 in the figure).

Some techniques for taking these effects into account were discussed by Fuller (1987, chapter 4) for multivariate regression, and by Carroll et al. (2006) for logistic regression and other generalized linear models. A more general formulation of linear models, which encompasses most of the models discussed by Fuller (1987), is given by structural equation models (SEM) (Jöreskog, 1970; Bollen, 1989).

In SEM, measurement error-related parameters and "structural" regression parameters can be estimated simultaneously. In this case standard errors of regression parameters will automatically take into account the estimation of the measurement error-related parameters. A comment in the literature on the effect of the level of measurement error on the standard error of structural parameters in the context of a SEM was made by Heise (1970, 15–18), who went on to state that "one might attempt to work out mathematically the relationship between measurement error [and] sampling error, (…) but resulting formulas would be complicated and difficult to interpret." (p. 16).

Chapter three proposes to leverage the theory of general SEM models to separate out the effect on the variance of estimates of measurement error, sampling error, interviewer clustering and other components. It provides a method of judging the percentage of variance contributed by each error source without the need for Monte Carlo simulation. A disadvantage of the method presented is that it is model-dependent. An advantage is that, conditional on the model, one can judge the relative importance of different survey error components. An example application is given.

As mentioned before, when measurement error and structural parameters are estimated simultaneously, standard errors automatically take into acocount the uncertainty in the estimation of measurement error. But there are several reasons why it is not always possible or practical to perform a simultaneous estimation. First, the model may become very large due to the need for multiple indicators for each of the latent variables of interest. In addition, while the main interest of a substantive researcher will lie in the structural relationships, such analyses require a certain amount of expertise on measurement error models. Finally, social surveys often provide only one rather than multiple measures of a particular concept, so that the estimation of measurement error is impossible.

When it is not possible or desireable to subsume the estimation of measurment error

into the model, it is still possible to correct for measurement error using SEM. An external estimate based on a previous study is then required to correct for measurement error. Such estimates are sometimes available for a particular question and population (e.g. Coromina et al., 2008; Alwin, 2007, appendix), or a prediction may be obtained from meta-analysis through the program SQP (Saris, van der Veld, & Gallhofer, 2004; Oberski et al., 2004; Saris & Gallhofer, 2007b). The correction can then proceed by specifying latent variables with single indicators and fixing the measurement error variance parameters to the estimated values (e.g. Hayduk, 1987).

The single indicators or 'two-step' approach is sometimes more convenient but also has a problem: standard errors of structural model parameters do not take into account that the measurement error estimates are only estimates. In general confidence intervals will be too narrow and inference is affected. This problem had not been solved to date for general structural equation models.

Chapter four solves the problem of underestimated standard errors in the single indicators case by providing an exact formula for the asymptotic variance-covariance matrix of SEM estimates obtained by the single indicators approach. The form and implications of this analytical solution are discussed and a Monte Carlo study shows that inference is only correct when this correction is applied, demonstrating its validity.

Chapter overview

Chapter one:

Published as:
D. L. Oberski, W. E. Saris & J. A. P. Hagenaars (2009)
"Categorization Errors and Differences in the Quality of Questions Across Countries". Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC). T. D. Johnson and M. Braun (eds.). New York: John Wiley & Sons.

Chapter two:

D. L. Oberski, J. A. P. Hagenaars & W. E. Saris
"The Latent Class Multitrait-Multimethod Model".
In review process.

Chapter three:

D. L. Oberski
"Joint Estimation of Survey Error Components in Multivariate Statistics".
In review process.

Chapter four:

D. L. Oberski and A. Satorra
"Measurement Error Models with Uncertainty about the Error Variance".
In preparation for submission.

Chapter 1

Categorization errors and cross-country differences in the quality of questions

Abstract

The European Social Survey (ESS) has the unique characteristic that in more than 20 countries the same questions are asked and that within each round of the ESS Multitrait-Multimethod (MTMM) experiments are built in to evaluate the quality of a limited number of questions. This gives us an exceptional opportunity to observe the differences in quality of questions over a large number of countries. The MTMM experiments make it possible to estimate the reliability, validity, and method effects of single questions (Andrews, 1984; Saris, Satorra, & Coenders, 2004; Saris & Andrews, 1991). The product of the reliability and the validity can be interpreted as the explained variance in the observed variable by the variable one would like to measure. It is a measure of the total quality of a question.

These MTMM experiments showed that there are considerable differences in measurement quality across countries. Because these differences in quality can cause wrong conclusions with respect to differences in relationships across countries, this paper studies the quality of the measures from the viewpoint of categorization. We assume that each category represents a range of scores on a latent continuous variable that have been grouped together, causing grouping errors. It depends on the distribution of values of the latent response variable in each category whether the intervals between the categories are equally far apart. If they are not, there is also transformation error. Both grouping and transformation are sources of measurement error due to categorization and therefore possible explanations for differences in the quality of questions. The results show that this effect is quite strong.

Introduction

Measurement error can invalidate conclusions drawn from cross-country comparisons if the errors differ from country to country. For this reason, when different groups such as countries are compared with one another, attention should not only be given to absolute levels of errors, but also to the differences between the groups. Different strategies have been developed to deal with the problem, for example within the context of invariance testing in the social sciences (Jöreskog, 1971), differential item functioning in psychology (Muthén & Lehman, 1985), and differential measurement error models in epidemiology and biostatistics (Carroll et al., 1995).

In the ESS a lot of time, money, and effort is spent to make the questions as functionally equivalent across countries as possible (Harkness et al., 2002) and to make the samples as comparable as possible (Häder & Lynn, 2007). Nevertheless, considerable differences in quality of the questions can be observed across countries. To study these differences is important because they can cause differences in relationships between variables in different countries which have no substantive meaning but are just caused by differences in quality in the measurement (Saris & Gallhofer, 2007a). In order to avoid such differences it is also important to study the reasons behind them.

In an earlier study, we investigated differences in translations, differences in the experiments' design, and differences in the complexity of the question as possible reasons for differences in question quality across countries (Oberski et al., 2007). Because these factors did not explain much of the differences we now consider differences in categorization errors as a source of differences between countries.

Categorization errors are part of the discrepancy between an unobserved continuous variable and a discrete observed variable that measures the unobserved continuous variable. Specifically, categorization errors are the differences between the score on the latent variable and the observed category that are due solely to the categorization process.

For example, suppose a person's age is known only to belong in one out of three categories, which are assigned the scores one, two, and three, but there are never any mistakes in this categorization. In spite of the absence of mistakes, there is still a discrepancy between the age of the person and the category she is assigned to; first, because people of different ages have been lumped together. And second, the distance between the categories in terms of average age may not be equal to the distances of unity between the numbers one, two, and three, assigned to the categories. This means that if one treats the observed variable as an interval level measure, the result of calculations such as correlations will differ also from what would have been obtained if the original age variable had been used.

In general, one can say that categorization errors arise when a continuous latent response variable is split up into different categories. This leads to two types of errors: grouping and transformation errors (Johnson & Creech, 1983). Grouping errors occur when different opinions are grouped together in the same category. Transformation errors occur when the differences between the numerical values of adjacent categories do not correspond to equal distances between the means of the latent response variables in those categories. If, for instance, the distances between categories are not the same in two different countries, this can lead to larger categorization errors in one country than another, leading in turn to lower question quality. This is why the distance between categories is a possible explanation for differences in question quality across countries.

The first section will discuss the models we use to estimate the measurement error co-

efficients of survey questions starting from a basic response model. We will then present the data from the European Social Survey that will be used. A short discussion of previous results follows. First the estimates from our previous research are shown. In a previous study, we already examined some possible explanations for the large differences in these estimates found across countries. These will be shortly reviewed. We then go on to present the model that will be the focus of this study, which accounts for categorization errors. It will be shown what we mean by such errors and how we compare the results we get from categorical models with those from continuous models. The statistical method of estimation is presented, after which we discuss our results. Since we have many such results, they are followed by a meta-analysis of the results. Finally, we discuss our general conclusions from this meta-analysis.

1.1 Theory

In Figure 1.1 we show the basic response model (Saris & Gallhofer, 2007a) we use as our starting point.

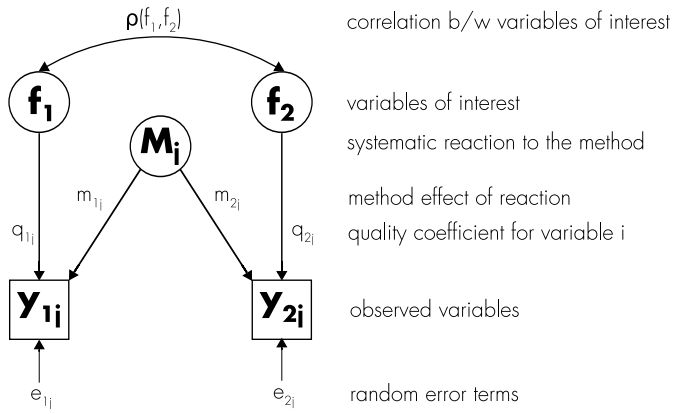


Figure 1.1: The continuous response model used in the MTMM experiments.

The difference between the observed response (y) and the variable of interest or concept by intuition (f) is both random measurement error (e) and systematic error due to the respondent's reaction to the method (M). This method effect is the only systematic error considered in the model.

An example of a method effect is when each respondent chooses her own reference points for all 11 point scales. For instance, an 11 point agree-disagree scale might label the highest category with the text 'disagree'. But in principle one can also disagree 'strongly' or even 'completely', although such opinions are not marked with a number on the answer scale. Thus it is up to the respondent to choose a location for these most extreme reference points. This choice will influence which category is finally chosen, given any opinion. Different reference points are generally chosen by different people if these points are not fixed by the question, causing non-substantive random variation. If the same reference points are chosen by the same people given the same answer scale, then there will also be a correlation between the answers to all 11 point scales that has nothing to do with the

respondents' opinions (Saris, 1988). This systematic variation can be considered method variance.

The coefficient q represents the quality coefficient and we call q^2 the total quality[1]. This quality–sometimes also called the reliability ratio–equals $\frac{Var(f)}{Var(y)}$: it can be interpreted as the proportion of variation in the observed variable that is due to the unobserved trait of interest. The correlation between the unobserved variables of interest is denoted by $\rho(f_1, f_2)$.

Several remarks should be made. The first is that the correlation $\rho(y_{ij}, y_{kj})$ between two observed variables measured with the same method is:

$$\rho(y_{ij}, y_{kj}) = \underbrace{\rho(f_i, f_k)}_{\textit{Correlation of interest}} \cdot \underbrace{q_{ij} \cdot q_{kj}}_{\textit{Attenuation factor}} + \underbrace{m_{ij} \cdot m_{kj}}_{\textit{Correlation due to method}} \tag{1.1}$$

where $i \neq k$ index the concepts by intuition and j a method.

This means that the correlation between the observed variables is normally smaller than the correlation between the variables of interest, but can be larger if the method effects are considerable. A second remark is that one can not compare correlations across countries without correction for measurement error if the measurement quality coefficients are very different across countries: this follows directly from the above equation (1.1). A third point is that one can not estimate these quality indicators from this simple design with two observed variables. In this model there are two quality coefficients, two method effects, and one correlation between the two latent traits, leaving us with five unknown parameters, while only one correlation can be obtained from the data. It is impossible to estimate these five parameters from just one correlation.

There are two different approaches to estimate these coefficients. The first is direct estimation from MTMM experiments. The second is the use of the prediction program SQP. SQP predicts the quality coefficient and method effect of a single question from many of its characteristics such as the topic, the number of categories, etc[2]. It is currently based on a meta-analysis of 87 MTMM experiments and 1028 different questions, while many more experiments are soon to be added (Oberski et al., 2004). In this study we use the MTMM approach.

Campbell and Fiske (1959) suggested using multiple traits and multiple methods to evaluate the quality of measurement instruments (MTMM). The classical MTMM approach recommends the use of a minimum of three traits that are measured with three different methods leading to nine different observed variables. An example of such a design is given in Table 1.1. Given the responses on all the variables, the coefficients described above can be estimated. A more elaborate introduction to MTMM and SQP can be found in Saris and Gallhofer (2007).

1.2 Data

The European Social Survey (ESS) has the unique characteristic that in more than 20 countries the same questions were asked and that within each round of the ESS Multitrait-

[1] One can also separate the reliability and method variance. This response model is known as the true score model and is more easily interpreted in terms of classical test theory, but mathematically equivalent to the classic MTMM model used here. For more details of the different models we refer to (Saris & Andrews, 1991)

[2] See the website `http://www.sqp.nl/`

Table 1.1: The classic MTMM design used in the ESS pilot study.

The three traits were presented by the following three items:

- *On the whole, how satisfied are you with the present state of the economy in Britain?*
- *Now think about the national government. How satisfied are you with the way it is doing its job?*
- *And on the whole, how satisfied are you with the way democracy works in Britain?*

The three methods are specified by the following response scales:

(1) Very satisfied; (2) Fairly satisfied; (3) Fairly dissatisfied; (4) Very dissatisfied

Very dissatisfied *Very satisfied*
0 1 2 3 4 5 6 7 8 9 10

(1) Not at all satisfied; (2) Satisfied; (3) Rather satisfied; (4) Very satisfied

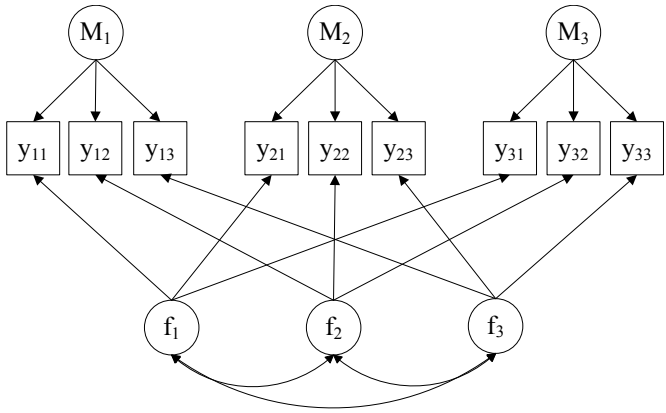


Figure 1.2: MTMM model illustrating the observed scores and their factors of interest.

Multimethod (MTMM) experiments are built in to evaluate the quality of a limited number of questions. This gives us an exceptional opportunity to observe the differences in quality of questions over a large number of countries. In this paper we have used the MTMM experiments of round 2 of the ESS, collected in 2004.

The questionnaires were administered by face to face interviewing in all countries. In Finland, France, Ireland, Italy, the Netherlands, Norway, and Sweden, the supplementary questionnaire with the repetition questions was self-completed with the interviewer present rather than asked face to face. This confounds mode effects with country effects for these countries. The countries we compare in subsequent sections all used face to face interviewing for both questionnaires, however. Therefore mode effects are not an issue in this particular study.

The topics of the four MTMM experiments from the ESS we will study were the following:

1. The social distance between the doctor and patients;

2. Opinions about job;

3. The role of men and women in society;

4. Political efficacy.

Concerning each of these topics three questions were asked and these three questions were presented in three different forms following the discussed MTMM designs. The first form, used for all respondents, was presented in the main questionnaire. The two alternative forms were presented in a supplementary questionnaire which was completed after the main questionnaire. All respondents were only asked to reply to one alternative form but different groups got different version of the same questions (Saris, Satorra, & Coenders, 2004). For the specific questions in the experiments we refer to the ESS website where the English source version of all questions are presented[3], and for the different translations we refer to the ESS archive[4].

Each experiment varies a different aspect of the method by which questions can be asked in questionnaires. The 'social distance' experiment examines the effect of choosing arbitrary scale positions as a starting point for agreement-disagreement with a statement. The 'job' experiment compares a four point true-false scale with direct questions using 4 and 11 point scales. In the 'role of women' experiment agree-disagree scales are reversed, there is one negative item, and a 'don't know' category is omitted in one of the methods. Finally, the political efficacy experiment pitted agree-disagree scales against direct questions.

A special group took care that the samples in the different countries were proper probability samples and as comparable as possible (Häder & Lynn, 2007).

The questions asked in the different countries have been translated from the English source questionnaire. An optimal effort has been made to make these questions as equivalent as possible and to avoid errors. In order to reach this goal two translators independently translated the source questionnaire and a third person was involved to choose the optimal translation by consensus if differences were found. For details of this procedure we refer to the work of Harkness et al. (2002).

Despite these efforts to make the data as comparable as possible, large differences in measurement quality were found across the different countries. Table 1.2 shows the mean and median standardized quality of the questions in the main questionnaire across the experiments for the different countries.

A remarkable phenomenon in this table is that the Scandinavian countries have the lowest quality of all while the highest quality has been obtained in Portugal, Switzerland, Greece, and Estonia. The other countries are in between these two groups. The differences are considerable and statistically significant across countries ($F = 3.19$, $df = 16$, $p < 0.001$) and experiments ($F = 92.65$, $df = 5, p < 0.0001$). The highest mean quality is 0.79 in Portugal while the lowest is 0.57 in Finland. If the correlation between the constructs of interest is 0.60 in both countries and the measures for these variables have the above quality then the observed correlation in Portugal would be 0.47 while the observed correlation in Finland would be 0.34. Most people would say that this is a large difference

[3]http://www.europeansocialsurvey.org
[4]http://ess.nsd.uib.no

Table 1.2: The quality of all 18 questions included in the experiments in the main questionnaire.

Country	Mean	Median	Minimum	Maximum
Portugal	0.79	0.81	0.63	0.91
Switzerland	0.79	0.84	0.56	0.90
Greece	0.78	0.79	0.64	0.90
Estonia	0.78	0.85	0.58	0.90
Poland	0.73	0.85	0.51	0.90
Luxembourg	0.72	0.73	0.53	0.88
United Kingdom	0.70	0.71	0.56	0.82
Denmark	0.70	0.70	0.52	0.80
Belgium	0.70	0.73	0.46	0.90
Germany	0.69	0.70	0.53	0.83
Spain	0.69	0.64	0.54	0.90
Austria	0.68	0.68	0.51	0.85
Czech Republic	0.65	0.60	0.52	0.87
Slovenia	0.63	0.60	0.46	0.82
Norway	0.59	0.59	0.35	0.83
Sweden	0.58	0.58	0.43	0.68
Finland	0.57	0.54	0.42	0.78

in correlations which requires a substantive explanation. But this difference can be expected because of differences in data quality and has no substantive meaning at all. Not all of these differences are necessarily due to categorization, however. Below we discuss other possible explanations for some the differences.

1.3 Explanations for cross-country differences in question quality

The previous section showed that in some cases large differences were found in question quality across the countries of the ESS. In a previous study, we examined a few possible explanations of these discrepancies (Oberski et al., 2007).

The first explanation we studied were errors in the translation. Although in the ESS a lot of care has been taken to ensure the correct translation of the questions, we found that a few questions in the supplementary questionnaire had not been translated in the way intended. In particular, one item in the 'social distance' experiment had been translated in all French questionnaires as 'Doctors rarely tell their patients the whole truth' rather than 'Doctors rarely keep the whole truth from their patients'. Since these sentences have opposite meanings, it is unsurprising that we should find a different relationship with the trait of interest.

Another alternative explanation for differences across countries is differences in the implementation of the experimental design. Here one difference existed between the implementations in Norway, Sweden, and Finland, and the other countries: in these countries respondents could send in the supplementary questionnaire containing the repetitions at a time chosen by themselves, while the general design used in other countries was that the supplementary questionnaire should be administered directly after the main interview.

Some respondents waited quite some time before answering the supplementary questions. In the time between the two interviews their opinions may have changed, or have been influenced by new considerations unique to that moment. An MTMM analysis of a sample split according to whether the questionnaire was returned within two days or later provided strong evidence that this was indeed the case. In fact, the sample of people who had returned the questionnaire on the same day was by itself very similar in the quality to other countries.

The third alternative we considered was that the language of the questions might be more complex in one language than in another. Previous meta-analyses found that language complexity can have an effect on the quality (Saris & Gallhofer, 2007b). However, we found no strong evidence that the complexity of the questions could explain the differences in question quality in this case.

Thus, in some cases we found artificial differences in quality which are likely to be due to an erroneous translation or different implementation of the experimental design–notably in the Scandinavian countries except Denmark and for one item in the French-speaking countries. However, these cases are not so numerous that they can explain the large overall variations in question quality found in the ESS. Therefore we now turn to the possibility that the distance between the categories in the categorical questions differs from country to country. Before we proceed to investigate the influence of categorization errors on the quality in different countries and experiments, we explain in more detail the model used to estimate the distances between the categories.

1.3.1 The categorical response model

The response model discussed so far makes no mention of the fact that many of the measures we use are in fact ordinal–that is, they are most likely ordered categories rather than measured on an interval scale. Broadly speaking, two types of measurement models have been proposed for this situation. The first assumes that there is an unobserved discrete variable, and that errors arise because the probability of choosing a category on the observed variable given a score on the unobserved variable is not equal to one. That is, the errors are modelled by the conditional chances of choosing a category on the survey question given the unobserved score. Such models are often referred to as latent class models (Lazarsfeld & Henry, 1968; Hagenaars & McCutcheon, 2002).

The second approach deals with the case where a continuous scale or 'latent response variable' (LRV) is thought to underly the observed categorical item. Such models are sometimes called latent trait models. Several extensions are possible, but we focus on a special case described by Muthén (1984). This is the model we will use in our subsequent analysis of the data (figure 1.3)[5].

Errors may arise at two stages. The first is the connection between the latent response variable (LRV_{ij} in figure 1.3) and its latent trait (f_i). This part of the error model is completely analogous to factor analysis or MTMM models for continuous data: the scale is modeled as a linear combination of a latent trait (f_i), a reaction to the particular method

[5]It can be shown that analysing polychoric correlations in an MTMM model is a special case of the model we use (Muthén & Asparouhov, 2002). However, we do not use polychoric correlations because it would be necessary to assume that the variances of the latent response variables are equal across countries. Since we try to separate categorization errors from differences in the continuous part of the model, this is not a desirable assumption. The model we use is equivalent to a multi-dimensional two parameter graded response model in item response theory (Muthén & Asparouhov, 2002).

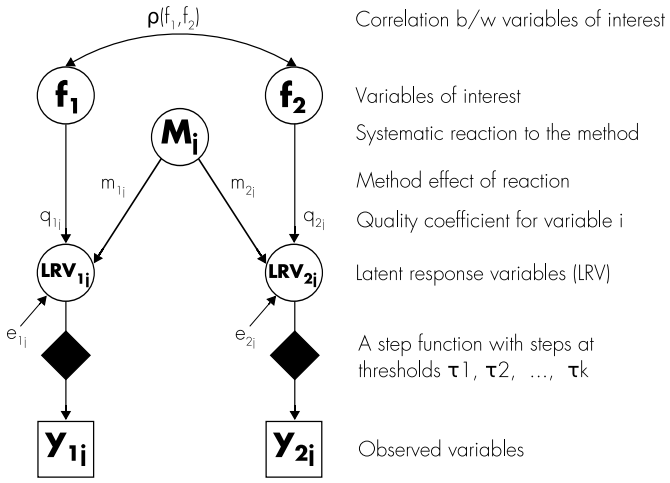


Figure 1.3: The categorical response model used in the MTMM experiments.

used to measure the trait (M_j), and a random error (e_{ij}), and interest then focuses on the connection between the trait and the scale (q_{ij}), which we again term the 'quality coefficient' (see also figures 1.1 and 1.2).

The second stage at which errors arise differs from the continuous case. This is the connection between the variables LRV_{ij} and y_{ij} in figure 1.3. Here the continuous latent response variable is split up into the different categories, such that each category of the observed variable corresponds to a certain range on the unobserved continuous scale. The sizes of these ranges are determined by threshold parameters. In figure 1.3 this step function has been represented by a black triangle. Examples of step functions are illustrated in figure 1.4.

In figure 1.4, the steps (solid line) show the relationship between the LRV and the observed variable, while the straight (dotted) line plots the expectation of the LRV given the latent trait. In the step function on the left-hand side, the LRV has been categorized using equal intervals. The error that is added by the categorization is the vertical distance between the dotted line and the step. That is, the distance between the dotted line and the horizontal segments of the solid line. It can be seen that the error is zero when the straight line crosses the steps, and that at each step, the error is the same (at 3, 6, and 9). The expectations within the categories have the same interval as the thresholds of unity, and so if the values 1, 2, 3, and 4 are assigned to the categories, no transformation occurs. Errors still occur, because the values along the dotted line have been grouped into the four categories formed by the solid line. Relationships of the observed categorical variable with other variables will therefore be attenuated.

Conversely, the right hand side shows a latent response variable that has been categorized with unequal steps. The figure shows that the distances between the thresholds τ_1, τ_2, and τ_3 are very different from each other. The consequence is that at the second step, i.e. in between τ_2 and τ_3, there is almost no extra error, while at the first and third steps the errors are much larger. Here a transformation occurs. Suppose that the categories are given the numerical values 1, 2, 3, and 4, as is often done. Then the distances between the expectations of the LRV in each of the categories do not equal unity, which is the distance

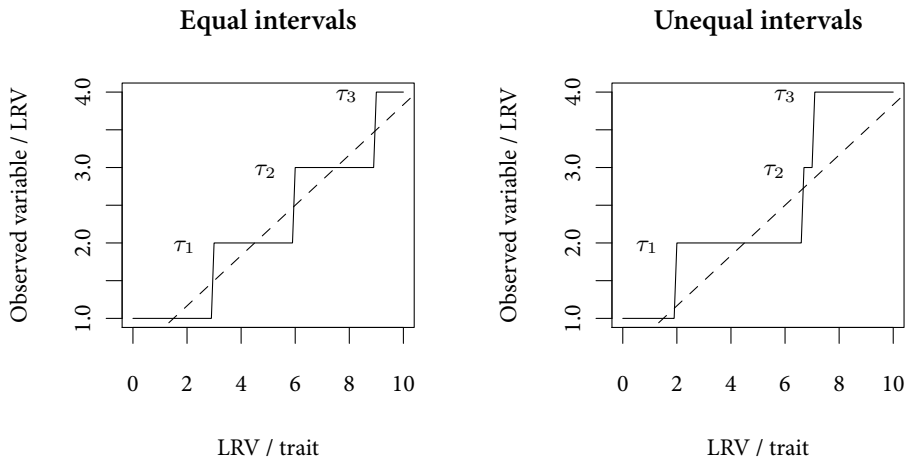


Figure 1.4: Two hypothetical step functions which result from categorization. The solid lines plot the observed categorical variable as a function of the latent response variable (LRV). The diagonal dotted lines plot the expectation of the LRV as a function of the latent trait on the same scale. The thresholds used for categorization are denoted by the symbols τ_1, τ_2, and τ_3.

between the values chosen for categories.

To sum up, two types of errors can be distinguished at this stage (Johnson & Creech, 1983):

1. *Grouping errors* occur because the infinite possible values of the latent response variable are collapsed into a fixed number of categories (the vertical distances between the diagonal line and the steps in figure 1.4). These errors will be higher when there are fewer categories;

2. *Transformation errors* occur when the distances between the numerical scores assigned to each category are not the same as the distances between the means of the latent response variable in those categories. This happens when the thresholds are not equally spaced, or when the available categories do not cover the unobserved opinions adequately.

We have described the categorization process here. Is is important to note, however, that normally this process is not observed and one only observes a discrete variable, which we then assume is the result of this process.

Categorization, then, can be expected to be another source of measurement error besides random errors and method variance. If these errors differ across countries, then so will the overall measurement quality, and differences in means, correlations, regression coefficients, and cross-tables across countries result which are due purely to differences in measurement errors.

Thus, the model we use allows to a certain extent for the separation of errors due to the categorization, errors due to the reaction to the method and random errors. In this paper

we take advantage of this separation to compare the amount of error due to categorization introduced across countries.

1.3.2 Categorization errors in survey questions

The previous sections showed that, using the MTMM design, it is possible to obtain a measure (q^2) of the total quality of a question. If a continuous variable model (hereafter referred to as CV model) is used, this quality is influenced by errors in both stages of the categorical response model: not only random errors and method effects are included, but also errors due to the categorization. For this reason Coenders (1996) argued that the linear MTMM model assuming continuous variables does not ignore categorization errors, but absorbs them to a certain extent in the estimates of the random error and method correlations. How this absorption functions exactly will depend on the model in use and is not extensively studied. The extent to which it holds in general is thus a topic that is still under discussion.

However, since the quality coefficient is estimated from the covariance matrix of the measures, it can be both reduced and increased by categorization errors. In general all correlations between measures increase after correction for categorization, but they need not all increase equally. If categorization errors are higher using the first method, the correlations between the latent response variables using this method will increase more relative to the observed correlations than the correlations of each variable with its repetition using a different method. In this case the amount of variance in the response variable due to the method will be larger in the categorical model than in the CV model, and the estimated quality of the measure in the categorical response model can become lower than the estimated quality in the continuous MTMM model. This is because there are method effects (correlated errors) on the level of the continuous latent response variables which do not manifest themselves in the observed (Pearson) correlations between the categorical variables. Categorization can therefore in some cases inflate estimates of the quality of categorical observed variables, even though, at the same time, it causes errors which reduce the quality. There are thus two processes at work, which have opposite effects on the estimates of the quality. a

As noted before, the quality of a variable is defined as the ratio of the true trait variance to the observed variance (see also figure 1.1 in the first section):

$$q^2 = \frac{Var(f)}{Var(y)}. \tag{1.2}$$

However, we have now seen that y is itself a categorization of an unobserved continuous variable (c), and therefore the above equation 1.2 can be 'decomposed' into

$$q^2 = \frac{Var(f)}{Var(LRV)} \cdot \frac{Var(LRV)}{Var(y)}. \tag{1.3}$$

The scale of LRV, the latent response variable, is arbitrary, except that it may vary across countries due to relative differences in variance (Muthén & Asparouhov, 2002). However, the ratio $Var(LRV)/Var(y)$ can easily be calculated once q^2_{con}, the quality from the continuous analysis, and $Var(f)/Var(LRV)$, the quality from the categorical MTMM anal-

ysis ($q^2_{cat.}$), have been obtained. So equation (3) shows that $q^2_{con} = q^2_{cat} \cdot c$ and

$$c = \frac{q^2_{con}}{q^2_{cat}},$$

where c is the categorization effect, or (assuming $q^2_{cat}, q^2_{con} > 0$)

$$\ln(q^2_{con}) = \ln(q^2_{cat}) + \ln(c).$$

This correction factor is a useful index of the relative differences between the quality estimates of the continuous and categorical models.

In the present study, we estimate this 'categorization factor' for different countries and experiments, and examine to what extent it can explain the differences in quality across countries.

1.4 Methods

In almost every country of the ESS, respondents were asked to complete a supplementary questionnaire containing the repetitions used in the experiments. Not all respondents completed the same questionnaire. The sample was randomly divided into subgroups, so that half of the people answered the first and second form of the questions, and the other half answered the first and third form.

This so-called split-ballot MTMM approach lightens the response burden by presenting fewer questions and fewer repetitions. Saris, Satorra, and Coenders (2004) showed that the different parameters of the MTMM model can still be estimated using this planned missing data design. If the different parts of the model are identified, so is the entire model. Since we can identify the necessary covariances in the categorical model, this is identified as well (Millsap & Yun-Tein, 2004).

For each experiment, two different models were estimated. The continuous analysis was conducted using the covariance matrices as input, and estimated using the maximum likelihood estimator in LISREL 8. The results presented in the tables below were standardized after the estimation.

The categorical model can in principle also be estimated using maximum likelihood. However, in order to deal with the planned missing data (split-ballot) a procedure such as full-information maximum likelihood would be necessary. This requires numerical integration in the software we used (Mplus 4), making the procedure prohibitively slow and imprecise. We therefore used an alternative two step approach, whereby in the first step the covariance matrices of the latent response variables were estimated, and in the second step the MTMM model is fitted to the estimated matrices. The estimation in the first step was done using the weighted least squares approach described by Flora and Curran (2004), and the second step again employed the maximum likelihood estimator[6].

This approach has the advantage that consistent and numerically precise estimates can be obtained within seconds rather than days (Muthén & Asparouhov, 2002). The disadvantages are that the standard errors of the estimates of the categorical MTMM model are

[6]We note here that the categorical MTMM model is equivalent to the 'graded response model' in item response theory. There is a simple relationship between the threshold and quality coefficients of our model and difficulty and discrimination parameters in IRT models: the quality coefficients are scaled discrimination parameters, while a scaled difficulty for each category can be obtained by dividing each threshold by the corresponding quality coefficient (Muthén & Asparouhov, 2002).

incorrect, and that the chi-square statistic and modification indices may be inflated. Although the problem could in principle be remedied by using the asymptotic covariance matrix of the covariances as weights in the estimation (Jöreskog, 1990), in the present paper we compare only the consistent point estimates of this model.

We model categorization errors using threshold parameters. These thresholds are the theoretical cutting points where the continuous latent response variable (LRV) has been discretized into the observed categories. If the thresholds are different across countries, the questions are not directly comparable, since differences in the frequency distribution are partly due to differences in the way the LRV was discretized. If the thresholds are the same across countries the questions may still not be comparable due to differences in linear transformations (loadings) and random errors. But in that case it is not categorization error that causes incomparability. A final possibility is that loadings, random errors, and thresholds are all the same across countries. In that case the frequency distributions can be directly compared.

In this paper we will perform only a basic invariance test on thresholds. If the thresholds are equal, categorization error is not a likely cause of differences in quality. However, we do not continue with tests for invariance on loadings and error variance, but will compare the results of the two different models.

The two models are the same with respect to the covariance structure of the response variables (the 'MTMM part' of the model). However, they differ in their basic assumptions about the 'observation part' of the model: the CV model assumes that the continuous response variables have been directly observed, while the categorical model assumes a threshold connection between the response variables and the observed ones.

Both models assume normality of the response variables, but the differences in basic assumptions cause the categorical model to be more sensitive to departures from normality. While in the CV model, under quite general conditions, violation of normality will not affect the consistency of the estimates (Satorra, 1990), this is not so in the categorical model. There, the threshold estimates are derived directly from quantiles of the normal distribution which the latent response variable is assumed to follow. Therefore, if the LRV's are not normally distributed, the threshold estimates will be biased. The MTMM estimates depend on the thresholds and can also change, though the precise conditions under which such estimates would change significantly have, to our knowledge, not been investigated analytically. It has been found in several different simulation studies that bias may occur especially when the latent response variables are skewed in opposite directions (Coenders, 1996).

Thus, while the categorical model may be more realistic in modelling the observed variables as ordinal rather than interval level measures, the CV model may be more realistic in that it is robust to violations of normality[7]. In any particular analysis, whether one or the other model provides a more adequate estimate of the quality of the questions therefore depends on the degree to which these assumptions are violated[8]. This should be kept in

[7]One important point to make here is that even when univariate distributions such as histograms and tables of the observed categorical variables are highly non-normal, this does not necessarily imply that the normality assumption of the categorical model is violated. The reason is that a very non-normally distributed observed variable may be the consequence of a perfectly normally distributed variable that has been categorized in a very uneven way.

[8]In principle the normality assumption on the latent response variables is testable. However, the question then still remains what impact any non-normality would have on the estimates. This question is beyond the scope of the present paper.

mind in the interpretations of the results.

We estimated the quality of the measures based on the CV model and based on the categorical model for four experiments which used an answer scale of five categories or less in the main questionnaire. For each experiment, the countries with the highest and the lowest qualities in the CV model were analysed. For each of the questions we took the ratio, called 'categorization factor', of the two different quality measures as an index of the effect that categorization has on the continuous quality estimates. The next section presents the results.

1.5 Results

1.5.1 Results of the experiments

The first experiment's results will be described in some detail, while we provide the results of the other experiments in the appendix.

The first experiment concerned opinions on the role of women in society (see table 1.3). We first turn to the hypothesis that all thresholds are equal across different countries. If this hypothesis cannot be rejected there is also little reason to think that the categorization is causing differences in the quality coefficients.

We selected the two countries with the highest and the country with the lowest quality coefficients. In this experiment, the wording of the question was reversed in the second method. For example, the statement 'When jobs are scarce, men should have more right to a job than women' from the main questionnaire was changed to 'When jobs are scarce, women should have the same right to a job as men' in the supplementary questionnaire. The countries with high quality coefficients were, in this case, Portugal and Greece. The lowest coefficients for this experiment were found in Slovenia. To be able to separately study misspecifications in the categorization part of the model, we imposed no restrictions on the covariance matrix of the latent response variables at this stage.

In the first analysis, all thresholds were constrained to be equal across the five countries. This yields a likelihood ratio statistic of 507 on 48 degrees of freedom. The country with the highest (128) contribution to this chi-square statistic is Portugal. When we examine the expected parameter changes, it also turns out that in this country these standardized values are very large with some values close to 0.9 while in other countries the highest obtained and exceptional value is 0.6. For some reason, the equality constraint on the Portuguese thresholds appears to be a particularly gross misspecification.

As it turns out, this particular misspecification is very likely due to a translation error. The intention of the experiment was to reverse the wording of the question in the second method. But in Portugal the reverse wording was not used, and the same version was presented as in the main questionnaire. To prevent incomparability when the MTMM model is estimated, we omit Portugal from our further analyses and continue with two countries.

The model where all thresholds are constrained to be equal yields a likelihood ratio of 351 and 36 degrees of freedom ($p < 0.00001$). This model should therefore be rejected: the thresholds are significantly different across countries.

We use the procedure of Saris et al. (2009) to determine whether misspecifications are present in the model. For this test we need the Expected Parameter Change (EPC), Modification Index (MI) and the power of the test. The EPC gives direct estimates of the size of

Table 1.3: The 'role of women' experiment: questions and threshold estimates (in z-scores).

'A woman should be prepared to cut down on her paid work for the sake of her family'

	1 *Agree strongly*	τ_1	2 *Agree*	τ_2	3 *Neither/nor*	τ_3	4 *Disagree*	τ_4	5 *Disagree strongly*
Slovenia		-1.4		-0.1		0.6		1.8	
Greece		-1.1		-0.2		0.5		1.4	

'A woman should not have to cut down on her paid work for the sake of her family.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-1.5		-0.0		0.6		2.0	
Greece		-1.5		-0.3		0.4		1.5	

'Men should take as much responsibility as women for the home and children.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-0.5		1.3		1.9		2.6	
Greece		-0.6		0.7		1.6		2.3	

'Women should take more responsibility for the home and children than men'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-1.7		-0.7		-0.2		1.2	
Greece		-1.6		-0.5		0.0		1.4	

'When jobs are scarce, men should have more right to a job than women.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-1.8		-0.8		-0.3		0.9	
Greece		-0.9		0.1		0.6		1.4	

'When jobs are scarce, women should have the same right to a job as men.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-0.8		0.7		1.1		1.9	
Greece		-1.1		-0.1		0.7		2.0	

the misspecification for all fixed parameters, while the MI provides a significance test for the estimated misspecification (Saris et al., 1987).

However, these two indices are not sufficient for determining misspecifications because the MI depends on other characteristics of the model. For this reason, the power of the MI test must be known in order to determine whether a restriction is misspecified. We use these quantities to incrementally free parameters that were indicated to be misspecified.

Using the modification indices and power as guides, we formulated a new model in which some thresholds were constrained to be equal, while others were freed to vary. Equality of thresholds is not required to estimate the relationships, but it is useful because the equality of thresholds allows for differences in variances of the response variables across the groups. This is in contrast with the use of polychoric correlations where the variances are constrained to be equal across the groups.

The resulting model has an approximate likelihood ratio of 2.8 on 2 degrees of freedom

Table 1.4: Quality (q^2) and method effects (m) according to the continuous and categorical models, with categorization factors for the experiment on opinions about the role of men and women in society.

		CutDown	'Women' Respnsib.	MenRight
Continuous analysis				
q^2	Greece	0.71	0.66	0.71
	Slovenia	0.54	0.25	0.68
m	Greece	0.15	0.15	0.15
	Slovenia	0.17	0.24	0.15
Categorical analysis				
q^2	Greece	0.51	0.35	0.48
	Slovenia	0.69	0.29	0.65
m	Greece	0.49	0.14	0.32
	Slovenia	0.33	0.75	0.19
Categorization factor				
	Greece	1.4	1.9	1.5
	Slovenia	0.8	0.9	1.0

($p = 0.24$)[9]. The resulting estimates of the threshold parameters are presented in table 1.3. These estimates have been expressed as z-scores in order to make them comparable.

Table 1.3 presents three different traits, each asked in two different forms. The first form of each trait is the form asked in the main questionnaire, while the second form was asked in the supplementary questionnaire (the third form has been omitted for brevity).

The thresholds in this model represent how extreme the 'agreement' has to be before the next category is chosen rather than the previous one. This strength is expressed in z-scores, i.e. standard deviations from the mean. Take, for instance, the third statement in the table: "Men should take as much responsibility as women for the home and children". Slovenians need to have an agreement differing from the country mean 2.6 times more than the standard deviation, before they will respond 'disagree strongly'.

Note that the threshold part of the relationship between LRV and observed response is deterministic. However, not all Slovenians with an *opinion* on the indicator of 2.6 standard deviations or more away from the mean will necessarily answer 'disagree completely'. This is so because the latent response variable is also affected by random measurement error. The combination of the threshold model and normally distributed random measurement error gives rise to a familiar probit relationship between indicator and response. Because the random error plays an important role in this relationship, not only the thresholds should be discussed here, but also the quality coefficients.

Looking at the first question, it can be seen that the distances between the thresholds are unequal for these two countries and different from one. One can also see that the endpoints are somewhat distant, especially in Slovenia: there the category 'disagree strongly' is 1.8 standard deviations or more away from the mean, reducing the number of scale points that are available for some people.

[9]It is also possible to free more parameters and put no restrictions at all on the model. This might lead us to find differences between countries more easily, since the parameters are allowed to vary. However, we prefer to aid our estimation by imposing these restrictions: if they do not hold in the population, this leads us to be conservative in ascribing differences between countries to the categorization.

The second form of the same question is similar to the first form in this respect, except that here both of the endpoints are rather distant in both countries, again reducing the number of scale points. As noted above, a reduction in scale points can be expected to increase grouping errors.

The second trait ('responsibility') presents a radically different picture. In both countries the 'disagree' and 'disagree strongly' categories are quite far away from the mean. This again reduces the number scale points, while, at the same time, the scale is cut off in this manner only from one side. Large transformation errors can be expected. Moreover, in Slovenia this effect is much worse than in Greece: the category 'neither disagree nor agree' is already 1.3 standard deviations or more away from the mean, reducing the amount of information provided by this variable in Slovenia even further.

The second phrasing of this question seems to provide a better coverage of the prevailing opinions on women and men's responsibility for the home and children.

For the third and last trait–the right to a job–the most striking feature of the thresholds is that in Slovenia, the first three categories represent opinions below the mean, while in Greece only the first category does. Beyond this, it is difficult to say which scale might produce fewer categorization errors. Surprising, however, is that the second form of the same question seems to produce much more comparable scales with respect to the thresholds than the first one.

It is also clear from the table that the two forms of phrasing are not exactly opposite in the way they are understood and/or answered. This is especially true for the 'right to a job' item. However, the choice for one phrasing or the other seems arbitrary. This particular way of phrasing a question is therefore inadvisable, because a decision that seems arbitrary is not arbitrary in its consequences. The key problem in this case may be the complex sentence structure in which men are compared to women, given an attribute (right to a job) under a certain condition (when jobs are scarce), and then a 'degree of agreement' with a norm ('should have') is asked. A more accurate way of measurement that may be less sensitive to such arbitrary shifts in response behavior might be to ask questions about the rights men and women should have according to the respondent directly.

The thresholds provide some insight into the nature of differences in categorization. However, the quality of the measure in the continuous model depends also on parameters of the categorical response model such as the method effects and the error variances, and on the latent response variable distribution.

Besides the thresholds also the correlations between the LRVs are estimated. Based on these correlations the MTMM model mentioned before has been estimated and so estimates of the quality and method effects of the measures corrected for categorization are estimated for all questions. The quality and method effects of the CV model have also been estimated. The results are presented in table 1.4. Based on these results the categorization effect can be derived because it is the ratio of the two coefficients. This result, too, is presented in table 1.4.

The top two rows of table 1.4 show that the quality in Greece was higher than in Slovenia using the CV model; this is, indeed, the reason we chose these particular countries to compare. The quality in Slovenia is lower for the first question, dramatically lower for the second question, and very similar for the third question. This is in principle in line with the descriptions given above of our expectations of categorization errors.

However, table 1.4 also shows that such interpretations of the possible influence of the thresholds are not as straightforward as they might seem. We fitted the MTMM model to

the estimated covariance matrix of the latent response variables, and obtained a model that fit reasonably well ($\chi^2 = 20$, $df = 10$, $p = 0.02$). While for the first and second questions the low qualities are indeed corrected upwards somewhat after the categorization has been taken into account, the opposite happens in Greece. In that country all quality coefficients are lower in the categorical analysis than they are in the continuous analysis.

A consequence of this is that, using the CV model, a higher quality is obtained in Greece than in Slovenia, while the reverse is true in the categorical model for the first and last items. This is rather striking given that, taken over all questions in the main questionnaire, Greece had a substantially higher quality estimate than Slovenia (see table 1.2).

The analyses of the other three experiments show that sometimes no large differences between the countries are found, while in others the thresholds are rather different. In particular we found several cases where the same question did not cover the distribution of the opinion in one country, but provided more information in another. We also found both examples of cases where differences in the quality do not go together with differences in the thresholds, and examples of cases where they do. A more detailed discussion of the results for the other three experiments can be found in the appendix.

Now that we have presented and discussed the results of one experiment in detail, the question remains whether there is a connection between the categorization factor and the quality of the question. The next section therefore presents the results of a meta-analysis we conducted on the categorization factors.

1.5.2 A meta-analysis of the results

Does the categorization factor affect the quality? Using the results presented in the previous sections, we constructed a data set consisting of the categorization factor for all questions–including those from the supplementary questionnaire not shown above–in the four different experiments for which this index was available. This yielded 72 cases in total.

As shown before, the categorization factor equals $c = q^2_{con}/q^2_{cat}$, and so $q^2_{con} = q^2_{cat}(c)$. If there were no effect of the categorization, then there would be no relationship between c and q^2_{con}, since q^2_{cat} would be higher or lower by a constant factor. If c and q^2_{con} are plotted against one another, one would then expect to find the points randomly distributed along a horizontal line. Figure 1.5 shows the scatter plot of these two quantities. Estimates from different experiments have been indicated with different symbols.

The clear relationship shown in the figure indicates that high quality coefficients from the continuous model tend to be lower in the categorical model, and vice versa. Figure 1.5 shows that categorization factors above unity were mostly found for questions with a high quality. We can estimate the relationship between the quality from the continuous model for each experiment easily by the transformation $\ln(q^2_{con}) = \alpha_k + \beta_k \ln(c)$. Here k indexes the four different experiments. We then fit a linear regression to the transformed variables. The resulting predictions for each experiment are shown in figure 1.6 on the original scales.

Figure 1.6 shows that both the intercepts and slopes for the 'efficacy' and 'job' experiments are rather similar, while the coefficients for the 'role of women' and 'social distance' experiments are completely different. The effect of the categorization factor is strongest in the 'social distance' experiment, where also some large differences between the threshold distances were found (see appendix). The experiment with the smaller number of categories, 'job', does not have a high coefficient.

We now turn to the question if these factors also differ between countries with 'high'

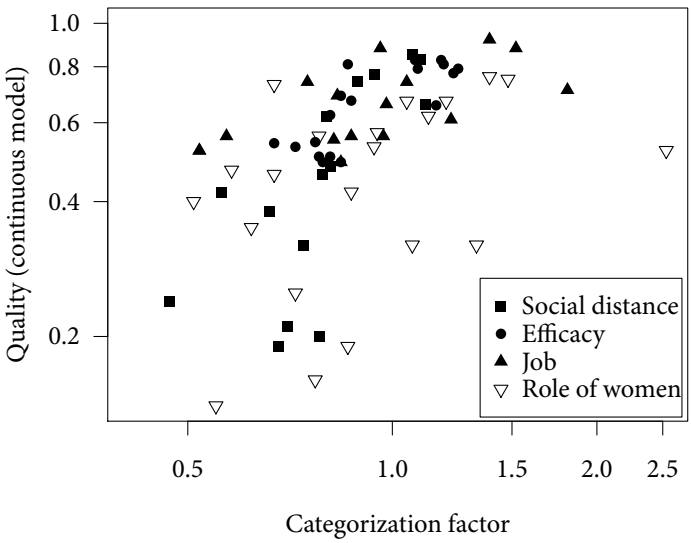


Figure 1.5: Scatterplot of the categorization factor (c) and the total quality of a measure (q^2_{con}) across the experiments. Note the log-log scales.

and 'low' quality coefficients. If the sample is split according to whether the quality was 'high' or 'low', the means of the categorization factors of the two groups are 1.25 and 0.85, respectively, for the questions in the main questionnaire ($t = 3.7$, $df \approx 18$, $p = 0.002$. For the questions in the supplementary questionnaire, the difference is in the opposite direction, but not statistically significant ($t = -1.70$, $df = 28$, $p = 0.10$). This suggests there is a considerable effect of the categorization, at least in the main questionnaire.

One possible explanation for the interaction effect found is that method factors were often constrained to zero for the main questionnaire. The questions in the main questionnaire were selected exactly because they were expected to have high quality and low method effects. The initial continuous analysis often indicated that the questions in the main questionnaire indeed had zero method variance. Since the categorical model tends to increase correlations, if the monomethod correlations for the main questionnaire go up more than the other correlations, it can happen that in the categorical model a method factor is found where none was found before. This will then lower the quality estimates.

A test was done of the hypothesis that questions for which the method effect was constrained to zero in the continuous model have the same categorization factor as other questions, controlling for country effects. This hypothesis was rejected ($p = 0.02$)[10]. The explanation that constraining the method factors to zero causes the interaction found above therefore seems plausible.

[10]Result of a hierarchical linear model fit using `R 2.6.1` with fixed effects of country and restricting the method to zero or not (0/1), and a random intercept across topics to account for the dependency among the observations.

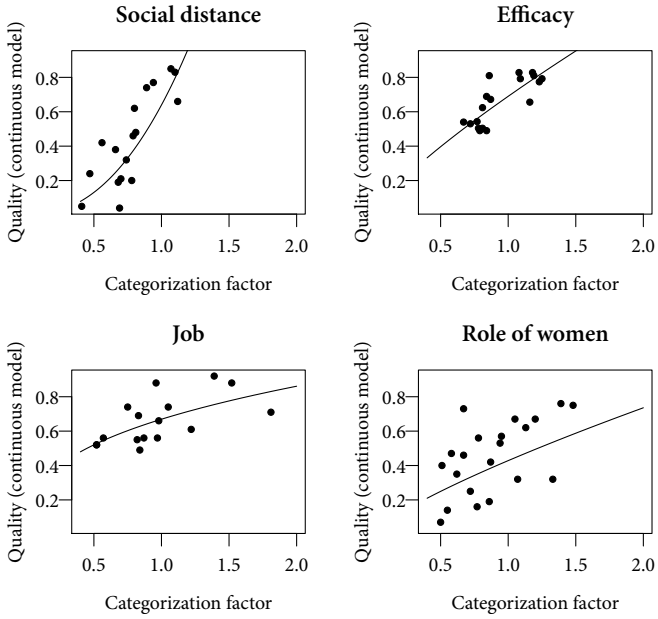


Figure 1.6: Scatterplot of the categorization factor (c) and the total quality of a measure (q^2_{con}) by experiment. The prediction line of the model $\ln(q^2) = \alpha + \beta \ln(c)$, as estimated for each experiment separately, is also given.. For the numerical estimates of these coefficients, please see the appendix.

1.6 Discussion and conclusion

Using the multitrait-multimethod design and model in the ESS, we found large differences between countries in the quality of survey questions. Because such differences can have important implications for cross-country research and survey design, we set out to discover whether these differences could not be attributable to errors due to the use of a small number of categories.

Overall, we found that categorization errors do occur besides random errors and method effects. These errors have two types of effects on the quality of the questions, which can work in opposing directions. The first is that the quality is lower when there is more categorization error. The second, that the categorization attenuates the relationships between different variables in the model differently, affecting not only the quality, but also the method effects and other parameters of the model. This in turn has as its consequence that the quality parameter under the CV model is not always smaller than the quality under the categorical model, as evidenced by the many 'categorization factors' above unity which we found.

A caveat should be added to the interpretation of this result, because a violation of the assumptions of the models (no categorization error versus bivariate normality) can have different consequences for the estimates. It is therefore not necessarily true that a categorization factor above unity indicates overestimation of the quality in the CV model. Several studies of the robustness of factor analysis models to categorization errors exist (see Olsson, 1979). However, we found that their results do not necessarily apply in the MTMM

model, which also includes method factors. Given the ubiquity of correlated errors in survey questions, it would be useful to study more closely the robustness of this particular type of measurement error model to categorization error. This, however, is beyond the scope of the present paper.

In a meta-analysis, we gathered the results from our four different experiments and analysed the relationship between the categorization factor and the quality in the continuous model. Effects were found for all four experiments.

If the categorization factors were equal for countries with the highest and lowest quality coefficients, they could not explain the differences in quality which we found earlier. The meta-analysis suggested that there is a considerable difference in the categorization factor between countries where the highest and the lowest quality coefficients were found given whether the question was part of the main or supplementary questionnaire.

The methods in the main questionnaire were chosen beforehand based on other experiments as the ones least likely to cause method effects. For example, direct questions rather than batteries were used. After re-examining the experiments on which the meta-analysis was based, it appears this is closely related to the interaction effect found there.

The main reason for the interaction effect we found in the meta-analysis appears to be that the method variance for the main questionnaire method was often close zero. The general rise in correlations that results from correction for categorization seems to have 'pushed' the monomethod correlations of the main questionnaire variable to the point where the method variance could not anymore be constrained to zero. And as the method variance rises, the quality must decrease in our model.

In other words, the correction for categorization has a negative influence on the quality. When the method factors were constrained to zero in the first instance, the effect was that the quality was in general lower in the categorical model than in the continuous model. This is contrary to what one might expect considering that all of the polychoric correlations are higher than their Pearson counterparts.

In this study we have shown that it is possible to split the measurement error model into three parts:

- A part due to random errors;
- A part due to systematic errors;
- A part due to splitting the variable into just a few categories: 'categorization error'.

This study has been largely descriptive of the effects of categorization error. Given our findings, it seems important to better judge the relative merits of the continuous and categorical models, and the effects that different question characteristics have, not only on quality and method effects, but also on the categorization errors.

Our study also has some limitations due to the assumptions made to attain the above separation. These are: normality of the latent response variables, linearity of the relationship between the latent traits and latent response variables, and interval measurement of the latent traits. In another paper these issues will be addressed by examining ways to relax the assumptions. Future research might also focus on finding other explanations for differences in quality across countries.

Chapter 2

Latent Class Multitrait-Multimethod Models

Abstract

The present paper suggests a statistical method, the latent class MTMM model, of estimating the quality of single questions while making fewer assumptions than have been made so far in such evaluations. The method is a combination of the multitrait-multimethod research design of Campbell and Fiske (1959), the basic response model for single questions of Saris and Andrews (1991), and the latent class factor model of Vermunt and Magidson (2004a, 227–230). The latent class MTMM model is thus not novel in itself, but combines an existing design, model, and method to improve the analysis of single questions in survey research.

A real experiment from the European Social Survey (ESS) is analyzed and the results are discussed at length, yielding valuable insights into the functioning of these questions.

Introduction

Since the late 19th century, psychometricians have studied the measurement quality of scales. With the advent of item response theory (IRT), the focus has shifted somewhat from scales *per se* to the quality of *indicators* as measurements of the scale (Hambleton et al., 1995). An IRT analysis of items provides more information about the functioning of the different indicators of the scale, separate from the properties of the scale as a predictor of behavior.

However, in some cases or disciplines, only one indicator may be available, an indicator may be used for different scales, or different countries must be compared with each other. Furthermore, a scientific interest exists among survey researchers in the effects of different design choices on the question quality, separate from the scaling properties of an indicator. In these cases, we argue, it is important to study the quality of *single questions* as a measurement of the indicator: the focus should then be shifted from indicators to single questions.

The present paper suggests a statistical method of estimating the quality of single questions as measurements of an indicator, while making fewer assumptions than have been made so far in the evaluation of single questions. The method is a combination of the multitrait-multimethod research design of Campbell and Fiske (1959), the basic response model for single questions of Saris and Andrews (1991), and the latent class factor model of Vermunt and Magidson (2004a, 227–230), originally formulated by Lazarsfeld and Henry (1968). The latent class MTMM model is thus not novel in itself, but combines an existing design, model, and method to improve the analysis of single questions in survey research.

The data obtained from multitrait-multimethod experiments (Campbell & Fiske, 1959) allow for a separation of systematic errors due to the method of asking a question and random measurement errors from the indicator of interest (Schmitt & Stults, 1986). By applying the latent class factor model, we obtain very precise information about the way responses are generated from underlying opinions on single indicators.

We discuss the method by applying it to a real dataset from a multitrait-multimethod experiment done in the European Social Survey (ESS). In an earlier study this experiment and several others were analyzed using the commonly applied confirmatory factor analysis and ordinal probit models (Oberski et al., 2007). The assumption of normally distributed latent response variables made in those analyses – that is, of an ordinal probit relationship between trait and indicator with parallel cumulative probability curves – may be false. The present application shows how this assumption can be relaxed. The latent class approach has the advantage that many assumptions that are usually made can be investigated. Among them are the measurement level (nominal, ordinal, or interval) of the observed variables, and the distribution of the latent variables. The model does not require the assumption of normally distributed latent variables, since the marginal distribution of the latent variables is left to be estimated.

The next section argues that it is essential to estimate the quality of single questions. We then explain the experimental design and the response model applied to analyze this quality. The rest of the paper applies this model to a real dataset, presented in the subsequent section. We then briefly note the software and methods used, after which the results of the analysis are discussed. Finally, conclusions are drawn from the analysis, showing the added value of our approach.

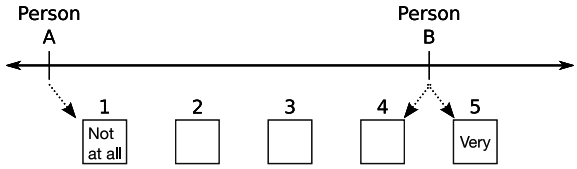


Figure 2.1: Theoretical true score and response options for the question 'How happy are you?'. The choices to be made when going from true score to response options are not always obvious.

2.1 Measurement error in single questions

Answers to survey questions cannot be taken for granted. There are random and systematic components in the answers given by respondents that have nothing to do with the opinion the question was supposed to measure. Such components are therefore measurement errors.

Systematic components arise because different people have their own idiosyncratic way of answering questions given their opinion (Saris, 1988). Some give extreme answers on five-point scales while others tend to choose the middle point, for instance (Hui & Triandis, 1989). Some are more sensitive to social desirability than others, causing differences depending on how the question is phrased (Crowne & Marlowe, 1960). One may also say that respondents 'satisfice', using simplifying answering strategies to reduce cognitive burden (Krosnick, 1991). These processes are distinct but have in common that they may cause two people with the same underlying opinion to give different answers, and will cause two answers to unrelated questions answered by the same person to correlate.

Such systematic ways of answering the question vary across people, but may be stable across questions. They therefore cause both error variance and spurious relationships between answers to questions asked in the same way. If the way of answering a question is specific to both person and method, it is called a 'method effect'. An example would always choosing to agree or disagree 'strongly' on agree-disagree scales, but not on other kinds of scales: this would be extreme response behaviour specific to the method. If the same respondent has a tendency to choose the extreme categories for *any* type of answer scale, the systematic error is called a 'style factor' (Jackson & Messick, 1958). Crucially, neither method effects nor style factors are related to the question content.

Random error is another source of measurement error; after the respondents have moulded their opinion into the form required by the question, some element of arbitrariness in choosing a response option may still remain. Consider the lines in figure 2.1. The possible opinions after correction for systematic effects or 'true scores' (Lord & Novick, 1968) of the respondent are represented as a line, while the response options below are categorical. Person A would presumably have no difficulty choosing 'not at all'. However he or she may make a mistake and accidentally mark option 2 rather than option 1. Person B, at the same time, could equally well choose options 4 or 5, and might do so at random from occasion to occasion. Both processes may occur at the same time and give rise to random measurement error.

This suggests that answers to survey questions contain random and systematic measurement errors. Estimating such errors (1) assesses the general quality of a question; (2) allows for the correction of study quantities of interest such as regression coefficients or group differences for the influence of errors; (3) assesses the cross-group comparability of

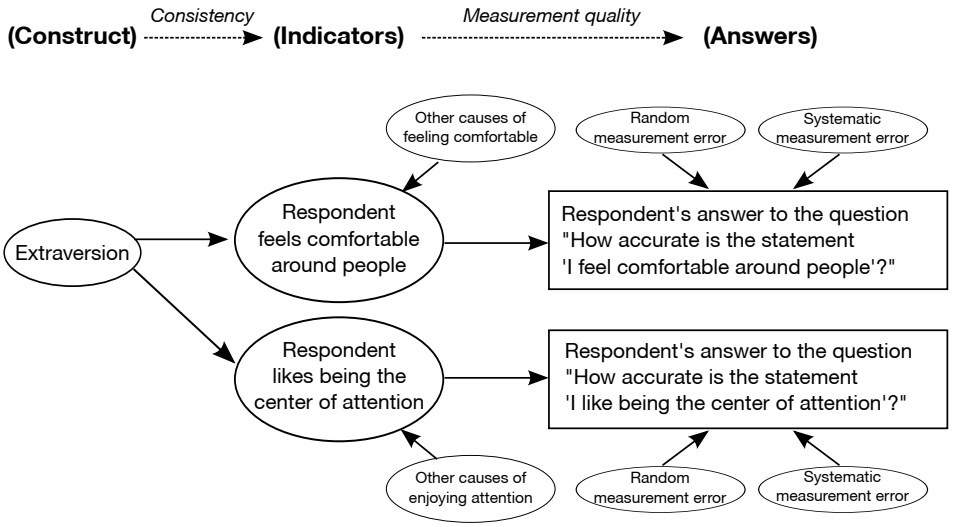


Figure 2.2: Illustration of the difference between pure measurement quality (the relationship between observed answer and unobserved indicator) and the consistency of indicators (the relationship between the unobserved indicator and the unobserved construct). In the present paper we will only study the connection between indicator and observed answer: pure measurement errors. The indicators are taken from the International Personality Item Pool (http://ipip.ori.org/)

.

quantities of interest.

The question has been asked for the purpose of measuring a construct. We term the degree to which the indicator, after correction for pure measurement error, measures this construct the 'consistency' of the indicator (Saris & Gallhofer, 2007a). The combination of measurement error and consistency has been called 'construct validity' by Andrews (1984). An illustration of the distinction between measurement errors and consistency is given in figure 2.2.

Assessing the general quality of items that form a scale and their cross-group comparability is a fairly common activity in psychological research[1]. This quality concerns both the degree to which an indicator is influenced by a construct ('consistency') and the pure measurement errors discussed above. There are, however, advantages to estimating the pure measurement errors separately rather than this combination.

First, there is a scientific interest among survey researchers in the effect on the quality of the questions of various choices to do with survey design. Such choices could refer to the number of response options, use of an agree-disagree scale, linguistic complexity, etc. (Saris & Gallhofer, 2007b). They can also refer to nonresponse (Olson & Kennedy, 2006), or the study of special populations such as immigrants or elderly people (Groves, 2002). In order to separate this effect on quality of survey design from effects on consistency with the construct, it is necessary to estimate measurement error separately.

[1]Dividing the number of matches in Google Scholar (http://scholar.google.com/) to each of the APA's 'core of psychology' four largest impact factor journals (including Psychological Methods) by the number of matches adding the term 'differential item functioning' suggests DIF is mentioned an average of 6%. If the percentages are weighted by the journal's impact factor in 2007, the average is about 4%. Although DIF is not mentioned very often, it is clearly a well-known technique.

Second, in studies that compare groups such as cross-national research, the measures must be invariant across groups: only measures with equal consistency across groups allow for comparisons. Having only the combination of measurement and consistency error available results in the stricter requirement that both must be equal across countries. Saris and Gallhofer (2007a) argued that such tests are unnecessarily strict and only the higher-order relationship between construct and indicators need to be invariant. Such a test requires separation of measurement error from consistency.

The third reason for estimating pure measurement errors is that in the social sciences, there are few standardized scales. Consequently, questionnaires often contain only one question instead of a number of questions to measure a single construct. A classic example in sociology is the question used to measure social trust: "Would you say that most people can be trusted, or that you can't be too careful in dealing with people?"[2]. Furthermore, it may also happen that different researchers construct different scales *post-hoc* using the same questions. Examples in political science are questions on citizens' trust in various political institutions. In such cases, estimating the extent of measurement errors in the question allows for the correction of the attenuation of relationships with other variables due to errors, and provides an upper bound for construct validity.

For these three reasons it is essential to estimate the quality of single questions. Two general approaches are possible: longitudinal designs using quasi-simplex models (Alwin, 2007) and 'multitrait-multimethod' (MTMM) experiments (Campbell & Fiske, 1959). We will discuss an approach of estimating the quality of single questions based on MTMM experiments that requires fewer assumptions than the approaches used so far for such data. An approach similar to quasi-simplex models for longitudinal designs such as the ones discussed in Alwin (2007) was discussed by Biemer and Bushery (2000)[3].

2.2 Multitrait-multimethod experiments

Campbell and Fiske (1959) suggested measuring multiple indicators ('traits') by multiple methods (MTMM). The correlations thus obtained were posited to follow a certain pattern. Later, different models were proposed to analyze these patterns, of which confirmatory factor analysis is the most commonly used (for a review, see Schmitt and Stults (1986)).

What the models applied to MTMM data have show is that the MTMM design can be used to separate the relationship between the indicator to be measured and the observed variable from random and systematic measurement errors. Note that here we mean by traits the indicators in the sense of figure 2.2 rather than the construct.

The classical MTMM approach recommends the use of a minimum of three traits that are measured with three different methods leading to nine different observed variables. An example of one trait measured with three different methods is given in figure 2.3.

Collecting data using this MTMM design, data for nine variables are obtained. These variables become the subject of a measurement or MTMM model. There is an ample literature about MTMM models using confirmatory factor analysis and the different choices

[2]The question was devised by Noelle-Neumann in 1948 in Germany. Later Rosenberg (1956) created a multiple item concept (scale) using this question. But to date, many questionnaires only contain the single question.

[3]It should be noted that the notion of reliability estimated by longitudinal models is different from that employed here: in the longitudinal studies mentioned unique considerations of the moment that form part of the true variance are included as measurement error (Veld, 2006).

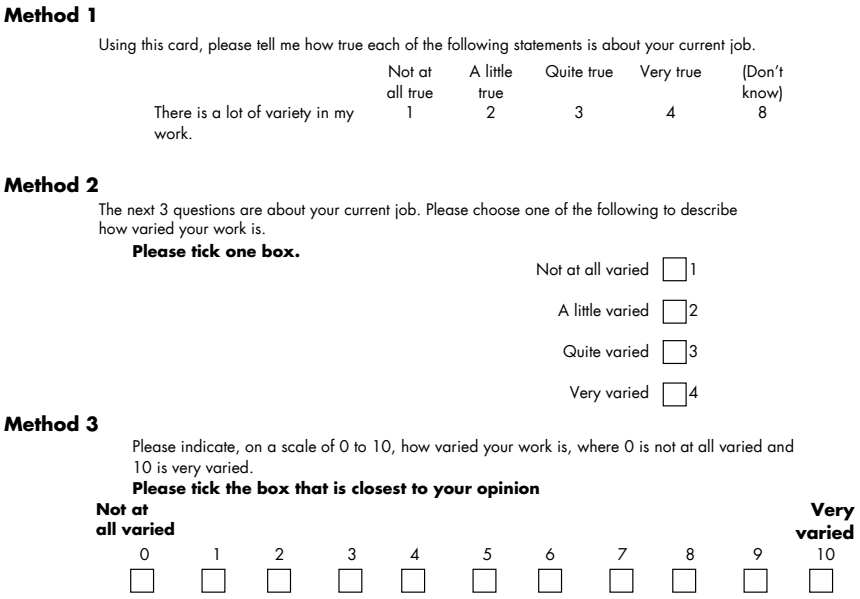


Figure 2.3: The trait 'perception of variety of job' measured by three different methods.

that can be made for such models. Here we wish to start from a more general model formulation that specifies the relationships between the latent and observed variables without necessarily being a confirmatory factor analysis.

2.3 The response model

Figure 2.4 specifies the relationships between the observed scores and their general factors of interest as a graph. The directed arrows indicate second-order effects, while the double-headed arrows indicate second-order relationships. The absence of an arrow implies conditional independence. We assume that in this graph no higher-order interactions exist. One response model for MTMM data that does contain interactions between the traits and methods was suggested by Browne (1984). However, when Corten et al. (2002) compared the fit of different models to data from many different MTMM experiments, no evidence of interactions was found. Thus there are empirical indications that the assumption of no higher-order interactions in these types of MTMM experiments is reasonable.

This figure shows that each trait (T_i) is measured in three ways. It is assumed that the traits are dependent but that the method factors (M_1, M_2, M_3) are independent.

In figure 2.4, y_{11} through y_{33} are the observed variables belonging to the experiment. The first digit (i) corresponds to the trait number and the second (j) to the method number. Following the graph, each trait is indicated with T_i and each method with M_j, In total there are $I = 3$ traits and $J = 3$ methods.

The quality of a measure is the strength of the relationship between the trait and the indicator that is supposed to measure it. The amount of systematic error or method effect depends on the strength of the relationship between the method factor and the indicators measured using that method. It should be noted here that a drawback of the MTMM

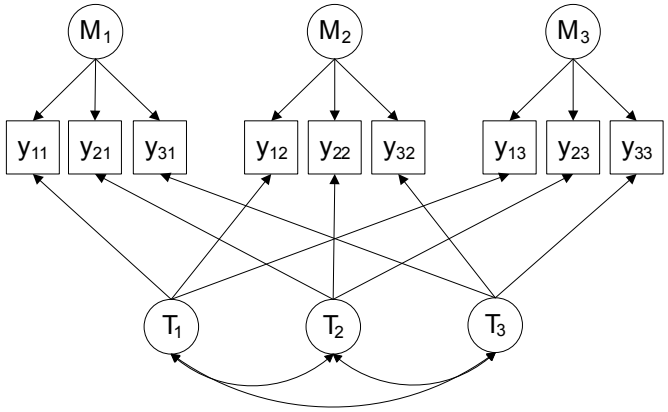


Figure 2.4: A model graph for multitrait-multimethod data. The method factors (M) represents different answering strategies used by the respondents that may be similar across questions. The trait factors (T) represent the opinion of the respondent after correcting for idiosyncratic response sets and random measurement error. Random error components for each observed variable are not shown here for clarity but can be imagined.

design we use is that one cannot separate method effects from other systematic errors. Thus an assumption is made that all systematic errors are specific to the method used. In an investigation of different explanations for correlated errors in MTMM data, Corten et al. (2002), provided some evidence that this assumption is reasonable.

The most common model applied to this graph is the continuous confirmatory factor analysis (CFA) model. In that case one can define the quality as the amount of variance explained in the indicator by its trait (Saris & Gallhofer, 2007a).

However, the assumption of continuous and interval measurement implicit in the CFA model may be false when responses with only a few categories are obtained. In that case the ordinal CFA (oCFA) model of Muthén (1984) is often applied (Scherpenzeel & Saris, 1997). This model is equivalent to Samejima's graded response model (Samejima, 1969) in item response theory. Such an analysis can be accomplished by applying the CFA model to so-called polychoric correlations, or by special software.

The oCFA model takes the discrete and ordinal nature of the responses into account, but at the cost of strong assumptions about the specific form of the relationship between latent and observed variable. In particular, it is assumed that there are continuous latent response variables that have been split up into just a few categories. These latent response variables are assumed to have a normal or logistic distribution, leading to the familiar probit or logit relationship between trait and observed variable (and between method and observed variable). More importantly, the slope parameters of the influence of the trait on the indicator are restricted to be equal for all categories. This implies that the cumulative probabilities of all categories are restricted to be parallel S-shaped curves.

It should be noted that although the normal distribution is a commonly used and computationally convenient choice for the latent response variables, other choices have also been suggested in the literature. Skew-normal (Roscino & Pollice, 2006), copula (Joe, 2005), and mixtures of normal distributions (Uebersax & Grove, 1993) have been suggested. Rost and Walter (2006) applied mixture Rasch models and the LLTM to MTMM

data. The alternative approach of optimal scaling should also be mentioned (Takane et al., 1977). These approaches do relax the assumption of normality, but express all relationships only in terms of the latent response variables, which does not allow for a full analysis of the relationship between the traits and observed variables we are interested in.

In this paper we will elaborate on a different approach: the latent class factor model (Vermunt and Magidson (2004a, 227–230); see also Vermunt and Magidson (2004b, 185)). This model (the LCM) derives from the latent structure model formulated by Lazarsfeld and Henry (1968). Goodman (1974) further developed the latent structure model and gave a method for maximum likelihood estimation of the parameters. Haberman (1979) discussed the parameterization in terms of log-linear coefficients used here, and suggested explicitly that different restrictions on these coefficients yield models for different measurement levels of the observed variables. Different applications of other variants of these models are discussed by Hagenaars and McCutcheon (2002).

The LCM specifies the following relationship between trait, method, and observed variable:

$$p(y_{ij} = k|T_i, M_j) := \frac{\exp(a_{ijk} + b_{ik}^{(t)} T_i + b_{jk}^{(m)} M_j)}{\sum_{l=1}^{K} \exp(a_{ijl} + b_{il}^{(t)} T_i + b_{jl}^{(m)} M_j)}; k, l \in \{1, ..., K\}, \qquad (2.1)$$

where K is the number of categories for the observed variable. The latent variables (traits and methods) are scaled to have equal-distance values lying between 0 and 1. Thus a trait with 5 categories will have scores $\{0, 0.25, 0.50, 0.75, 1\}$ for category numbers 1 through 5. The log-linear parameters a and b are set to sum to zero over all categories of y. This is an arbitrary restriction necessary for identification. By $b_{ik}^{(t)}$ we mean the slope for trait number i and category k, whereas the $b^{(m)}$ are slopes for the method.

The b_k parameters in the model of equation 2.1 are the associations or log-linear effects and the a_k parameters are intercepts for each category. The effects b_k differ by category. Each effect can also be written as $b_k = bk$, meaning there is only one slope b for the effect of the latent variable on the observed one, and k is the category score. The category scores can be restricted to increase by a certain number for each category, or they can be freely estimated. By different restrictions a different assumption about the measurement level of the observed variable results.

If the observed variable is assumed to be have interval level, the category scores can be assumed to be of equal distance, in general increasing by unity. A common choice is to use the scores 1, 2, 3, 4, 5 for the first, second, third, etc. categories. The effects b_k in equation 2.1 then become $b, 2b, 3b, 4b, 5b$ for each category k. This model is also known as the linear by linear association model, since the local odds ratios of adjacent rows and columns in the cross-table of the latent and observed variables have the same value everywhere, namely $\exp(b)$ (Agresti, 2002, pp. 369-370). In item response theory this formulation is equivalent to the partial credit model (Thissen & Steinberg, 1986).

The category scores k can also be estimated by the model. In this case a linear relationship between trait and the log-odds of choosing a given category is still assumed, but the scores of the observed variables' categories can take on any value. This includes values that do not increase or decrease monotonically with the category number. The observed variables would then have nominal measurement level. In practice often the scores do increase monotonically with the category number, yielding an ordinal measurement level. This ordinality is, however, not a restriction imposed by the model but may or may not be found

	Latent variable	
	Nominal	Interval
Observed variable		
Nominal	(*Classical LCA*)	Row/column association model
Ordinal	(*Classical LCA with constraints*)	(*RC association with constraints*)
Interval	(*Row/column association model*)	Uniform association model

Table 2.1: Different measurement levels for latent and observed variables can be accomodated within the latent class model (Heinen, 1996). We consider only the models where the latent variable is interval and the observed variable either nominal or interval. Other possible models are indicated in brackets.

in practice. Since the effects b_k equal bk, one cannot determine whether the differences in b_k stem from different categories or different slopes (or both). Thus for this model we will only report their combination b_k.

The model where the observed variables have nominal or ordinal measurement level is also known as the row (or column) association model (Hagenaars, 2002, 243), after the set of loglinear models for observed variables of the same name (Agresti, 2002, pp. 373-4). It can also be described as a latent class version of the nominal response model from item response theory (Bock, 1972).

The LCM can thus accommodate different measurement levels of the latent and observed variables (see table 2.1; see Heinen (1996) for further discussion). Both latent and observed variables are always regarded as discrete, but one can impose the restriction of interval measurement on latent and/or observed variables. We use this possibility to examine whether the responses can be taken to have been measured at interval level or not, and whether the assumption of ordinal categories is warranted. In this table 'classical LCA' indicates the model where no restrictions are placed on the pairwise loglinear parameters.

Again the quality and amount of systematic error of the observed variable can be defined in terms of the relationship between the trait and method variables and the observed variable. Where in the CFA model the quality is the amount of explained variance, in the LCM the relationship is more complex and depends on the value of the latent trait. It is determined by the log-linear a and b parameters of equation 2.1, which can be used to express the effect of the trait or method on the observed variable in odds ratios. We can say that the quality of the measure is zero for all values of the trait if the relative odds of choosing any category do not increase or decrease with the trait. This is the case only when all b coefficients are zero. In contrast, a good measure has a high quality for values of the trait that cover as much of its distribution in the studied population as possible. Note that typically very high and very low (and unlikely) values of the trait will still be inaccurately measured, even by a high-quality measure.

The odds ratios aid in understanding the model, but make it more difficult to interpret in terms of probabilities. We will therefore also examine the probability of each category given the trait (item category characteristic curves). We will further evaluate the quality of the questions by plotting the amount of information that each item provides on its own about its latent trait. The so-called 'item information' is a generalization of the concept of reliability and used in test construction in IRT (Hambleton et al., 1995). In sum, although the relationship between latent variable and indicator is more complex than in the linear CFA model, this relationship can still be examined, and in great detail. In general the LCM

provides much more detailed information about the use of the scale than the other models mentioned above.

The LCM approach also has disadvantages. First, in our previous conceptualization, the latent variables were continuous and measurement errors arise partly because answers are obtained only in categories. To put it another way, only an unknown range of values on an underlying continuous variable is observed, and the latent response variables are discretized into the observed variables. In the latent class model this aspect of the errors is not modeled: the latent variables are still discrete. Thus, if the real variable of interest is continuous the latent classes still contain measurement error. Whether the classes of the LCM provide an accurate enough approximation to this continuous distribution is a topic for discussion and research.

The LCM approach may also appear to have the disadvantage that the models have many parameters. However, models with linear or ordinal parametric relationships *can* be formulated in this framework using testable restrictions, and are thus special cases of the LCM. Therefore, if a complex model is necessary, the LCM can be used to estimate it. And if a parsimonious model fits the data just as well, such a model can be found using the same approach as well. Here we will use the Bayesian Information Criterion (BIC), which penalizes extra parameters, to investigate whether the more complex or the simpler model is necessary.

2.4 Data

The European Social Survey (ESS) has the unique characteristic that in more than 20 countries the same questions were asked and that within each round of the ESS Multitrait-Multimethod (MTMM) experiments are built in to evaluate the quality of a limited number of questions. This gives us an exceptional opportunity to observe the differences in quality of questions over a large number of countries. In this paper we have used the MTMM experiments of round 2 of the ESS. The topics of the 6 MTMM experiments in the second round of the ESS were (1) Time spent on housework; (2) The social distance between the doctor and patients; (3) Opinions about job; (4) The role of men and women in society; (5) Satisfaction with the political situation; (6) Political trust.

Concerning each of these topics 3 questions were asked and these three questions were presented in 3 different forms following the discussed MTMM designs (Campbell & Fiske, 1959). The first form, used for all respondents, was presented in the main questionnaire. The two alternative forms were presented in a supplementary questionnaire which was completed after the main questionnaire. All respondents were only asked to reply to one alternative form but different groups got different version of the same questions (Saris, Satorra, & Coenders, 2004). For the specific questions for the 6 experiments we refer to the ESS website where the English source version of all questions are presented[4], and for the different translations we refer to the ESS archive[5].

Each experiment varies a different aspect of the method by which questions can be asked in questionnaires. The 'housework' experiment compares numeric estimates by respondents with other scales. The 'doctors' experiment examines the effect of choosing arbitrary scale positions as a starting point for agreement-disagreement with a statement.

[4]http://www.europeansocialsurvey.org
[5]http://ess.nsd.uib.no

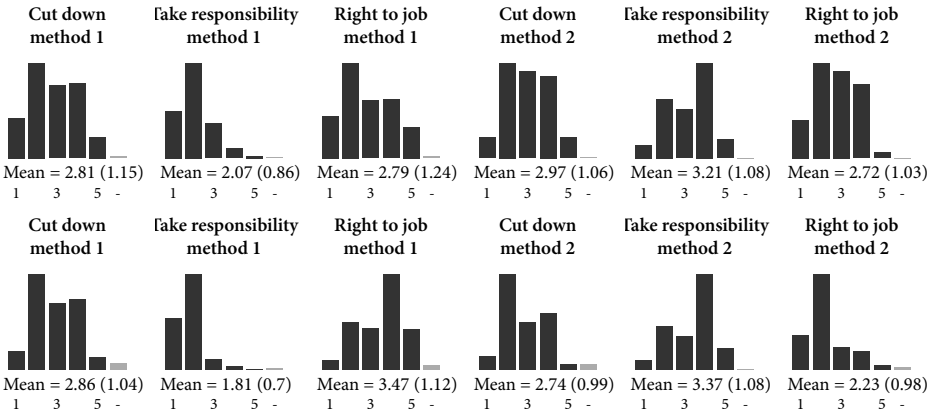


Figure 2.5: Top: Greece. Bottom: Slovenia. Barplots, including proportion of missing data (-) of the observed variables for the first two methods of the 'role of women' experiment in two countries. Below each barplot the mean and standard deviation (in brackets) are given.

The 'job' experiment compares a 4 point with an 11 point scale and a true-false scale with a direct question. In the 'women' experiment agree-disagree scales are reversed, there is one negative item, and a 'don't know' category is omitted in one of the methods. The 'satisfaction' experiment varies the extremeness and number of fixed reference points of the scale. And finally, the experiment on political trust was meant to investigate the effect of repeating the same question in the same format.

The questions asked in the different countries have been translated from the English source questionnaire. An optimal effort has been made to make these questions as equivalent as possible and to avoid errors. In order to reach this goal two translators independently translated the source questionnaire and a third person was involved to choose the optimal translation by consensus if differences were found. For details of this procedure we refer to the work of (Harkness et al., 2002).

We applied the LCM model specified above to the experiment on the role of men and women in society. Three traits were measured in this experiment, namely:

1. "A woman should be prepared to cut down on her paid work for the sake of her family;"

2. "Men should take as much responsibility as women for the home and children;"

3. "When jobs are scarce, men should have more right to a job than women."

These traits were measured using a five category agree-disagree scale in different phrasings of the question for the first two methods, and using an item-specific scale as the third method (see appendix for question formulations). Barplots and descriptive statistics for the variables we will study are given in figure 2.5.

In order to be able to compare countries on the quality of measurement, we selected the country with the highest and the country with the lowest quality for the questions, as estimated in the confirmatory factor analysis model. In this experiment Greece (n=2406) had the highest qualities and Slovenia (n=1442) the lowest. More information can be found together with the precise quality estimates for all countries in Oberski et al. (2007).

2.5 Methods

We used the program Latent Gold 4.5[6] (Vermunt & Magidson, 2005a) to estimate the following models which result from different assumptions about the relationships in figure 2.4:

Observed variable measurement level							
Interval				Nominal			
	Traits (no. classes)				Traits (no. classes)		
Methods	3	4	5	Methods	3	4	5
2	×	×	×	2	×	×	×
3	×	×	×	3	×	×	×

In all models we take the latent variables to be of interval level measurement, while the observed variables may be interval or nominal. We also investigate models with different numbers of classes, to the extent allowed by the amount of information in the data and the estimation procedure[7]. In order to limit the number of possible models, we vary the number of classes for all traits at the same time and for all methods at the same time. We do not consider models with 5 classes for one trait and 3 for another trait, for instance.

No restrictions are imposed on the associations between the latent traits, except that there are no third-order interactions. This implies that the associations between any two latent traits may be of any form but do not vary across levels of the third trait.

Although not shown here, we also estimated models with no method factors and differing numbers of classes for the traits. In all cases the fit indices indicated a strong need to introduce method factors.

In the analysis, aside from dealing with the planned missing data design that can be considered missing completely at random (MCAR), we also take into account the design weights provided by the ESS, interviewer clustering effects on estimates and standard errors, and data missing at random (MAR). The solutions are obtained by the EM algorithm with at least 10 random starting values in order to find the global optimum, switching to Newton-Raphson at the end of optimization.

2.6 Results and discussion

Model selection

We estimated the latent class MTMM model described above with different numbers of classes and different assumptions about the measurement level of the observed variables. That is, a so-called linear-by-linear association model and a row association model. Table 2.2 shows the resulting BIC model selection criteria and selected models[8]. Lower numbers indicate a better and more parsimonious fit to the data. Note that models with differing number of classes are not nested and cannot be compared using a likelihood ratio test. For

[6] http://www.statisticalinnovations.com/products/latentgold_v4.html

[7] For example, estimating the nominal model with 5 trait classes and 3 method classes for Greece took 2.5 days on our computer.

[8] The BIC based on the log-likelihood of the model $-2 \ln L + (\ln N) npar$ is reported. An alternative is the BIC based on the L^2. The two measures always yield the same model choice, however (Vermunt & Magidson, 2005b, 60-1)

Greece

Observed variable measurement level							
Interval				Nominal			
	Traits (no. classes)				Traits (no. classes)		
Methods	3	4	5	Methods	3	4	5
2	31438	31209	31017	2	30922	30595	***30478***
3	31083	30852	30880	3	.	30502	30498

Slovenia

Observed variable measurement level							
Interval				Nominal			
	Traits (no. classes)				Traits (no. classes)		
Methods	3	4	5	Methods	3	4	5
2	17417	17342	17335	2	.	**16149**	*16160*
3	17427	17332	.	3	.	16158	.

Table 2.2: BIC for the different models estimated on the 'role of women' experiment. The model selected by the BIC is shown in **bold face**. The model selected by the AIC (not shown) is shown in *italics*. For Greece BIC and AIC select the same model.

such comparisons the BIC can be used (Raftery, 1995). The selected model according to the AIC criterion is also indicated.

In both countries, the BIC and AIC indicate that models including method factors fit the data much better than models without method factors. Therefore the criteria indicate that method factors must be introduced. Also for both countries, the observed variables cannot be taken to be measured at interval level: a model with nominal (or ordinal) level observed variables fits the data much better. This brings into question the assumption of interval level measurements made by the confirmatory factor analysis model. The degree of the difference between the equal and unequal interval models can be deduced from the parameter estimates discussed later.

In Greece the AIC and BIC select the same model, which has 5 classes for the traits and 2 classes for the method factors. The observed variables are measured at nominal level in this model. The model has 2257 degrees of freedom. In Slovenia the AIC selects this same model (1246 degrees of freedom), while the BIC selects the more parsimonious 4-class solution for the traits (1249 df). In the interest of being able to compare the two countries, we will select the same solution for both countries, choosing the model with 5 classes for the traits and 2 for the methods for both Greece and Slovenia.

Quality of the questions

Parameter estimates The results for the selected model for Greece and Slovenia are shown in table 2.3. This table only shows the parameter estimates for the questions measured using the first method (i.e. the questions asked in the main questionnaire of the ESS).

The model selected has a separate parameter for each category of the observed variable. This parameter can be seen as a varying the effect of the trait on the observed variable (see equation 2.1). The parameters can be interpreted in terms of odds ratios: if the latent trait increases by one category, the odds of choosing category 2 over category 1 of the first item

		Greece				Slovenia			
		Traits b_{kt}	s.e.	Method b_{km}	s.e.	Traits b_{kt}	s.e.	Method b_{km}	s.e.
Cut down									
Agree strongly	1	-21.82	(4.02)	2.76	(1.51)	-6.77	(2.30)	3.33	(0.89)
Agree	2	-17.40	(3.49)	-1.22	(1.39)	-5.85	(1.32)	-1.92	(0.67)
Neither agree nor disagree	3	-5.14	(4.13)	-1.58	(1.14)	0.22	(1.03)	-2.73	(0.66)
Disagree	4	17.78	(2.58)	-2.37	(0.44)	4.53	(1.15)	-2.04	(0.53)
Disagree strongly	5	26.59	(10.36)	2.40	(3.78)	7.87	(3.18)	3.36	(1.34)
Take responsibility									
Agree strongly	1	-29.17	(4.36)	1.37	(0.70)	5.34	(1.87)	0.88	(0.54)
Agree	2	-13.05	(1.99)	-2.24	(0.57)	0.68	(2.08)	-3.94	(0.40)
Neither agree nor disagree	3	12.76	(1.78)	-1.11	(0.43)	-2.64	(1.74)	-3.49	(0.87)
Disagree	4	12.65	(1.94)	-1.06	(0.61)	-6.25	(5.69)	-1.97	(0.59)
Disagree strongly	5	16.80	(4.21)	3.03	(1.18)	2.86	(3.18)	8.52	(0.93)
Right to job									
Agree strongly	1	-25.33	(9.97)	0.06	(2.00)	-18.65	(4.68)	3.40	(0.93)
Agree	2	-20.17	(2.77)	-2.73	(0.46)	-14.11	(2.64)	-0.71	(0.57)
Neither agree nor disagree	3	-7.91	(3.10)	-3.45	(1.13)	1.77	(2.30)	-1.42	(0.33)
Disagree	4	11.05	(4.16)	-2.69	(0.97)	10.94	(2.50)	-2.72	(0.62)
Disagree strongly	5	42.37	(6.33)	8.81	(1.21)	20.05	(2.55)	1.45	(0.82)

Table 2.3: Estimates of the log-linear effects of the traits on their respective observed variables (column 2 for Greece and 6 for Slovenia, with robust standard errors in columns 3 and 7), and of the relationships between the method factors and the observed variables (columns 4 and 8). For the sake of brevity only the three questions from the main questionnaire (method 1) are shown.

in Greece, for instance, increase by 20. This is so because a one category increase of the latent trait is scored as 0.25 and $0.25(e^{-17.4}/e^{-21.8}) \approx 20$. So for each one-category increase in the trait the odds of choosing category two rather than one increase 20-fold.

The model does not restrict the items to be of ordinal measurement level. Ordinality *may* hold, however. An item is ordinal if the estimated log-linear effects (b) of the trait on the observed variable are all increasing (or all decreasing) numbers. The table therefore shows that ordinality holds for all observed variables shown here except for the 'Take responsibility' item in Slovenia, although the difference between the offending coefficients is not statistically significant. This item has an exceptionally low measurement quality in all analyses we have performed.

The same effects are also very unevenly spaced. In some cases, such as categories three and four ('neither agree nor disagree' and 'disagree') for the 'Take responsibility' item in Greece, they are almost equal for two different categories. This suggests that these categories represent much the same opinion and that therefore these items can not be taken to be of interval measurement level.

Turning to the effects of the method factors, it can be seen that these represent an 'extreme versus middle response' factor. Take, for example, the first item in Greece (column three in the table). If a person were to go from class 1 to class 2 on the method factor, their odds of choosing 'agree strongly' rather than 'agree' increases about 50-fold, *keeping the trait score constant*. At the same time, their odds of choosing 'disagree strongly' rather than just 'disagree' increase about 100-fold. Considering that disagreeing with the statement is the obviously socially desirable answer, higher scores of the method factor are associated with answers that are extreme, but more so on the socially desirable side. This finding holds for all items and methods, including the ones not shown here.

The table shows also that the parameters have been estimated with considerable uncertainty. This uncertainty includes sampling design and interviewer effects, since we have included these in the model estimation procedure. In spite of the large standard errors, most coefficients have been estimated with sufficient precision to distinguish between the parameter values. The highest uncertainty is associated with categories one and five, which were chosen much less often than the other three categories. The lack of data points for these categories, which, as we shall see, is a consequence of the poor quality of some of these items, aggravates the uncertainty inherent in the analysis.

Item characteristic curves The parameters shown in table 2.3 can be used to compare the countries on the relationships mentioned. But they do not provide a complete picture of the quality of the indicators. We are primarily interested in the conditional probabilities of belonging to each category of the observed variables given the latent class, and these probabilities are also determined by the intercepts in equation 2.1. As a clarification, one can consider that in an analysis with two classes and two observed categories the conditional probabilities would be the true positives and false negatives rate. To shed more light on the precise relationships the traits have with the indicators, therefore, we also provide plots of the so-called item characteristic curves[9]. These are sometimes also called 'item-category response functions'.

Figure 2.6 provides the curves describing the conditional probability of belonging to each category, given the score on the latent trait the variable is supposed to measure.

[9]Note that here we show the conditional probabilities rather than the cumulative probability often graphed.

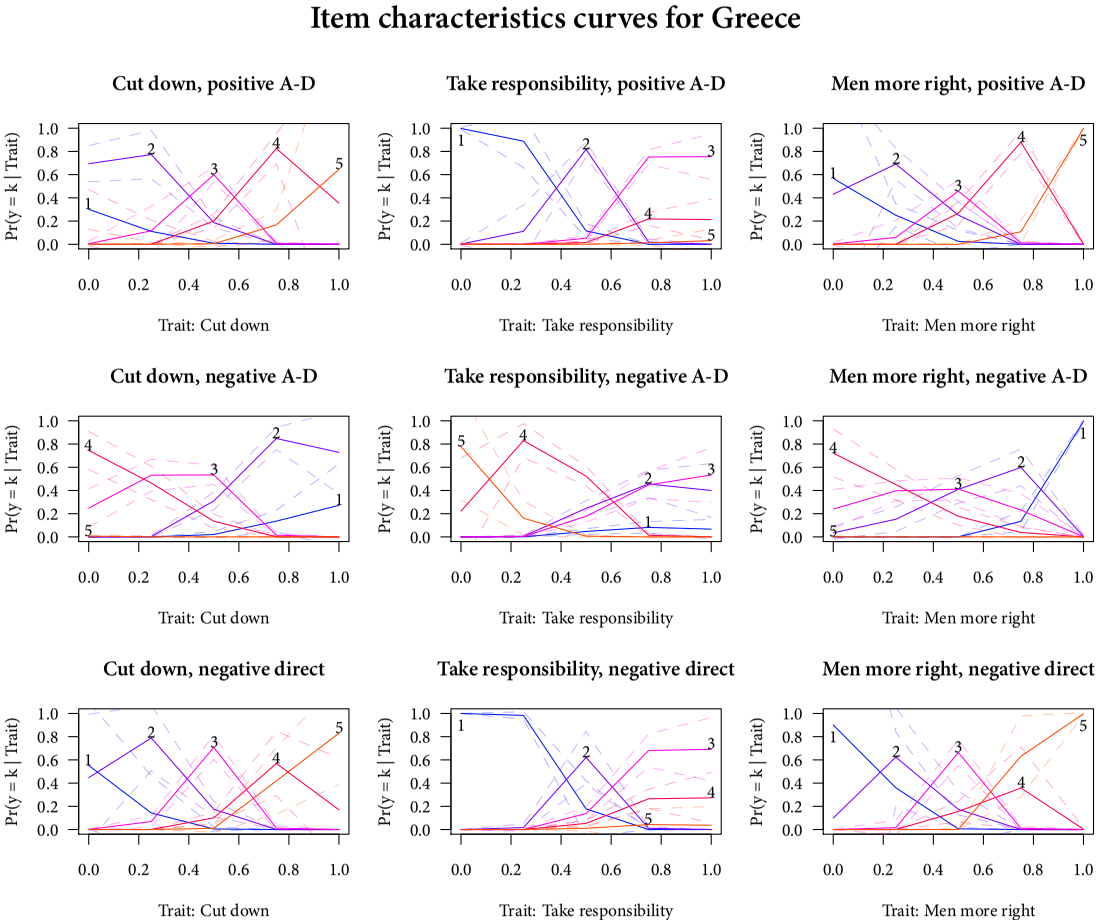


Figure 2.6: Item characteristic curves for Greece. The lines indicate the probability of choosing a category, given that value for the trait to be measured. Each line is marked with its category number at the point where this probability is highest (its peak). The dotted lines are approximate 95% confidence intervals around the probabilities. The columns show measurements of the three different traits, while the rows show measurements using the three different methods.

In the figures, the three methods of asking the question correspond to the rows, so that the first row contains the three graphs of the ICC's for the first method (main questionnaire), and the second and third rows the graphs for the supplementary questionnaires. The columns and graph titles correspond to the three traits described above.

The solid lines in the graph correspond to the probability of choosing a category, given the trait score. The lines have been marked with a color and the number of the category. This category number is moreover plotted at the point where the conditional probability of choosing that category is highest, i.e. at the peak of the item characteristic curve. If an item is ordinal then the ICC's peak in succession, and one will read either '1 2 3 4' or the reverse from left to right. It is also of interest whether the peak is high (close to one) or not, as this is an indication of the specificity of the category. Last, the peak should ideally not be underneath another curve.

The dotted lines provide approximate 95% confidence intervals around the ICC's. This is an example of the richness of the output that can be obtained from Latent Gold. The uncertainty noted in table 2.3, which is considerable for the extreme categories, is again reflected here.

The top left graph for Greece shows that item 'cut down' from the main questionnaire has very good measurement properties in this country. The category curves peak in succession, meaning that all categories provide information about the score on the latent trait. Moreover these peaks, which can be likened to the probability of true positives or sensitivity in two category models, are quite high, in the $Pr(y = k|T_1 = k, M = E(M)) = 0.8$ range, except for the first category. Since all the curves are steep, the probability of choosing any other category than the modal one–false negatives or (un)specificity–decreases sharply. This graph is highly similar to the same graph as calculated from the probit IRT model by the program Mplus (Muthén & Muthén, 1998) based on our previous research. Thus for this indicator the probit IRT or categorical factor analysis model may describe the relationship between trait and indicator adequately.

The same graph for Slovenia (figure 2.7) is quite different. Here it is clear information is only being obtained from the three middle categories. The extreme categories are hardly used at all. For these middle categories, however, the peaks are successive and relatively high for categories 2 and 4. Thus, although not all categories are used, resulting in a loss of information, the discriminating power of the three middle categories is quite good.

This is not the case for the same item measured by the second method in Slovenia. Here the quality is extremely low, as almost no discriminating power exists except for choosing the second category versus all the others. In general the measurement quality for the second method is much worse in both countries. The third method fares better in Greece then it does in Slovenia, where the measurement quality is disastrous; in the second item only choosing the first versus all the other categories provides any information.

It can also be seen that in general the measurement properties in Slovenia are worse than in Greece. This is in line with the findings from CFA and ordinal probit models; in fact, it was the reason these two countries were selected. As an example one can compare item 2 'take responsibility' in the main questionnaire across the countries. In Greece again only the three middle categories provide good measurement properties. Thus this item has intermediate quality. But in Slovenia the middle category is equally likely to be chosen for all values of the trait, and in fact just as likely as categories 4 and 5. The only differentiation one can make between people on the latent trait comes from a distinction between

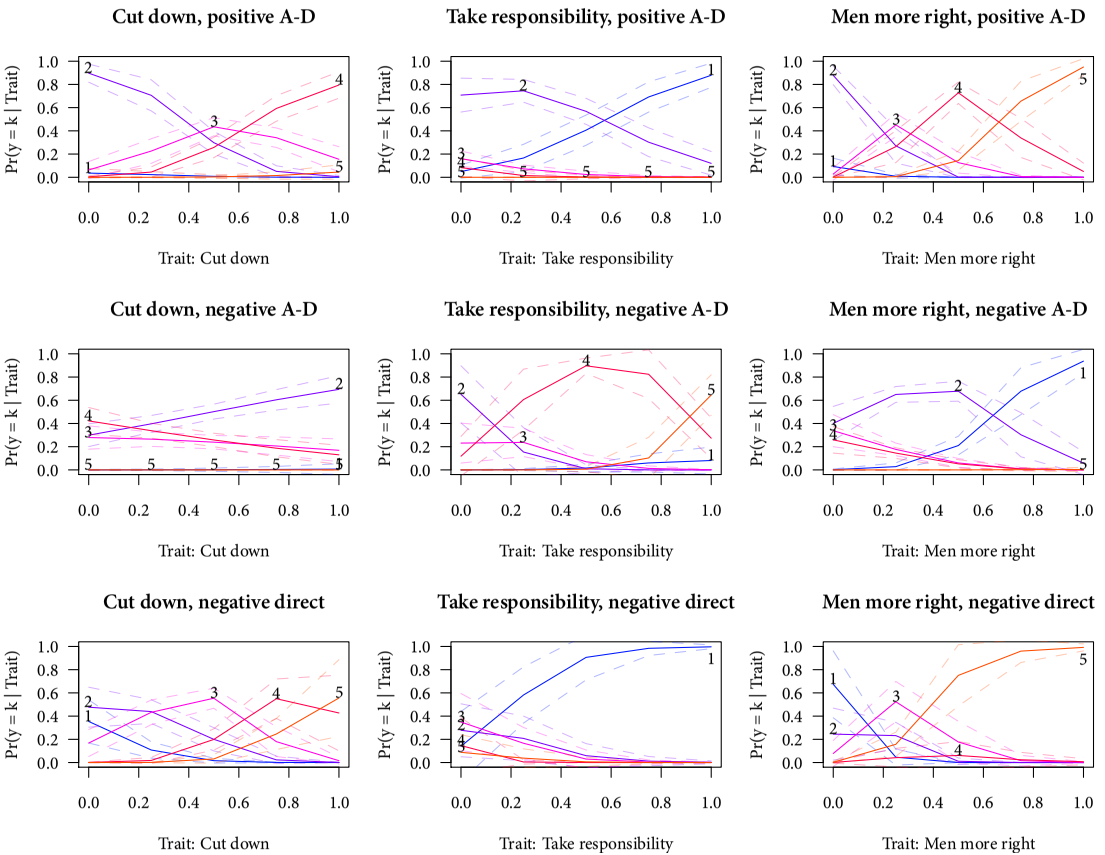


Figure 2.7: Item characteristic curves for Slovenia. The lines indicate the probability of choosing a category, given that value for the trait to be measured. Each line is marked with its category number at the point where this probability is highest (its peak). The dotted lines are approximate 95% confidence intervals around the probabilities. The columns show measurements of the three different traits, while the rows show measurements using the three different methods.

categories one and two[10]. This item has an extremely low quality in all models we have examined for these data so far; in the continuous CFA model the percentage of variance explained in the item by the trait was estimated at 25%. An explanation is now found in the extremely limited use of the scale. In the estimated marginal distribution (prevalence) of this opinion in Slovenia the proportion of people in categories of the latent trait associated with disagreement is below 0.10.

When the comparisons described above are made across the methods it can be seen in figures 2.6 and 2.7 that the second row consistently has worse measurement properties than the other two rows. The first and third methods have comparable measurement quality. The same conclusion was also drawn in the categorical CFA analyses that were conducted earlier. The linear CFA analysis suggested that the first method was slightly better than the other two.

Item information A more direct measure of the quality than has been used so far is the item 'information'. It is the inverse of the error variance of the maximum likelihood estimate of the trait that one can get from each item, and can be seen as a generalized reliability. The information function $I(T)$ is a measure of precision in the estimation of the trait T: $\sigma(\hat{T}) = 1/\sqrt{I(T)}$. Thus, as the curve approaches zero, less and less can be said about the person's trait score. The item information functions are shown for all items in figures 2.8 and 2.9. For more details about the information function and how it was computed we refer to the appendix.

Because of the non-linear specification of the model, and contrary to CFA, the information varies across levels of the trait[11]. Instead of a single number, a plot is obtained across the range of the trait. One can also obtain the marginal or average information in a particular country by averaging over categories of the trait, weighting the information at each category by the prevalence of that category (e.g. Donoghue, 1994). This average information is a single number that provides the expected information for that country. It is important to note, however, that it depends on the marginal distribution of the trait: items in two countries with the same information curve but different marginal distributions will in general provide different average information.

In the figures the information has been plotted on a log scale to allow for comparison of the different items, which vary widely in information provided. Therefore any visible differences in height of the curves are usually substantially large. One can appreciate the absolute values of the curves by considering for example that an information value of 74 (the average for the direct version of item 1 in Greece) implies that the best estimate of the latent trait for a particular person that one can obtain with this item will have a standard deviation of 0.12, on a scale of 0 to 1. One can also compute the relative efficiency of two items as the ratio of their information (Hambleton et al., 1995). The average information in the country has been indicated at the top of the graphs.

The agree-disagree versions of the 'cut down' item in Greece are clearly asymmetrical. This implies that opinions against the 'feminist direction' are measured much less accurately than 'pro-feminist' opinions. The item-specific scale is much better overall and also provides better coverage of the entire range of opinions. It is for the population studied

[10]Note that the scale of this latent trait has been reversed relative to Greece. This ordering of the classes is arbitrary and does not affect the results.

[11]A complication omitted here is that it also varies across levels of the method. The curves shown provide the marginal information collapsed over categories of the method. See the appendix for an explanation.

Figure 2.8: Model-based item information functions for **Greece**. Note the log scales.

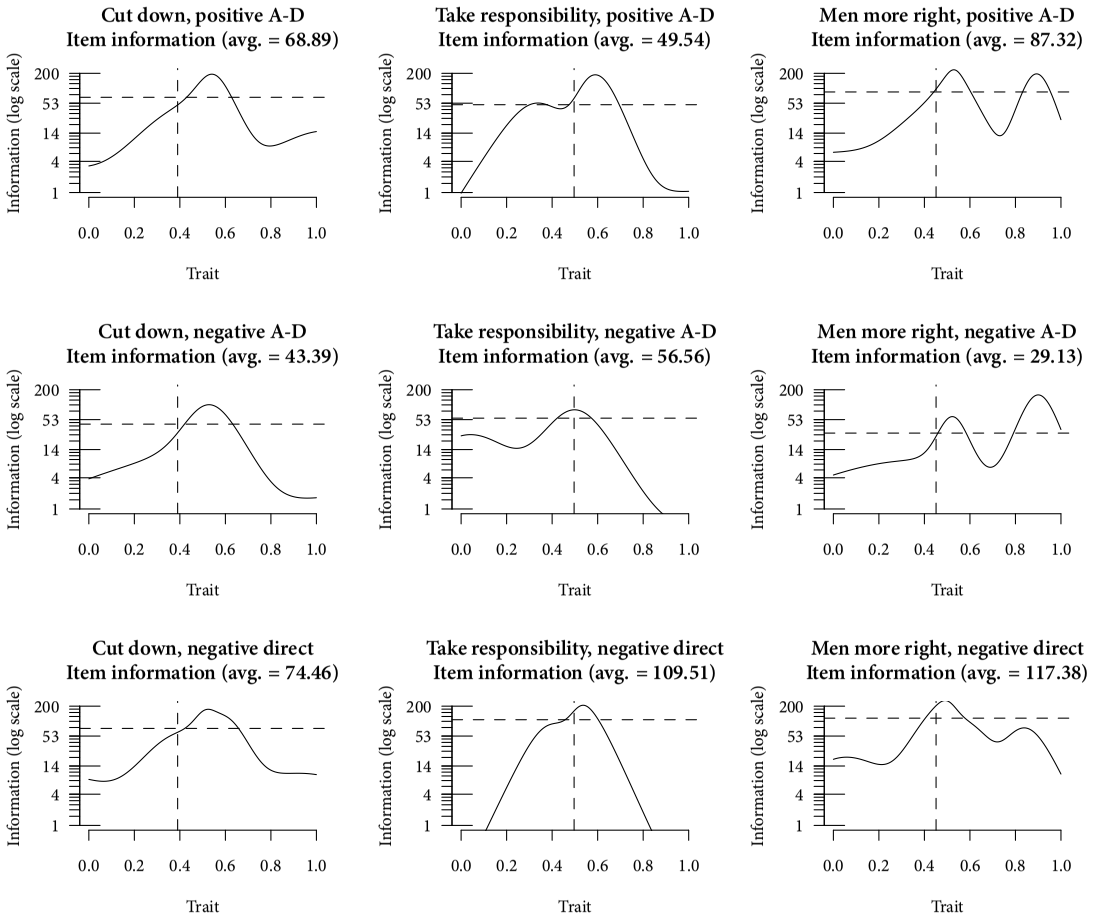
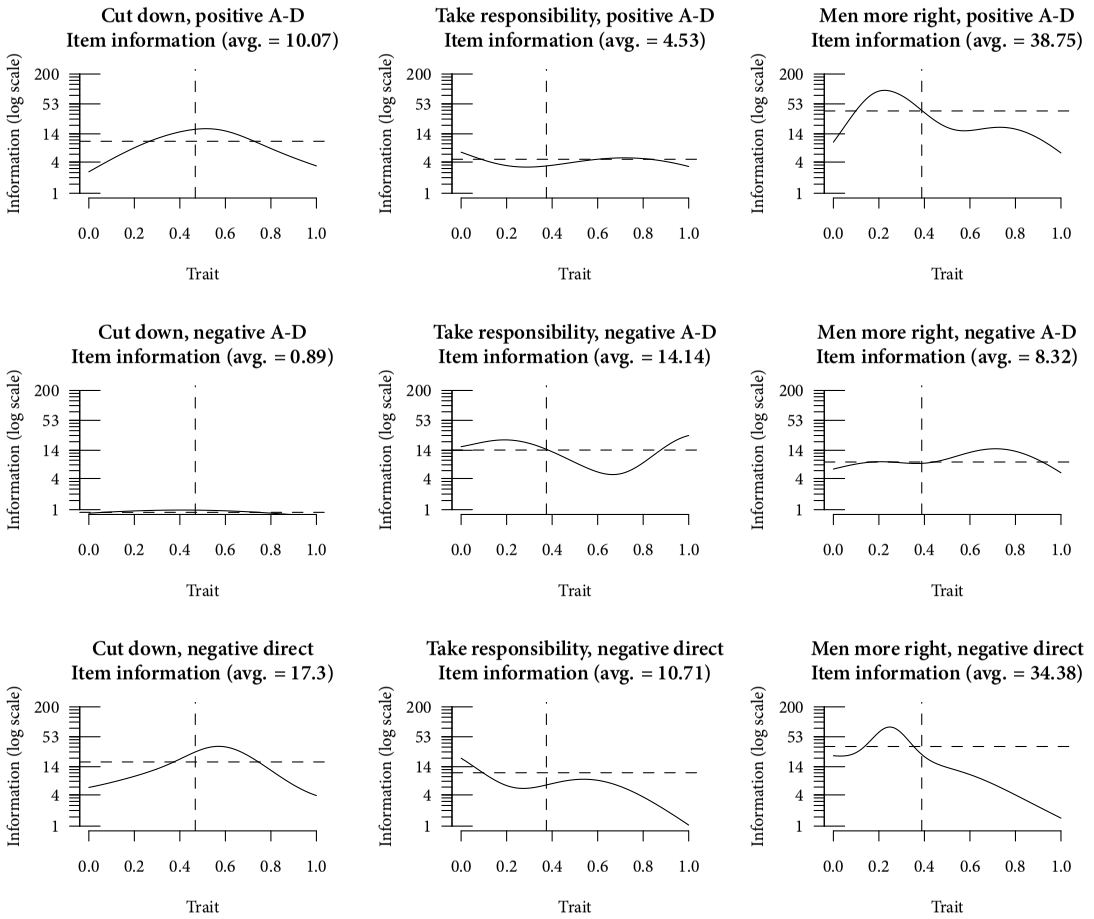


Figure 2.9: Model-based item information functions for **Slovenia**. Note the log scales.



slightly (1.1 times) more efficient than the first method and 1.7 times as efficient as the second method.

The direct version of the 'take responsibility' item has a very high peak and provides much more information about opinions close to the average Greek opinion than the agree-disagree scales. However, away from the average the information provided is much higher for the first two methods. In principle these are therefore better adjusted to measure relatively 'feminist' or 'anti-feminist' opinions than the item-specific scale, where 'feminist' opinions are again better measured than 'anti-feminist' ones. On average the item-specific scale is still about twice as efficient as the agree-disagree scales for the population studied.

The 'men more right' item has high quality overall, and covers the whole range of opinions quite well. In this respect the item-specific scale again does much better than the agree-disagree scales, whose information curves are skewed towards the measurement of 'feminist' opinions. On average the item-specific scale is 1.3 times more efficient than the positive agree-disagree version, and much (4 times) more efficient than the negative agree-disagree version.

The graphs for Slovenia immediately reveal the much lower overall measurement quality of the items in that country. The median item in Greece is 4.3 more efficient on average than in Slovenia. This is more than the largest information ratio in Greece. Thus the differences between the countries are much larger than the differences between the items within each country.

'Men more right' is also the better item in Slovenia. There the measurement is skewed towards measurement of 'pro-feminist' opinions for both the direct and positive agree-disagree versions. Contrary to the pattern found in Greece that the item-specific scale provided more equal measurement across the whole range of the scale, in Slovenia the item-specific scale's information curve is more skewed than the other two method's curves. The average information for the Slovenian population is not very different, however, reflecting the highly skewed marginal distribution of the trait in that country.

The 'take responsibility' trait is not well measured in Slovenia. The negative agree-disagree is slightly (1.3 times) better than the item-specific scale, and seems to be able to pick up also negative opinions somewhat. This is the only item where the negative agree-disagree scale is better than the other two methods.

'Cut down' negative agree-disagree is the worst item of all. It has a variance which is higher than the entire range of the trait scale. This means one knows about as much about a Slovenian's opinion after asking this question as before asking it. The other measures are better, the direct version being the best of the three.

Overall we found that the item-specific scales were better than the agree-disagree versions (3.5 times more efficient on average), and that the positively formulated items were better than the negatively formulated ones (1.8 times on average). This finding is in line with Saris et al. (frth). The difference in quality between the countries that motivated the choice of countries in the first place was also clearly found.

Method effects

So far we have only discussed the relationship between the traits to be measured and their observed variables, that is, the quality of the questions as indicators of the trait they are supposed to measure. As discussed previously, another important part of answers to survey questions can be described as 'method effects'. These have been modeled in our case as

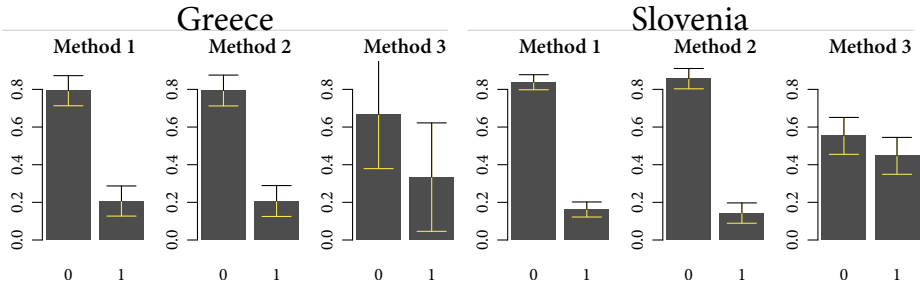


Figure 2.10: Estimated histograms of the latent method factors with approximate 95% error bars.

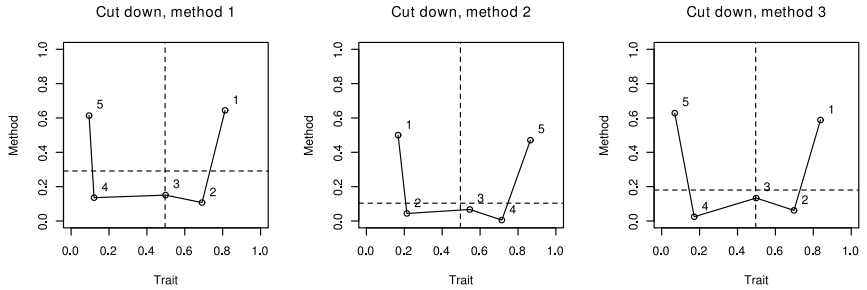


Figure 2.11: Bi-plots for the first item, 'Women should be prepared to cut down on their paid work', for all three methods in Greece. Plotted is the conditional mean of the trait (horizontal axis) and method factors (vertical axis) influencing the item given that a particular category (the points labeled with a category number) was chosen.

latent variables that affect the answers in the same way for questions asked in the same way, but are unrelated to the traits to be measured.

The coefficient estimates for these method factors are shown in table 2.3 and were already discussed. The method factors found in both countries represent a contrast between only using the middle categories or giving extreme answers, more so on the socially desirable side. The fact that class 1 on the method factor is more associated with disagreement is not in line with the hypothesis that respondents tend to 'acquiesce', that is, to tend to agree with any statement.

The estimated proportion of people (with approximate 95% confidence intervals) in each of the two categories of the method factors is shown in figure 2.10. The majority of people are in the class which uses only the middle categories. But a substantial proportion of people also are extreme responders. For the third method there are more extreme responders, with the difference in proportions of extreme and middle responders not statistically significant. This may be a consequence of the fact that the first two methods were fully labeled scales while the third method has only the two extremes labeled, perhaps attracting more responses.

An instructive way of examining the relationship an item's categories have with its trait and method factors is through a bi-plot (Magidson & Vermunt, 2001). Bi-plots for the first item ('women should be prepared to cut down on paid work') for Greece are shown in figure 2.11. The plots shown in this figure plot the conditional mean of the trait and method factors given a choice for each of the five categories of the items. The plot again

makes the meaning of the method factors readily apparent: categories 1 and 5 versus the rest.

When one projects the points (categories) onto the trait axis, it can be seen that the categories are quite unevenly spaced as was already remarked. For the first method categories 5 and 4 are much closer together than categories 1 and 2, suggesting the scale is not symmetric. The same happens in the opposite direction for the second method. For this method choosing the middle category represents an above-average opinion, suggesting the phrase 'neither agree nor disagree' does not have its intended neutral meaning in this case. In all three methods two of the categories are much closer to the middle category than the other two. It can also be seen that the method and trait factors represent very different things as they are unrelated.

External validation of the method factors The model as estimated so far appears to give valid inferences about the items. However, the meaning of the method factors has been assumed rather than checked by using external data. We now do this, demonstrating how one can use the factor score estimates of the latent class MTMM model to perform additional analyses.

It was suggested that the method factors represent an 'extreme response style' (ERS), and, to a much smaller extent, social desirability. This conclusion was based on the coefficient estimates. We now test the same conclusion with data not used in the model. Similar studies have been done using CFA by Billiet and McClendon (2000) and using a latent class factor model by Kieruj and Moors (frth).

One important reason for this is that the literature on ERS suggests that it is a stable personality trait that is different for different people but the same across all questions for each person (Billiet & McClendon, 2000). In the MTMM model developed above, however, the method factors are independent, suggesting that ERS on one set of items does not imply ERS on another. Therefore the correlation with external measures can be seen as a validation of the MTMM model.

We selected 39 variables from the ESS main questionnaire that had an answer scale on which extreme response was possible. In order to prevent confounding of variables, the items on position of women studied here were excluded. A measure of extreme response style was constructed by counting the number of times each respondent chose the most extreme possible categories on the answer scale (the minimum or the maximum). This variable, called 'stylesum', had mean 7.1, median 6 and interquartile range 6.

The method factor scores of the three methods was estimated for each person and added to the data set with the variable 'stylesum'. The modal (most likely) method scores (0 or 1) were also added.

We then computed the correlation between the method factor scores and the extreme response style (ERS) measure that was computed completely independently of the 'role of women' variables. We also computed the polyserial correlation between the modal category of the method and the ERS measure. The results are shown in table 2.4.

It can be seen that the first and third method factor scores correlate significantly with the independent ERS measure. Method 1 correlates much higher with this measure than the other two methods. This shows that extreme response style works differently for different items; a person who answers one type of item in an extreme manner does not necessarily do the same for another. The differences in correlation can in part also be explained by the amount of time between the questions; most questions used to measure ERS were

	Correlation with ERS		
	Method 1	Method 2	Method 3
Pearson correlation with factor scores	-0.222*	-0.027	-0.076*
Polyserial correlation with modal category	-0.261*	-0.069	-0.086*

*Significantly different from zero ($p < 0.01$)

Table 2.4: Correlations between the method factor scores and the external extreme response style (ERS) measure. This measure is constructed as the number of extreme responses on 39 other questions.

asked in the main questionnaire, closer to the questions used to estimate the factor scores of method 1. The factor scores for methods 2 and 3 were estimated from questions asked in the supplementary questionnaire, approximately one hour after the start of the main interview. This suggests that respondents may also change their reponse style during the interview.

Another method of testing the suggestion that extreme response style is a stable personality trait is to correlate it with other stable personality traits. To this end we correlated the method factor scores and modal categories, as well as the ERS 'stylesum' measure, with 26 questions from the Schwartz 'human values scale' asked in the supplementary questionnaire of the ESS (Schwartz, 1992). These correlations were very small; we found none above 0.1.

From the significant correlation of -0.26 above we can conclude that to some extent the method factors do indeed measure a response style independent of the content of the questions. However, the low correlations with other methods, and of the methods with a person's values, it would appear that the this response style is not a stable personality trait but can vary across methods and even during the interview. Thus the model with separate method factors used here appears more warranted than a model with one style factor that represents a personality trait of extreme response tendency. It should also be noted that it would be very difficult to model extreme response using a traditional or ordinal factor analysis model.

It was clearly shown that the method factors represent a middle versus extreme response. It was also suggested that, to a much smaller extent, they represent susceptibility to socially desirable answers to some extent. So far this claim does not rest on much more than the fact that one of the positive log-linear coefficients for each method factor is larger than the other one, and this happens on the socially desirable side. However, it can also be validated directly.

In the European Social Survey besides the data from the main and supplementary questionnaires data was also gathered in interviewer questionnaires. These included a question for the interviewer on whether 'anybody [was] present, who interfered with the interview'. If the method factor truly represents social desirability in part, then the probability of belonging to the classes should be influenced by the presence or absence of another person during the interview. It is not completely that simple, however; given the content of the questions men and women should show opposite behavior depending on whether their partner is present. Also, presumably, religion plays a role in what is considered desirable.

When each of the items are regressed on explanatory variables and the presence of another person during the interview, it is clear that this variable has a statistically significant

	Method 1			Method 2			Method 3		
	Est.	s.e.	t	Est.	s.e.	t	Est.	s.e.	t
(Intercept)	1.58	0.24	6.65	3.15	0.47	6.76	2.37	0.29	8.19
Other person present	-0.04	0.51	-0.08	-0.07	1.03	-0.07	1.13	0.81	1.39
Female	-0.12	0.13	-0.93	-0.63	0.30	-2.13	0.24	0.17	1.41
Religion	-0.14	0.22	-0.61	0.21	0.42	0.49	0.37	0.26	1.41
Church attendance	0.10	0.05	1.94	0.26	0.11	2.35	-0.12	0.06	-1.96
Married	0.01	0.13	0.07	-0.21	0.28	-0.76	-0.26	0.17	-1.53
Present×Female	0.00	0.28	0.00	-0.60	0.73	-0.83	-0.86	0.34	-2.53
Present×Religion	-0.17	0.48	-0.36	1.49	0.80	1.87	-1.53	0.78	-1.96
Present×Married	0.29	0.29	0.99	-0.57	0.85	-0.67	0.69	0.35	1.99
N		2401			2401			2401	
$Deviance$		2118			608			1460	
$-2LLR(Model\chi^2)$		6.52			18.03*			24.52*	

Table 2.5: Logistic regression of the probability of belonging to the first class on each method as influenced by the presence or absence of a third person during the interview, mediated by different variables that influence what is socially desirable: gender, religion, and marital status.

influence. For women the expected mean of the 5-point scale increases by 0.5 when their partner is not present compared to when they are. For men this effect is in the opposite direction but much smaller[12]. Thus it is clear that social desirability effects are present in the items. It remains to be investigated, however, whether the method factors estimated in our analysis account for the effect of social desirability.

For this reason we created a data set that combines the original variables from the ESS main, supplementary, and interviewer questionnaires with the factor scores (modal classes and probabilities) obtained from our LCM analysis. We then regressed the logit of the probability of belonging to the first class of each method factor on presence of another person, as well as gender, religion, marital status, and their interactions with the presence or absence of another person during the interview. The results of this analysis for the three method factors for Greece are shown in table 2.5.

It can be seen in the table that the effects for the first method are all non-significant. For the second method there are main effects of church attendance and gender, and the interaction effects, though not statistically significant, are in the expected direction. For the third method the model is clearest. All of the interaction effects as well as the main effect of church attendance are statistically significant.

To give an example of the meaning of the above analysis, consider the estimates for the third method. For a woman who is religious, the probability of moving into class 1 of the method factor increases from 0.80 to 0.93 if somebody is present at the interview. This in turn increases her chances of saying that 'a woman should be prepared to cut down on her paid work', for instance. Incidentally it also decreases the chance that she will use an extreme category considerably. All these effects happen, in our model, while keeping her trait score constant. Thus any difference in the answers provided by respondents differing on the characteristics in table 2.5 has nothing to do with a change in their underlying opinion.

[12]The model controls for age, gender, education, religion, living with a partner, and marital status. The analysis is not shown here but can be obtained upon request from the first author.

It should be noted that the social desirability effects found on the method factors are small relative to the effects found on the items themselves. This suggests that there is an element of social desirability, different across respondents, that still remains to be explained. The model could be expanded to include an acquiescence style factor, for instance (Billiet & McClendon, 2000). However, it is questionable whether a model with such an extra latent variable can be estimated with the experimental design used here. This remains a topic for further investigation.

2.7 Conclusion

The goal of this study was to show how more general measurement models can be formulated, and in particular to demonstrate the use of latent class models for analysis of the quality of single questions.

We have formulated latent class factor models from our general graphical model and applied these models to a multitrait-multimethod experiment on the role of women in society. Furthermore we compared the results for two countries, one of which was previously estimated to have low question qualities (Slovenia) and the other (Greece) high ones.

We investigated the quality of the questions using the item characteristic curves and information functions. To our knowledge this paper is the first to provide formulas for the item information function of latent class factor models (see appendix).

The investigation of question quality using the LCM yielded a wealth of information about the functioning of the questions. It was established that for the agree-disagree scales only the middle three categories (out of five) provide information about the traits to be measured, and that for the items with exceptionally low quality the number of categories providing information is reduced even more.

The quality in Slovenia was again found to be much lower than in Greece, in line with previous findings. It was also clear that the positively worded items were better than the negatively worded ones, and that the item-specific scales provided much more information than the agree-disagree format. With two exceptions, they also provided more equal information across the whole range of the trait. The agree-disgree versions provide more accurate measurement of 'pro-feminist' than of 'anti-feminist' opinions.

This finding is important: items with an approximately equal amount of information across the range of the trait are desirable, especially in cross-national research.

An item with much skew in its information function is less likely to be useful for cross-national comparisons. This is so because even if the information functions were the same in all countries, countries with higher average opinion would have a higher measurement quality[13]. As measurement errors affect the analysis of means and regression (Fuller, 1987), differential measurement errors across countries invalidate comparisons of means and relationships.

In the present analysis we found that the information functions for Slovenia and Greece

[13]Skew in the amount of information will also bias regression analyses with interactions. As an example, consider the 'take responsibility' item's information function in Greece. A regression of a dependent variable on 'take responsibility', a third variable, and their interaction is formulated. Figure 2.8 shows that agreement is measured accurately, while disagreement is not. Thus two groups of Greeks which have a different average opinion on the trait will have different amounts of measurement error in this item and therefore different correlations with other variables. Therefore a regression analysis which includes both a main effect and an interaction with the opinion on this item will give biased estimates.

were very different. Thus it is not clear that the lower quality in Slovenia is due to a difference in the average opinion. The analysis shows that the difficulty of categories indicating 'anti-feminist' opinions was far higher in Slovenia. Thus most Slovenians are left with only two choices, 'agree' versus 'neither agree nor disagree', which are used differently by different people as evidenced by the method effects. Clearly this item does not measure opinions in a way equivalent to the way they are measured in Greece. If answers are to be compared or the items in Slovenia to be analyzed, therefore, improvements should be made. One suggestion would be to rephrase the question so that it is less extreme.

We examined the method effects. Bi-plots showed clearly what the method factors represent: a distinction between extreme versus middle responses. Most people (about 80%) were found to use only the middle categories. This is a strong indication of satisficing; the question might not be clear enough or too cognitively difficult to answer. The parameter estimates suggested that the method factors also represent a susceptibility to answering in a socially desirable way. This was investigated by regressing the estimated method factor scores onto the presence or absence of a third person during the interview. Effects were found for the third and second methods, but not the first. Considering that the effects of this variable on the items are much larger precisely for the first method, there may still be room in the model for a social desirability or style factor. Such an study would also shed light on the plausibility of the assumption we have had to make that all systematic errors are specific to the method of asking the question. This is, however, is outside the scope of the present study.

The latent class analysis elaborated in this paper provides much information about the precise workings of the items, as well as suggestions for their improvement. Furthermore this was achieved without any assumption of normality or of parallel probability curves. Indeed these assumptions, made in (ordinal) confirmatory factor analysis models usually applied to MTMM data, were found not to hold. Therefore we hope to have shown the utility of this approach for the evaluation of categorical items using multitrait-multimethod designs.

Appendix: Phrasing of the questions used in the 'role of women' experiment

Figure 2.12: Main questionnaire (method 1)

CARD 59 I am now going to read out some statements about men and women and their place[81] in the family. Using this card, please tell me how much you agree or disagree with the following statements.

		Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
G6	A woman should be prepared to cut down on her paid work for the sake of her family. [82]	1	2	3	4	5	8
G7	Men should take as much responsibility as women for the home and children.	1	2	3	4	5	8
G8	When jobs are scarce, men should have more right[83] to a job than women.	1	2	3	4	5	8

Figure 2.13: Supplementary questionnaire

Please indicate how much you agree or disagree with each of the following statements about men and women and their place in the family.

iS8[26] "A women should not have to cut down on her paid work for the sake of her family."
Please tick one box.

- Agree strongly ☐ 1
- Agree ☐ 2
- Neither disagree nor agree ☐ 3
- Disagree ☐ 4
- Disagree strongly ☐ 5

iS9[27] "Women should take more responsibility for the home and children than men."
Please tick one box.

- Agree strongly ☐ 1
- Agree ☐ 2
- Neither disagree nor agree ☐ 3
- Disagree ☐ 4
- Disagree strongly ☐ 5

iS10[28] "When jobs are scarce, women should have the same right to a job as men."
Please tick one box.

- Agree strongly ☐ 1
- Agree ☐ 2
- Neither disagree nor agree ☐ 3
- Disagree ☐ 4
- Disagree strongly ☐ 5

(a) Method 2

iS22[57] If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.
Please tick one box.

	1	2	3	4	5	
A woman should be prepared to cut down on her paid work for the sake of her family	☐	☐	☐	☐	☐	A woman should not have to cut down on her paid work for the sake of her family

iS23[58] If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.
Please tick one box.

	1	2	3	4	5	
Men should take as much responsibility as women for the home and children	☐	☐	☐	☐	☐	Women should take more responsibility for the home and children than men

iS24[59] If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.
Please tick one box.

	1	2	3	4	5	
When jobs are scarce, men should have more right to a job than women	☐	☐	☐	☐	☐	When jobs are scarce, women should have the same right to a job as men

(b) Method 3

Appendix: Item information function for the polytomous Latent Class Factor Model

In this section we explain how to obtain the variance of the best estimate of the trait T that one can obtain from an observed item y. This variance is also called the item information, as it equals the Fisher information in the likelihood of the item given only the trait (Hambleton et al., 1995).

The observed variable has a multinomial likelihood:

$$L = \prod_{k=1}^{K} P_k^{U_k},$$

where U_k is an indicator function that equals 1 if $y = k$ and 0 otherwise.

The item-category response function for the model used is

$$p(y = k|T, M) := P_k(T, M) = \frac{\exp(a_k + b_k T + m_k M)}{\sum_{c=1}^{K} \exp(a_c + b_c T + m_c M)}. \tag{2.2}$$

For succinctness we will refer to $P_k(T, M)$ simply as P_k. This model is very similar to a generalized partial credit model (PCM, see Muraki (1993)). Indeed, if for a given relationship one replaces the category scores for the observed variable in the PCM by the slope for that category and sets the discrimination parameter to unity, identical first and second derivatives result.

The equation above gives the conditional probability of choosing category k, given both the trait and the method. In total there are K categories and item-category response functions. These functions are also called the item characteristic curves.

The item information function (IIF) is now equal to the Fisher information in the item, with respect to the trait:

$$I(T) = -E(\frac{\partial^2 \ln L(T)}{\partial T^2})$$

For any given value of M we can derive the second partial derivative of the item likelihood with respect to T as

$$\frac{\partial^2 \ln L}{\partial T^2} = \sum_{k}^{K} [U_k(\lambda^2 - \nu)],$$

where

$$\lambda = \sum_{k}^{K} [b_k P_k];\ \nu = \sum_{k}^{K} [b_k^2 P_k].$$

A proof follows from the derivation for the PCM in the appendix of Donoghue (1994), replacing in that paper the quantities D and a by 1 and all category scores (k, c) by the slope b_k for that category.

Noting that $E(U_k|T, M) = P_k(T, M)$ and $E(x|T) = \sum_l^L E(x|T, M)p(M = l)$ for any random variable x, we can conclude that the information in the item about T, conditional on M is

$$I(T|M) = \sum_{k}^{K} \beta_k^2 P_k(T, M) - [\sum_{k}^{K} \beta_k P_k(T, M)]^2. \tag{2.3}$$

and the marginal information then equals

$$I(T) = \sum_{l}^{L} [I(T|M)\ p(M = l)], \tag{2.4}$$

where the index l runs over all scores of the method factor. In the model selected in this paper $l \in \{0, 1\}$.

One can also calculate a trait score estimate for each person based on the parameters. The standard error of the estimation of this score then equals $1/\sqrt{I(T)}$ (Hambleton et al., 1995).

Chapter 3

Joint estimation of survey error components in multivariate statistics

Abstract

We outline a procedure for simultaneously estimating the "design" effects of different survey error components, in the context of structural equation models. The effects of clustering, measurement error, and non-normality, are jointly estimated for an example multivariate model involving reciprocal effects, instrumental variables, correlated error terms, and measurement error. The example is estimated on real data from the European Social Survey 2008.

It is shown how estimates of the effects of these different survey error components can be obtained. In the example given, it is also shown that the relative sizes of these effects are very different than commonly found in the estimation of means and totals. In particular, measurement error is an important factor in our example.

Finally, it is remarked that our general knowledge of the relative importance of different survey error components for multivariate statistics could be greatly increased by the application of the method discussed in this paper to a cross-section of real analyses.

3.1 Introduction

The total survey error literature is rich in discussions of the effects of different error components on the bias and variance of mean estimators (starting points for the abounding literature are Cochran, 1977; Groves, 1989). Effects on parameters of multivariate models are also discussed, though less often (e.g. Scott & Holt, 1982; Lyberg, 1997; Biemer et al., 2004). Such discussions, however, focus mostly on bias in the multivariate statistics due to survey errors, a notable exception being Kish and Frankel (1974). Here we will attempt to take also into account the effect on the variance of the estimates, with a focus on the simultaneous estimation of the effect of different error sources.

The same error components that affect means may also affect regression coefficients and other multivariate statistics such as factor loadings, and latent variable variances. The influences and relative influences they have, however, cannot be expected to be similar. For example, the design effect due to clustering for a mean is

$$1 + (c - 1)\rho, \tag{3.1}$$

where ρ is the intraclass correlation (icc) and c is the common cluster size (e.g. Cochran, 1977). Compare this with the same design effect for a simple regression coefficient, which approximately equals

$$1 + (c - 1)\rho_\epsilon \rho_x, \tag{3.2}$$

where ρ_ϵ is the icc for the residuals and ρ_x the icc for the independent variable (Scott & Holt, 1982): clearly the design effect due to clustering for the regression coefficient is in general smaller than the design effect on means.

Random measurement error, meanwhile, does not add systematic errors and can therefore be ignored in the analysis of bias in means and totals[1]. On the other hand, regression coefficients are well-known to be biased by measurement error if left uncorrected (Fuller, 1987, 3). For example, in a simple linear regression of a dependent variable Y on an observed independent variable X, denote the linear regression slope of Y on X by γ. If X is not a perfect measure, but contains measurement error, X has a reliability ρ_{xx}, sometimes termed 'reliability ratio' in the survey error literature (e.g. Groves, 1989). If the true regression slope is denoted by β, the relationship between the regression slope of observed variables, the true slope, and the reliability is (Fuller, 1987, 5):

$$\gamma = \rho_{xx} \cdot \beta. \tag{3.3}$$

Thus the bias in the regression slope uncorrected for measurement error is multiplicative in the reliability. Since reliabilities of 0.7 are not uncommon (Saris & Gallhofer, 2007b; Alwin, 2007), large biases can occur.

Measurement error in the independent variable, when left uncorrected, will also decrease the explained variance and thus increase the variance of the estimated (standardized) coefficient. Since the mean square error (MSE) of the regression coefficient is the sum of bias squared and variance, it follows that measurement error influences the MSE through both bias and variance increase.

[1] Obviously measurement error biasing effects also exist in the form of systematic errors ('relative bias'); these do affect mean estimation (see also Biemer & Trewin, 1991).

Depending on sample size, cluster size, and effect size, the effect on the mean square error of measurement error may exceed that of clustering or vice versa. Thus, careful attention should be given to the relative sizes of the effect of different error sources: they cannot be assumed similar to those for means and totals.

We will in turn discuss different error components, both sampling and nonsampling, and how their impact on multivariate statistics may be measured. We do this in a structural equation model context developed in the following section. Due to the generality of structural equation models, our discussion will cover (multivariate) regression, factor analysis, longitudinal models, and models with ordinal, count, and censored variables among others (Muthén, 1984).

3.2 Structural equation models

Suppose a sample of J clusters has been drawn from a population, and in total n persons have been selected within the clusters by random sampling with probability weights w_i. In this way, n observations of the measures y are observed. These observed variables are assumed to be imperfect measures of a vector of variables η, with possibly correlated random measurement error ϵ. Systematic stochastic measurement errors such as method and style effects can also be incorporated through the latent variable structure (see e.g. Werts & Linn, 1970; Jöreskog, 1970).

A structural equation model (SEM) can then be specified as

$$\eta = B_0\eta + \zeta, \tag{3.4}$$

$$y = \Lambda\eta + \epsilon, \tag{3.5}$$

$$\forall_{k\in K}\forall_{l\in L}(E(\epsilon_k, \zeta_l) = 0), \tag{3.6}$$

with $\Phi_{K\times K}$ the covariance matrix of the latent variable disturbance term vector ζ and $\Psi_{L\times L}$ the covariance matrix of the measurement error variables ϵ. This specification is known as the "LISREL all-y model" (Jöreskog, 1970). Other well-known model formulations are the Bentler-Weeks model (Bentler & Weeks, 1980) and the RAM model (McArdle & McDonald, 1984). All three models can be re-written into equivalent specifications to fit the form of the other models. The parameters of the model are collected into a vector θ.

We assume there is a matrix of observed variances and covariances S on the p observed variables that converges in probability to a population covariance matrix Σ. The $p(p+1)/2$ unique elements of S can be collected into a vector $s := \text{vech } S$.

The implied variance-covariance matrix by the model above is then

$$\Sigma(\theta) = B^{-1}\Lambda\Phi\Lambda' B^{-1} + \Psi, \tag{3.7}$$

where $B := I - B_0$. We collect the unique elements of $\Sigma(\theta)$ into a vector $\sigma(\theta) := \text{vech } \Sigma(\theta)$.

Given the above assumptions, the parameters of the model can be consistently estimated by minimizing the weighted least squares fitting function

$$F = (s - \sigma(\theta))' V (s - \sigma(\theta)), \tag{3.8}$$

where V is a positive definite, possibly stochastic, weight matrix (Satorra, 1989).

The weighted least squares fitting function will be equivalent to maximum likelihood estimation if V is chosen as the inverse of the model-implied fourth-order moments under normality. Thus the discussion given here encompasses normal-theory maximum likelihood as well as generalized least squares and 'asymptotic distribution free' estimation methods, among others.

Consistency is not affected by the choice of V, as long as V does not violate identification conditions. Only one choice of V is asymptotically optimal, however; namely that V such that V converges in probability to Γ^{-1}, where Γ is (a function of) the matrix of fourth-order moments of y. The consistency of the estimates also does not depend on any assumption about the distribution of y.

Under the model the asymptotic variance of the estimates $\hat{\theta}$ is

$$avar(\hat{\theta}) = n^{-1}(\Delta' V \Delta)^{-1} \Delta' V \Gamma V \Delta (\Delta' V \Delta)^{-1}, \tag{3.9}$$

where Δ is the first derivative of the implied covariance matrix $\Sigma(\theta)$ with respect to the parameters θ, and Γ the matrix of fourth-order moments (e.g. Satorra, 1989). An expression for Δ in terms of the parameters of the model was given by Neudecker and Satorra (1991). We will employ this expression to calculate the asymptotic variance under different conditions.

The matrix Γ in equation 3.9 will play an important part in the discussion that follows: it is the matrix of fourth-order moments of y, and the primary means by which survey error components affect the variance of the estimates $\hat{\theta}$. In general, if there is no clustering a consistent estimate of Γ under general conditions is given by

$$\hat{\Gamma} = n^{-1} \sum_{i=1}^{n} (b_i - \bar{b}_\cdot)(b_i - \bar{b}_\cdot)', \tag{3.10}$$

where $b_i = D^+ \vec{(}y_i - \bar{y})(y_i - \bar{y})'$ (e.g. Fuller, 1987, 332) and D^+ is the Moore-Penrose inverse of the duplication matrix (Magnus & Neudecker, 2002).

With clustering and weighting the estimate of Γ can be obtained by first aggregating to the level of the clusters while using the sampling weights. Then the estimate becomes

$$\hat{\Gamma}^{(c)} = \frac{J}{n^2(J-1)} \sum_{j=1}^{J} (b_j - \bar{b})(b_j - \bar{b})', \tag{3.11}$$

where b_j is the weighted sum of all b_i's in cluster j, replacing $\bar{y}$ with the weighted sample mean (Muthén & Satorra, 1995).

The matrices $\hat{\Gamma}$ and $\hat{\Gamma}^{(c)}$ provide consistent estimates of the fourth-order moments regardless of the distribution of y. If, however, y can be assumed to have a multivariate normal distribution, then the Γ matrix can be consistently estimated by

$$\hat{\Gamma}^{*} = 2D^{+}(S \otimes S)D^{+'}. \tag{3.12}$$

The fourth order moments are then a function only of the variances and covariances.

As we have remarked earlier, the choice of V determines the estimation procedure. By replacing Γ in equation 3.9 with $\hat{\Gamma}^{*}$, $\hat{\Gamma}$, or $\hat{\Gamma}^{(c)}$, normal-theory variances, variances robust to non-normality, or cluster and weighting-corrected variances are obtained. Here

it should be noted that the cluster-corrected variances of Muthén and Satorra (1995) are also robust to non-normality.

The elements of the measurement error variance matrix Ψ can, if identification conditions have been met, be estimated simultaneously with the 'structural' parameters B_0 and Φ. However, in practical applications, not enough information may be available to estimate the measurement error from the sample or doing so simultaneously might lead to very large models (Saris & Gallhofer, 2007a). In such cases an estimate of the measurement error in the measures y may be obtained from other sources, such as published evaluations of psychometric properties of scales or meta-analyses of measurement error estimates (Saris & Gallhofer, 2007b; Alwin, 2007). Correction for measurement error can then proceed by fixing the elements of Ψ (the 'single indicators' approach)[2].

Now that we have developed some results for the general structural equation model we will discuss an application of a structural equation model from the literature. We then proceed to separately estimate the magnitude of the error components in the variance of parameters of an analysis of this real data set.

3.3 Application of a structural equation model to real data

Saris and Gallhofer (2007a) provide an analysis of a structural equation model of social and political trust with corrections for measurement error. A simplified adaptation of their model is shown in figure 3.1. The model shown in the figure can be expressed as:

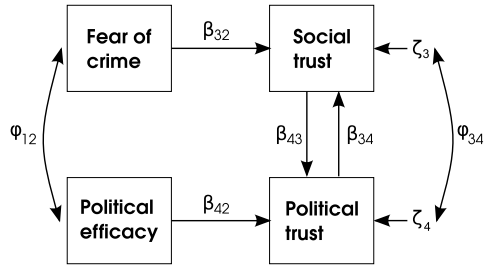


Figure 3.1: Structural equation model adapted from Saris & Gallhofer (2007a).

$$\begin{aligned} \text{SocTrust} &= \beta_{34}\ \text{PolTrust} + \beta_{31}\ \text{Fear} + \zeta_3 && (3.13)\\ \text{Poltrust} &= \beta_{43}\ \text{SocTrust} + \beta_{42}\ \text{Efficacy} + \zeta_4 && (3.14)\\ E(\phi_1\zeta_3) &= E(\phi_1\zeta_4) = E(\phi_2\zeta_3) = E(\phi_2\zeta_4) = 0. && (3.15) \end{aligned}$$

It can be seen that the model cannot be estimated with ordinary linear regression because it contains a reciprocal effect between social and political trust, which is identified by the "instruments" fear of crime and political efficacy. In line with standard practice in econometrics we allow for the possibility of a covariance between the disturbance terms ζ_3 and ζ_4.

[2]An alternative method that will not be discussed here is the so-called 'covariance reduction' approach (Saris & Gallhofer, 2007a).

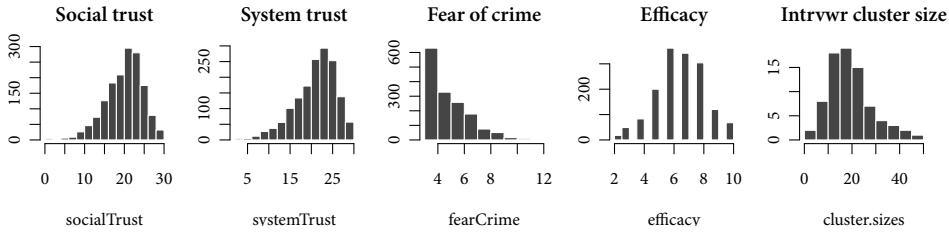


Table 3.1: Histograms of the observed variables and number of interviews per interviewer (cluster size) in Denmark.

Variable	Mean	Std dev	Skewness	Kurtosis	icc	$\hat{\rho}_x$ (s.e.)	$\hat{\psi}$ (s.e.)
socialTrust	20	4.7	-0.70	-0.80	0.25	0.73 (0.01)	6.0 (0.22)
systemTrust	21	4.8	-0.79	-0.54	0.11	0.77 (0.01)	6.3 (0.24)
fearCrime	5	1.7	0.77	-0.31	0.17	0.57 (0.02)	1.3 (0.04)
efficacy	7	1.7	-0.25	0.05	0.11	0.64 (0.03)	1.2 (0.07)

Table 3.2: Summary statistics for the Denmark dataset. The total sample size was 1610. The icc shown is the intra-interviewer correlation coefficient.

Whether this model is reasonable is not the topic being discussed here. We will assume that an interest exists in estimating the model among users of a survey and show how the effect of different survey error components on the parameters of interest can be estimated using the theory outlined in the previous sections.

The variables shown in the model are not observed variables, but rather they are constructs defined as influencing the answers to certain survey questions. For each of these constructs we can obtain at least two measures from the European Social Survey (ESS) round 4, conducted in 2008 (Jowell et al., 2007). As an illustration we will select one country only, Denmark, because it had a simple random sampling design, simplifying the discussion needed below. We stress, however, that our methods can be equally easily applied to designs with unequal inclusion probabilities.

The variables used to measure each of the four constructs can be found in the appendix. An estimate of the composite scores was constructed by taking the simple sum of indicators. Table 3.2 shows the histograms and summary statistics for the resulting sum scores.

In order to estimate the reliability of each construct and the associated error variance, we estimated a four-factor model using the software EQS 6.1 (Bentler, 1995). We obtained the standard errors of the reliability and error variance through a non-parametric bootstrap [3]. The resulting reliability and error variance estimates are shown in the last two columns of table 3.2.

The data collection mode was computer-assisted personal interviewing in the home of the respondent. Each interviewer conducted at least 4 and at most 48 interviews. Below we will investigate the effects on the variance of the estimates of the model due to correlation between the answers of different respondents interviewed by the same interviewer. The sixth column of table 3.2 shows the univariate intra-interviewer correlation coefficients[4] It

[3]The procedure we used is similar to that of Raykov (2009), except that in contrast with the approach discussed there, we also allow for randomness in the loadings in estimating the error variance and reliability.

[4]The intra-interviewer correlation coefficient was estimated using `R 2.11.0` by fitting a multilevel linear

can be seen that considerable intra-interviewer correlations exist. The average number of interviews per interviewer was 20; the distribution of the number of interviews per interviewer ranged between 4 and 48 and is shown in the rightmost histogram.

Using the statistical software `R 2.11.0` (R Development Core Team, 2010) and the package `OpenMx` (Boker et al., 2010), we estimated the model[5] shown in figure 3.1. The parameter estimates are shown in table 3.4.

The parameters of most interest to substantive researchers are the direct effects of social trust on political trust and vice versa, as well as the so-called 'total effects'. The direct effect of social trust on political trust is stronger than the converse effect. The total effects of social trust and political trust can be calculated as 1.0 and 0.40, respectively. This implies that for a given amount of change in social trust, political trust, which was measured on the same scale, is expected to increase by the same amount. The reverse is not the case, as social trust can be expected to increase by only 40% of the change in political trust. This is largely in correspondence with the literature on social capital (e.g. Putnam, 2001).

We do not comment on whether the model discussed is correct. This must be assessed by thorough investigation into a combination of appropriate theoretical considerations and model fit to observed data, which is outside the scope of this paper. We only show how the contribution of different survey error components to the variance of the coefficients can be estimated, also for models involving such complexities as instrumental variables, reciprocal effects, and a correlated error term.

3.4 Estimation of sampling and non-sampling errors in SEM

In the previous section we presented the structural equation modeling framework as well as an parameter estimates of a model found in the literature using real data. We will now discuss how different survey error components affect the variance of those estimates, and proceed to separately estimate their effects in the analysis presented.

3.4.1 Complex sampling and (interviewer) clustering

Muthén and Satorra (1995) provided an in-depth discussion of the estimation of structural equation models under complex sampling. Our discussion of this topic largely follows their results.

Unequal selection probabilities necessitate the estimation of the covariance matrix by a weighted estimator. We will denote this estimator as $S^{(c)}$. The variance of the parameter estimates can be obtained by the normal variance estimator if an adjustment is made to the estimated Γ matrix of fourth-order moments. The effect of unequal sampling weights on the variance, operating through the Γ matrix, is therefore in general multiplicative. Indeed, all survey error components that affect the fourth-order moments have a multiplicative effect on the variance of the estimates.

Clustering does not affect the estimator needed for the estimation of Σ. The variance of the estimates is affected, however. Muthén and Satorra (1995) discuss two separate ways

model with a random interviewer intercept to each variable. The icc was then estimated as the square root of the ratio of the random intercept variance to the residual variance.

[5] In formulating this model it was simpler to parameterize the error covariance as an effect of a latent variable with fixed loadings. This model is mathematically equivalent to the model shown in the figure but implies that the error variance parameters ϕ of social and political trust equal $\phi_i - \phi_{ij}$.

to take clustering into account: one design-based and one model-based. The model-based method proceeds by specifying a random effects (multilevel) model with the clusters (for example, PSU's or interviewers) as second-level units. Their design-based solution provides an adjustment to the Γ matrix that takes both clustering and stratification into account. Their method also allows for complex sampling designs on levels lower than the PSU through aggregation to the level of the PSU's.

In the discussion that follows we will adopt the design-based approach to variance estimation in the presence of clustering. The resulting variance estimator can be seen as obtained through the Taylor linearization method (Muthén & Satorra, 1995, 284).

3.4.2 Non-normality

The asymptotic distribution of estimators of the mean of a variable are free of the distribution of that variable (e.g. Neyman, 1934). Such is not the case, however, for regression coefficients and other parameters of multivariate models. The variance and the form of the distribution of these parameters depends on the fourth-order moments of the observed variables (Satorra, 1989; Muthén & Satorra, 1995).

It has been shown that even under non-normality, normal-theory maximum likelihood estimates of structural equation models are still consistent (e.g. Satorra, 1989). Indeed, this result holds for the entire family of minimum distance-estimators discussed in the preceding section. Thus, non-normality does not cause asymptotic bias in the estimates.

When the y vector does not follow a multivariate normal distribution, this does affect the variance of structural equation model parameters. The Γ^* matrix of equation 3.12, which is a function purely of the observed variances and covariances, no longer provides a consistent estimate of the matrix of fourth-order moments. In this case the general Γ matrix of equation 3.10 must be used; or, in the case of complex samples, $\Gamma^{(c)}$ of equation 3.11. It should be noted that the default behavior of all commonly used structural equation modeling software is to provide the variance estimators assuming normality.

Thus, the effect of non-normality is in general to change the matrix used as an estimate of Γ used in equation 3.9. Therefore non-normality, similarly to complex sampling, has a multiplicative effect on the variances and covariances of the parameter estimates.

3.4.3 Measurement error and its estimation

The effect that measurement error has on regression coefficients is well-known (Fuller, 1987; Biemer & Trewin, 1991). For general structural equation models the effect will depend on the structure of the model, which can be deduced from the matrix of first derivatives of the population covariances with respect to the parameters. For regression with a single predictor, the regression coefficient will be biased downwards. In multiple regression the bias is not necessarily downwards. Bias in multiple regression coefficients can be upwards or downwards, depending on the correlations between the predictors and the relative amount of measurement error in each of them.

If the measurement error was correctly estimated, either in the model itself or in an earlier analysis, the estimates are consistent even under non-normality, adding no asymptotic bias component to the total error. This is because the measurement error has already been corrected for. As will be shown, however, the correction does add variance.

Structural equation models can be used to simultaneously estimate measurement errors and correct for them. In fact, the distinguishing characteristic of structural equation models is that they are a marriage of psychometric (factor analysis) and econometric (regression) models (Jöreskog, 1978).

In practice, however, simultaneous estimation of measurement and 'structural' parameters may be impractical or impossible.

The inclusion of both a measurement and structural part in the model simultaneously may cause a model to become prohibitively large. In the example discussed above the model size would not be exceedingly large as the four constructs are estimated from 11 indicator variables. This is close to the minimum needed of 8 indicators. On the other side of the extreme, educational and psychological scales may have hundreds of indicators each.

Often, moreover, repeated measures or validation data are not available in the same study as that used for the estimation of the structural model, although the measurement properties of the variables used have been estimated in other studies. In such cases simultaneous estimation is impossible, and the structural model must be corrected for measurement error using the estimates from previous studies. Another approach is the prediction of measurement error from meta-analyses of reliability based on characteristics of the question (Oberski et al., 2004; Saris & Gallhofer, 2007a).

3.4.4 Nonresponse, coverage, and survey mode

To the extent that other error components such as nonresponse, coverage, and mode bias the covariance matrix, the parameters θ will be correspondingly biased.

The theoretical effect of nonresponse bias on the covariances was discussed by Groves and Couper (1998, chapter 2): it is a function of the nonresponse bias, nonresponse rate, and the difference in variances between respondents and nonrespondents. A similar result can be developed for coverage errors. The theory therefore suggests that nonresponse and coverage errors can in principle bias multivariate parameter estimates.

One can conclude from the results developed that for a bias to exist, there must be an interaction between variables correlated with nonresponse and the variables under study. Such might for example occur when, in a simple regression of social trust on fear of crime, there would be an interaction with living in a city or not, so that the relationship between fear of crime and trust were different for city dwellers. Since urbanicity is a commonly found correlate of nonreponse (Groves & Couper, 1998) this would imply a bias caused by nonresponse in the simple regression coefficient. Thus, for nonresponse bias to occur in a regression coefficient or other function of covariances mean, a 'third order' interaction must exist, whereas for nonresponse bias in means and totals a bivariate relationship or 'second order' interaction with participation correlates suffices.

The general focus in studies examining such effects in real surveys is on the estimation of means (De Leeuw & Van der Zouwen, 1988; Groves, 2002; Groves & Peytcheva, 2008). Due to this focus very few studies examine the extent of such biasing effects on multivariate statistics. An exception for nonresponse is Voogt (2004), who used official record data and found nonresponse bias in political variables' means to be high but did not find bias for logistic regression coefficients. Recently Révilla and Saris (frth), comparing a web survey with the face-to-face ESS, investigated possible mode effects and found no differences in the correlations between repeated measures and other variables.

	Distribution of y	Clustering / weighting	Measurement error	Γ	$\hat{\Sigma}_\eta$	Σ_Ψ
1	Normal	-	-	$\hat{\Gamma}^*$	S	0
2	Normal	-	Fixed	$\hat{\Gamma}^*$	$B^{-1}\Phi B^{-1}$	0
	Normal	-	Estimated	$\hat{\Gamma}^*$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$
	Normal	Yes	-	$\hat{\Gamma}^{(c)*}$	$S^{(c)}$	0
	Normal	Yes	Fixed	$\hat{\Gamma}^{(c)*}$	$B^{-1}\Phi B^{-1}$	0
	Normal	Yes	Estimated	$\hat{\Gamma}^{(c)*}$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$
	Non-normal	-	-	$\hat{\Gamma}$	S	0
3	Non-normal	-	Fixed	$\hat{\Gamma}$	$B^{-1}\Phi B^{-1}$	0
	Non-normal	-	Estimated	$\hat{\Gamma}$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$
	Non-normal	Yes	-	$\hat{\Gamma}^{(c)}$	$S^{(c)}$	0
4	Non-normal	Yes	Fixed	$\hat{\Gamma}^{(c)}$	$B^{-1}\Phi B^{-1}$	0
	Non-normal	Yes	Estimated	$\hat{\Gamma}^{(c)}$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$

Table 3.3: Different error components and their effect on the choice of estimators of the parameters θ and their variance. The effect of weighting is not shown separately because in our subsequent example simple random sampling was employed. However, the procedure for examining its effect is identical to that for clustering. The final variance is always obtained by equation 3.9.

The record of nonresponse, coverage, and mode effects on multivariate statistics is thus rather incomplete. Theoretically biases can exists, but more is required than for means and totals. Two studies could not find any biases, but a generalization cannot be made. It is worthy of note, however, that the study of such biases requires a special data collection design (e.g. Biemer, 2001), which is not available to us in the example we discuss. Therefore we are forced to ignore the possible biasing effects of nonresponse, coverage, and survey mode in our subsequent discussion.

3.4.5 Decomposition of the variance of multivariate statistics

The variance of the parameter vector θ was given in equation 3.9. Clustering, unequal sampling weights, and non-normality are all factors that affect the choice of the matrix Γ necessary for obtaining correct variance estimates. Sampling weights and measurement error also affect the necessary choice of a covariance matrix estimator.

Each row of table 3.3 yields a different estimator of the variance (or standard errors) of the parameter estimates of the model. A simple model, based on the observation that the effects of the conditions are in general multiplicative, is then to assume that each of the variance vectors under the different conditions is the result of a multiplication of the effects of non-normality, clustering, measurement error and a general scaling constant of the variance.

$$\text{var}_{\text{condition}} = vNCM, \tag{3.16}$$

where N, C, and M are the multiplicative 'design' effects of non-normality, clustering, and measurement error, respectively.

These effects can be estimated by estimating the four numbered rows shown in table 3.3. We then report the square roots of the encountered effects (defts), since these are on

	Estimate	$\sigma_1(\hat{\theta})$	$\sigma_2(\hat{\theta})$	$\sigma_3(\hat{\theta})$	$\sigma_4(\hat{\theta})$	$\text{deft}_{\text{measerr}}$	$\text{deft}_{\text{non-normality}}$	$\text{deft}_{\text{clustering}}$
soctrust → poltrust	0.77	0.08	0.16	0.17	0.17	1.92	1.06	1.02
efficacy → poltrust	0.51	0.08	0.16	0.18	0.23	1.94	1.11	1.30
poltrust → soctrust	0.30	0.11	0.19	0.21	0.25	1.77	1.10	1.17
fearcrim → soctrust	-0.68	0.12	0.22	0.25	0.25	1.88	1.10	1.03
ϕ(poltrust, soctrust)	6.54	1.60	3.17	3.52	4.20	1.98	1.11	1.19
ϕ(efficacy)	1.64	0.06	0.10	0.10	0.11	1.71	0.99	1.11
ϕ(fearcrime)	1.71	0.06	0.11	0.12	0.14	1.78	1.07	1.16
cov(efficacy, fearcrime)	-0.60	0.06	0.10	0.08	0.09	1.73	0.77	1.10

Table 3.4: Parameter estimates, standard error estimates under various conditions, and square root design effects (deft) for the example analysis.

the scale of the parameters rather than their square.

We now apply this decomposition to the example multivariate analysis discussed earlier, reporting standard errors under the various conditions as well as the square root of the interviewer clustering, measurement error, and non-normality effects.

3.4.6 Estimation of error components in the example

The analysis of the structural equation model of political and social trust presented earlier can be used to show how the effects of different survey error components can be estimated.

Table 3.4 shows the standard error estimates under different conditions. Four conditions are shown, corresponding to the numbered rows of table 3.3. From these standard error estimates the 'design effects' of measurement error, non-normality, and clustering can be estimated[6]. The square roots of these design effects (defts) are shown in the last three columns. These show the percentage increase in the standard error due to each factor. It can be seen that, in general, measurement error is the primary concern for the variance of the estimates, as it almost doubles the standard errors of the structural parameters of the model. The smallest measurement error deft is a 71% increase in the standard error for the variance of the independent variable efficacy.

The effect of interviewer clustering is less than the effect of measurement error but also considerable, with defts ranging between 1.03 for the standard error of the effect of 'fear of crime' on 'social trust' and 1.30 for the effect of 'political efficacy' on 'political trust'. The relative sizes of the defts of clustering may appear surprising considering equation 3.2; since the icc of social trust is rather large (0.25), in a simple regression we would expect the effect of social trust as a predictor to have the largest design effect. However, there are two important differences between the model of Scott and Holt (1982) leading to equation 3.2 and our model: the cluster sizes are not equal for all interviewers as shown in table 3.2, and we are not dealing with simple regression but with a complex structural equation

[6]We have designated the within-interviewer clustering effect an 'interviewer effect'. However, because the sample was not interpenetrated, some correlation between interviewer and region may exist.

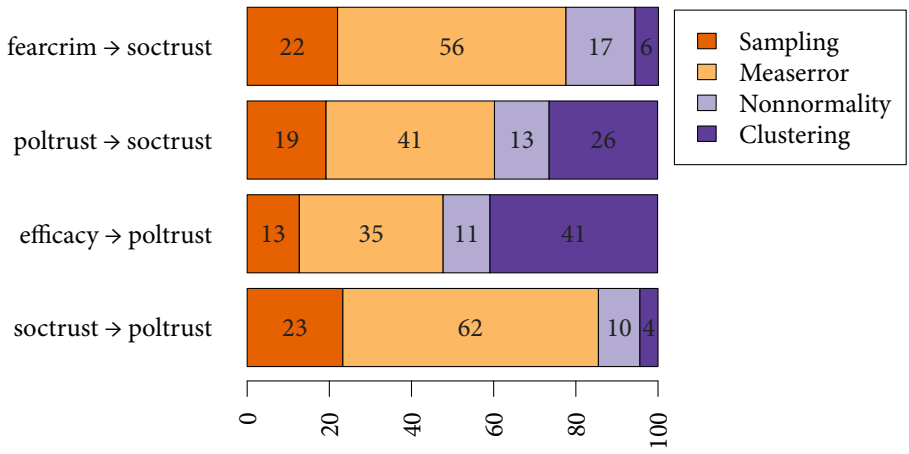


Figure 3.2: Percentage of variance of the regression coefficients of the model (see fig. 3.1) contributed by each error component.

model. This shows that approximations such as equation 3.2 cannot always be used in more complex situations.

Finally, it can be seen in table 3.4 that non-normality in general increases the variance of the estimates. This is not always the case in our sample, however, as the 'robustness effect' for the covariance between efficacy and fear of crime is smaller than unity. This most likely reflects the fact that the Γ matrix used is not the true population matrix, but a sample-based estimate. It may also reflect the fact that the second order sample moments may be negatively correlated. In general percentage increase in the standard errors due to non-normality is modest compared with the effects of interviewer clustering, and even more so when compared with the effects of measurement error.

The previous sections showed that the effects of measurement error, interviewer clustering, non-normality, and the sampling design in general are multiplicative in the variance of the estimates. For this reason the design effect or its square root is the more appropriate summary of their effects on the variance. However, for a given solution one can also calculate the percentage of variance due to each factor. Here a caveat should be added that such a measure is necessarily conditional on the sample, sample size, model, and parameter values encountered and cannot easily be generalized. However, it can be instructive for a given analysis to examine how much of the variance in the regression coefficients can be attributed to sampling, measurement error, non-normality, and interviewer clustering.

Figure 3.2 shows the amount of variance in the four structural (regression) coefficients of the model attributable to each of the four survey error sources studied here. Due to the simple random sampling scheme employed for the data collection, the factor "sampling" indicates purely the sample size. It can be seen that only a fifth of the total variance is attributable to sampling for all four parameters. It is also clear that measurement error is another important source of uncertainty about the parameter estimates. Interviewer clustering contributes in about equal parts with measurement error for two of the parameters, while not apparently playing any large role in the other two. The percentage of variance

	$\hat{\theta}_{\text{correct}}$	$\hat{\theta}_{\text{naive}}$	$\widehat{\text{bias}}^2$	$\sigma^2(\hat{\theta}_{\text{naive}})$	$\sqrt{\text{MSE}}$	% MSE	
						Bias	Var.
soctrust → poltrust	0.77	0.83	0.00	0.02	0.16	15%	85%
poltrust → soctrust	0.30	0.44	0.02	0.04	0.25	32%	68%
efficacy → poltrust	0.51	0.28	0.05	0.02	0.26	76%	23%
fearcrim → soctrust	-0.68	-0.31	0.13	0.01	0.38	89%	10%

Table 3.5: If the model is not corrected for measurement error the parameter estimates of interest will be biased. The Mean Square Error (MSE) of the naive (not corrected for measurement error) estimate then equals $\text{bias}^2 + \sigma^2(\theta_{\text{naive}})$. The last two columns show the approximate percentage of mean square error in the naive estimates due to the bias and the variance, respectively. (Percentages not adding up to 100 are due to rounding errors.)

due to non-normality is slightly less that that simply due to sampling.

From the above discussion it is clear that sampling is by no means the only or even the most important factor contributing to the inferential uncertainty about the parameters in this example. It also shows that the error sources are far from equal and show a different pattern from that typically found in the estimation of means and total. This finding cannot be generalized to other models, but does show that in an analysis of real data using a model found in the literature such differences can occur.

So far we have assumed that the practitioner obtains estimates of the measurement error variances and correct for them. If the correction is applied and assuming that the model is correct, the asymptotic bias is zero, even under non-normality and clustering. However, such corrections are not always applied. Therefore we will briefly show the effect that not correcting for measurement error has for this example.

Table 3.5 repeats, in the first column, the consistent estimates of the four regression coefficients while correcting for measurement error. The second column shows the estimates obtained by incorrectly assuming no measurement error. The square of the difference between these two is shown in the column labeled "$\widehat{\text{bias}}^2$". As discussed above, the mean square error of the naive estimates will equal the sum of the bias squared and the variance of the estimates. The last column shows the amount each of bias and variance contribute to the mean square error.

Without correction for measurement error the estimates of the model are biased. Therefore the root mean square error shown in the column $\sqrt{MSE}$ in table 3.5 is a function of both this bias and the variance of the estimate. This root mean square error without correction for measurement error can be compared with the column labelled σ_4 in table 3.4. This is so because the model in that table has been estimated with correction for measurement error so that the estimates are unbiased (asymptotically and under the null hypothesis). In that case the root mean square error of an estimate will equal the right-most standard error shown in table 3.4.

When thus comparing the root mean square errors with and without correction for measurement error, it is clear that the MSE of this model without correction is larger for all parameters than the MSE of the corrected model estimates. The bias caused by ignoring measurement error causes the MSE to exceed that of the unbiased estimates in this example.

Both from the point of view of obtaining unbiased estimates and that of minimizing the mean square error it is therefore necessary to correct for measurement error.

3.5 Discussion and conclusion

This paper studied the effects of total survey error sources on multivariate statistics.

The first sections developed the structural equation modeling framework, which allows for the formulation of a wide range of common and specialized multivariate models; multiple regression, factor analysis, instrumental variables, multilevel models, and longitudinal models are among some of the wide variety of possible models that can be formulated within this framework. An example analysis with a model involving reciprocal effects, correlated errors, and measurement error demonstrated the use of structural equation models.

A review of existing studies showed that both theoretical and empirical considerations suggest such effects may differ greatly in relative size from their effects on the more commonly discussed estimation of means and totals.

It was discussed how each of the error sources (interviewer) clustering, unequal probability sampling, non-normality, and measurement error can be taken into account in structural equation models. It was shown exactly how each source influences the variance of the estimates of such models.

The relative effects of each error source on the estimates of our example analysis were then shown in terms of (root) 'design effects' (defts) and percentages. It was clear from this exercise that in the example given, estimated on real data from the 2008 European Social Survey, the effects of measurement error were the most pronounced, leading almost to a doubling of the standard errors relative to the variance the estimates would have had if there had been no measurement error. Clustering was another important factor, with non-normality leading to relatively smaller differences.

This result might lead one to think that it may not be worthwhile to correct for measurement error. However, the bias introduced by assuming no measurement error in the same analysis caused the mean square error to exceed that of the corrected model for all estimates. Therefore even if the goal is to obtain estimates that have the smallest mean square error, but that are not necessarily unbiased, the choice of preference should be the measurement error-corrected estimator.

One limitation of our example is that we have spoken of interviewer effects, while the interviewers were not randomly assigned to respondents (interpenetrated). Therefore it is possible that the clustering effects found were not (solely) due to the interviewer, but, for example, due to region. For a design allowing for the separation of such effects, see Bassi and Fabbris (1997). Another limitation is that in our example it was not possible to simultaneously estimate the effects of nonresponse, noncoverage/overcoverage, and survey mode. Again, a special study design is necessary for the study of such effects. The procedure presented in this paper, however, can easily be extended to encompass such effects.

A more fundamental caveat should be added about the use of the term "design effect". This term usually is taken to mean the variance of an estimator under the sampling scheme used, relative to the variance under simple random sampling. For measurement error, clustering, and non-normality, we have employed the same term, but the comparison is not with simple random sampling *per se* but with a design without measurement error, clustering, or non-normality. Thus, there is a strong analogy with the pure "design effect" but the two measures are not exactly the same.

We hope to have shown that it is possible to simultaneously estimate the effect on multivariate statistics of different survey error sources. We have given one example analysis.

The relative importance of different survey error sources in general, across different studies, remains a topic of considerable interest. The methods discussed in this paper could be applied to enable such a study.

Appendix: Questions used in the example analysis

Here we show the questions used to measure the four constructs analyzed above. All questions come from the European Social Survey Round 4 and were translated into each country's respective language (in our case Danish). The first item of political efficacy and the last two items of fear of crime were reverse-coded so higher scores indicated more efficacy or more fear, respectively.

3.5.1 Social Trust

```
Using this card, generally speaking, would you say that most people can be
trusted, or that you can't be too careful in dealing with people? Please tell
me on a score of 0 to 10, where 0 means you can't be too careful and 10 means
that most people can be trusted.

                                                          Most
  You can't                                              people
   be too                                                can be
   careful                                               trusted

   00    01   02   03   04   05   06    07  08   09     10

 Using this card, do you think that most people would try to take advantage of
   you if they got the chance, or would they try to be fair?

    Most people                                      Most people
    would try to                                     would try to
  take advantage                                       be fair
       of me

   00    01   02   03   04   05   06    07  08   09   10

Would you say that most of the time people try to be helpful or
   that they are mostly looking out for themselves? Please use this card.

    People                                            People
  mostly look                                        mostly try
    out for                                           to be
  themselves                                         helpful

   00   01   02   03   04   05   06    07  08   09   10
```

3.5.2 Political efficacy

```
How often does politics seem so complicated
   that you can't really understand what is going on?
   Please use this card.

                          Never        1
                        Seldom         2
                    Occasionally       3
                       Regularly       4
                     Frequently        5
```

```
B3 CARD 7 How difficult or easy do you find it to make
  your mind up about political issues? Please use this card.

                    Very difficult      1
                         Difficult      2
     Neither difficult nor easy         3
                              Easy      4
                      Very easy         5
```

3.5.3 Political trust

```
Using this card, please tell me on a score of 0-10 how much you
personally trust each of the institutions I read out. 0 means you do not trust
an institution at all, and 10 means you have complete trust. Firstly...


                        No trust                                    Complete
                         at all                                       trust

  B4     ...[country]'s   00   01   02   03   04  05   06  07  08  09  10
       parliament?
  B5     ...the legal     00   01   02   03   04  05   06  07  08  09  10
       system?
  B6     ...the           00   01   02   03   04  05   06  07  08  09  10
       police?
```

3.5.4 Fear of crime

```
How safe do you – or would you - feel walking alone
  in this area after dark? Do – or would – you feel... READ OUT...

         ...very safe,    1
                safe,     2
              unsafe,     3
    or, very unsafe?      4

How often, if at all, do you worry about your home being burgled?
        Please choose your answer from this card.
       All or most of the time  1
       Some of the time         2
       Just occasionally        3
       Never                    4

How often, if at all, do you worry about becoming a victim of
violent crime? Please choose your answer from this card.

All or most of the time       1
     Some of the time         2
     Just occasionally        3
               Never          4
```

Appendix: EQS input for reliability analysis

```
/TITLE
    Factor analysis of construct indicators ESS round 4 Denmark
```

```
/SPECIFICATIONS
    DATA='denmark.ess';
    VARIABLES=46; CASES=1610; GROUPS=1;
    METHOD=ML; ANALYSIS=COV; MATRIX=RAW;

/LABELS
...

/EQUATIONS
    V2 =   1F1 + E2;
    V3 =   *F1 + E3;
    V4 =   *F1 + E4;
    V6 =   1F2 + E6;
    V7 =   *F2 + E7;
    V8 =   1F3 + E8;
    V9 =   *F3 + E9;
    V10 =   *F3 + E10;
    V17 =   1F4 + E17;
    V18 =   *F4 + E18;
    V19 =   *F4 + E19;

    F5 = V2 + V3 + V4;
    F6 = V6 + V7;
    F7 = V8 + V9 + V10;
    F8 = V17 + V18 + V19;

/VARIANCES
     F1 = *;
     F2 = *;
     F3 = *;
     F4 = *;
     E2 = *;
     E3 = *;
     E4 = *;
     E6 = *;
     E7 = *;
     E8 = *;
     E9 = *;
     E10 = *;
     E17 = *;
     E18 = *;
     E19 = *;

/COVARIANCES
    F1,F2 = *;
    F1,F3 = *;
    F2,F3 = *;
    F1,F4 = *;
    F2,F4 = *;
    F3,F4 = *;
/PRINT
    TABLE=EQUATION;
    COVARIANCE=YES;
    CORRELATION=YES;

/SIMULATION
  bootstrap = 1610;
  replication = 2000;
  seed = 123456789;
```

```
/OUTPUT
  parameters;
  standard deviation;

/END
```

All replications converged. The output from this analysis was read in using R 2.11.0, and the reliability calculated as

$$\frac{(\sum \lambda)^2 \text{var}(\text{F}_i)}{\text{var}(\text{F}_{i+4})}; i = 1, 2, 3, 4,$$

in each of the 2000 replications. The averages and standard deviations across replications are shown in table 3.2.

Chapter 4

Measurement error models with uncertainty about the error variance

Abstract

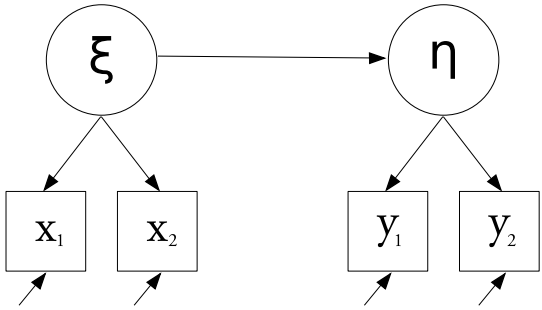
Measurement error biases regression estimates, making it necessary to correct for measurement error whenever it is present. Structural equation modeling has been developed to deal with this problem, allowing for the estimation of a very wide class of models with correction for measurement error.

Often the amount of measurement error cannot be estimated directly but must be obtained from external analyses. In this situation one can still correct for measurement error by fixing error variance parameters of the model to the estimated values. So far the only recourse has been to assume these values are perfect estimates of the true amount of measurement error present in the observed variables. Usually, however, this assumption is false since the fixed values represent estimates based on an external study.

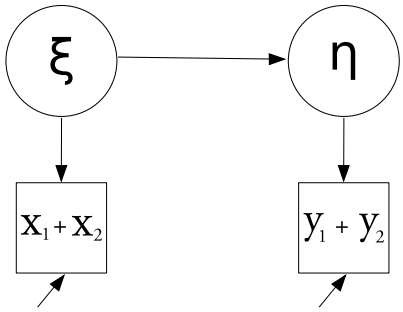
In this paper we show that this procedure can cause the standard errors of the model to be smaller than they should be, causing inference to be incorrect. Even though the parameter estimates are still consistent, confidence intervals based on standard program output will be too small.

We also provide a solution to this problem for general structural equation models. This solution comes in the form of an explicit analytical expression that should be added to the standard errors found in the standard program output.

The implications of the results are discussed, while a Monte Carlo study confirms their validity.



(a) Simple regression with multiple indicators.



(b) Regression with single indicators.

Figure 4.1: Two different ways of correcting for measurement error in a simple regression using SEM

Introduction

Measurement error in variables is a serious problem for the estimation of regression models (e.g. Fuller, 1987; Bollen, 1989). Error in dependent variables will bias R^2 and standardized regression coefficients, while error in independent variables will, in addition, bias unstandardized regression estimates. Ignoring such errors can severely affect estimates and conclusions. An important issue in the analysis of (simultaneous) regression equations is therefore how to correct for measurement error.

Structural equation modeling (SEM) was developed precisely to deal with this issue (Bollen, 1989). SEM allows the researcher to estimate simultaneous regression equations while also correcting for measurement error. The most direct way of correcting is by simultaneously specifying the "measurement part" and the "structural part" of the model, using multiple indicators of the latent variables whose relationships are of interest.

Figure 4.1(a) shows the simplest possible example of such a model: the latent independent and dependent variables of interest are respectively denoted ξ and η and their indicators x_1, x_2, y_1, and y_2. The arrows at the bottom signify the influence of measurement error. Estimating this model the researcher obtains estimates of the variance of the measurement error in the indicators, as well as the "structural" relationship between ξ and η corrected for measurement error. Thus the correction for measurement error is subsumed in the model.

A second possibility, recommended in various textbooks on SEM (Bollen, 1989; Hay-

duk, 1987; Schumacker & Lomax, 2004), can be characterized as the "two-step" procedure. The latent variable ξ is defined as having the sum $x_1 + x_2$ as a single indicator, while η has the sum $y_1 + y_2$ as a single indicator. Means or weighted sums are also used as observed composite scores. The resulting model, shown in figure 4.1(b), is not identified. The variance of the measurement error in the composite scores must be calculated separately based on the error variances of the indicators (Saris & Gallhofer, 2007a), or obtained from other sources such as published reliability studies. Step one is then to fix the error variance parameters for the two composite scores to these estimates. In the second step the structural model is estimated while correcting for these fixed measurement error variances.

The two-step solution to correction for measurement error has at least two advantages. First, it can dramatically reduce the size of the model and the number of possible parameters the researcher has to deal with. Since at least two indicators are usually needed for each latent variable to subsume the measurement error in the model, the number of variables is at reduced by at least one half.

A second advantage of the two-step method is a separation of labor between reliability studies and more substantive research. To estimate measurement error variances one requires specific research designs that are adequate for this purpose. Often the models that must be applied to these designs are rather complex (Alwin, 2007; Saris & Gallhofer, 2007a). To subsume the measurement error into the analysis, it is therefore necessary that the research design both addresses the substantive question and allows for the estimation of measurement error. It is also required that the researcher has both the substantive knowledge to formulate the correct structural model, and is in addition well-versed in the analysis of measurement error models. The two-step solution allows substantive researchers to concentrate on the substantive model while still correcting for measurement error.

Examples of studies employing the two-step approach abound in the literature. Some examples from different fields of application are Beckie and Hayduk (1997, 30), Varki and Colgate (2001, 236), Small et al. (2003, 170), and Rhodes et al. (2006, 3150). Each of these studies employs an estimate of the error variance from a previous study or separate analysis, and fixes the corresponding parameters to these estimates. The model of interest is then estimated keeping these values fixed.

The two-step solution has advantages, but also introduces a problem: the error variance parameters are fixed to certain values, but these values are themselves estimates from previous studies. The uncertainty about these estimates is not taken into account in the second step, where they are held fixed. Because the variance of the fixed parameters is not taken into account, the standard errors of the "structural" parameters will be too low.

All of the studies using this method published so far suffer from the problem of downwards biased standard errors. This implies that up until now, the only method to correct for measurement error while performing correct inferences is to subsume the measurement error directly into the model as shown in figure 4.1(a).

The purpose of this paper is therefore to provide a solution to the problem of downwards biased standard errors in the two-step method of correction for measurement error. The solution takes the form of a simple correction formula, which is straightforward to implement in standard SEM software. We show that when our solution is applied, the standard errors allow for correct inferences in the face of uncertainty about the fixed parameters. This paves the way for inferentially correct applications of the two-step method of correction for measurement error.

The first two sections discuss the general problem of measurement error in the estimation of structural equation models. It is shown how such uncertainty in the reliability can impact the standard errors of estimates of a correlation coefficient. A correction to the standard errors in the presence of uncertainty about the amount of measurement error is therefore needed. An example of SEM analysis with correction for measurement error is given in the form of a relatively complex model which cannot be formulated as a multiple regression model. This analysis does not take into account the uncertainty about fixed measurement error variance parameters.

Section three proposes a general solution to this problem in the context of structural equation models. First it is shown that the variance of parameters in two-step models equals the sum of a "standard" variance and an "extra" variance due to the uncertainty about the fixed parameters. This shows that under certain conditions, the current standard SEM program output will underestimate standard errors and confidence intervals. We then present our correction to the standard errors, taking into account the uncertainty about fixed measurement error variance parameters. An explicit analytical formula for the correction in the context of SEM is provided. Given the parameter estimates of the model, this correction to standard errors can be readily applied.

Section four elaborates our correction by application to a multiple regression model with two uncorrelated regressors. The effect of uncertainty about fixed parameters on the standard errors of the regression is shown. It is also shown how the form of the model determines the size and presence of the effect.

If our assertions are correct, models with uncertainty about fixed measurement error variance parameters should underestimate standard errors without our correction, while applying the correction should bring confidence intervals in line with the nominal coverage rate. The fifth section shows that this is indeed the case in a Monte Carlo study that varies the standard errors of the measurement error variance parameters over a wide range.

The final section provides a conclusion, and discusses some limitations and needs for further investigation.

4.1 The problem of uncertainty about the reliability estimates

Uncertainty about the reliability is a potentially important source of variability in estimates corrected for measurement error using the two-step method. A short simulation shows why this should be the case, and why the uncertainty about measurement error should be taken into account.

One of the simplest and most well-known procedures of correction for measurement error is the classical correction for attenuation of the correlation coefficient (e.g. Fuller, 1987):

$$\rho(\eta, \xi) = \frac{\rho(y, x)}{\kappa_1 \kappa_2}, \tag{4.1}$$

where y and x are observed indicators of respectively η and ξ. This equation states that the correlation between the variables of interest η and ξ equals the correlation between the observed variables y and x divided by the product of their reliability "coefficients" (Saris & Gallhofer, 2007a) or "ratios" (Fuller, 1987). This holds true for a population of subjects; when a sample is obtained, the sample correlation coefficient serves as an estimate of the

population correlation between the observed variables:

$$\hat{\rho}(\eta, \xi) = \frac{\hat{\rho}(y, x)}{\kappa_1 \kappa_2}. \tag{4.2}$$

The only difference between this equation and the previous one is that the population correlation coefficient was replaced by its sample counterpart. The estimate of the corrected correlation between η and ξ depends on the sample estimate of the correlation between y and x. It is therefore easy to see that as the sampling variation of $\hat{\rho}(y, x)$ increases, so will the variation in the estimate $\hat{\rho}(\eta, \xi)$.

However, in equation 4.2 the reliability coefficients κ_1 and κ_2 were assumed perfectly known and equal to their true population values. If they are not exactly known but estimated in a separate analysis, the estimates $\hat{\kappa}_1$ and $\hat{\kappa}_2$ replace their population counterparts in equation 4.2. It can again be seen, then, that the variation in the estimate $\hat{\rho}(\eta, \xi)$ increases with increased variability in the reliability estimates $\hat{\kappa}_1$ and $\hat{\kappa}_2$.

We show the effect that variability in the reliability estimates has on the estimate of the corrected correlation coefficient $\hat{\rho}(\eta, \xi)$ by constructing a simulation. In each simulation the correlation between the observed variables $\rho(y, x)$ is perfectly known and held constant at the true population value, but a random draw is taken from the sampling distribution of the reliability estimates $\hat{\kappa}_1$ and $\hat{\kappa}_2$. The resulting variation in the corrected correlation is therefore exclusively due to variation in the reliability estimates: the effect of uncertainty about the reliability estimates is shown in the limit of an infinite sample used to estimate the correlation between the observed variables.

In the simulations the correlation between the observed variables $\rho(y, x)$ was chosen to equal 0.40. The reliabilities κ_1^2 and κ_2^2 in the population were chosen as $\kappa_1^2 = 0.75$ and $\kappa_2^2 = 0.65$. These choices imply that in the population the true correlation between the two constructs of interest $\rho(\eta, \xi) \approx 0.573$. We assume the sampling distributions of the reliability estimates are normal with a certain standard error, and drew 500 samples from this distribution of $\hat{\kappa_1}^2$ and $\hat{\kappa_2}^2$. On average the estimates of $\hat{\kappa}_1^2$ and $\hat{\kappa}_2^2$ equaled their population values 0.75 and 0.65, but in any given sample the estimate differed from the population value due to variation in the reliabilities. For each sample drawn, we then compute the corrected correlation coefficient $\hat{\rho}(\eta, \xi) = 0.40/\left(\hat{\kappa}_1 \hat{\kappa}_2\right)$, obtaining 500 estimates of the correlation between the latent variables.

This process was repeated using different standard errors for the reliabilities. Increasing amounts of uncertainty about the reliability were used by setting the standard errors of the reliabilities to 0.001, 0.01, 0.05, and 0.1. Figure 4.2 shows the boxplots of 500 draws from four distributions of the corrected correlation for these different standard errors of the estimates of the reliability.

The dotted line in figure 4.2 shows that the mean of all of these distributions equals the true corrected correlation ~ 0.57. It can be seen that all of the estimates are far away from the population correlation between the observed variables of 0.4; if the correction had not been made then a very precise but biased estimate would have been obtained.

When the standard errors of the estimated reliability are very small (.001), almost all of the corrected correlations are very close to the true correlation .57. Small standard errors of .01 already cause a greater variability in these corrected values. For medium and larger standard errors – of .05 and .10 – the corrected correlations vary considerably. The boxplot shows that corrected correlations varied between .42 and .87 when the standard error was .10. Of the corrected correlations 95% lay between .49 and .70.

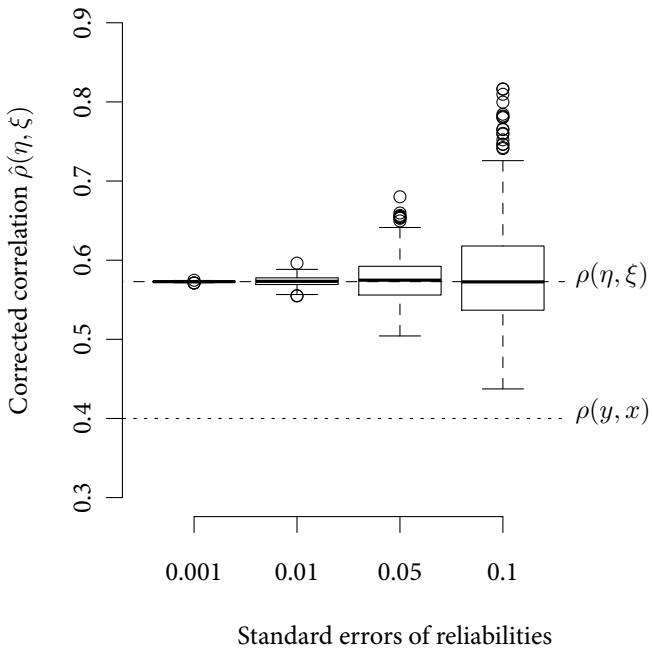


Figure 4.2: Estimates of the true correlation of interest can vary widely when variability in the reliability is medium or large. Each box represents 500 draws from the distributions of the two reliabilities for increasing standard errors. The sample size used to estimate the correlation between the observed variables is assumed to be large, so that the boxplots represent only variation due to the uncertainty about the reliabilities.

This short simulation study shows that treating the reliability as a known constant is only warranted when it has been precisely estimated with a very small standard error. In all other cases, the estimation of statistics which take measurement error into account can be affected quite a bit by uncertainty in the measurement error variance or reliability. Therefore the assumption of zero uncertainty about fixed measurement error variances can have negative consequences for inference.

4.2 Measurement error in structural equation models: an example

The previous section showed that uncertainty about the reliabilities can be a problem in the two-step method of correction for attenuation of correlations and simple regression models. The same problem occurs for more complex structural equation models.

This section provides an example analysis of a structural equation model with correction for measurement error. The model chosen contains reciprocal effects and correlated errors, making it impossible to formulate as a multiple regression. After formulating the structural part of the model, the two-step procedure to correction for measurement error

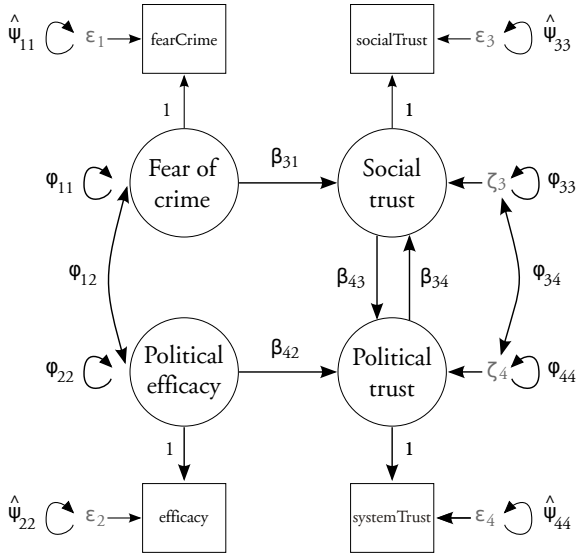


Figure 4.3: Structural equation model adapted from Saris & Gallhofer (2007a). Variance parameters are shown as circular arrows.

is applied using previously obtained estimates of the variances of the error variables.

The relationship between political and social trust is a central point of interest in the study of social capital. While Putnam (2001) has argued that social trust engenders good government and subsequently trust in politics, others have argued that political trust may in turn affect social trust: the correlation between social and political trust may be due to effects in both directions. Saris and Gallhofer (2007a) discuss an analysis of the reciprocal effect between social and political trust with correction for measurement error. A simplified version of their model is shown in figure 4.3.

The model shown in figure 4.3 contains a reciprocal effect between the variables "social trust" and "political trust", as well as a covariance between the disturbance terms of these two variables. The exogenous variable "fear of crime" affects only "social trust" and not "political trust", while "political efficacy" affects only "political trust" and not "social trust". These restrictions are enough to identify the reciprocal effect between "social trust" and "political trust", as well as the disturbance term covariance.

Whether this model is correct must be studied by careful examination of the underlying theory and the model fit to actual data, which is not the topic of the present example. Here we will discuss only how measurement error can affect the analysis of such a complex model.

All four variables in the model are complex concepts, measured as simple sum scores of several survey questions[1]. Data from the European Social Survey round 4 (2008) in Denmark are used. The sample size was 1610 Danish residents, surveyed by computer-assisted personal interviewing (CAPI) in their home. Table 4.1 shows the summary statistics for the resulting sum scores. Also shown is the estimated reliability for each sum score, and the

[1]For the full questionnaire we refer to http://ess.nsd.uib.no/ess/round4/fieldwork.html

Variable	Scale	Mean	Std dev	$\hat{\text{Rel.}}$ (s.e.)	$\hat{\psi}$ (s.e.)
fearCrime	3–12	5	1.7	0.57 (0.02)	1.3 (0.04)
efficacy	2–10	7	1.7	0.64 (0.03)	1.2 (0.07)
socialTrust	0–30	20	4.7	0.73 (0.01)	6.0 (0.22)
systemTrust	0–30	21	4.8	0.77 (0.01)	6.3 (0.24)

Table 4.1: Summary statistics for the Denmark dataset.

	Uncorrected			Corrected		
	Est.	s.e.	t	Est.	s.e.	t
β_{43}	0.86	(0.15)	5.9	0.77	(0.16)	4.9
β_{34}	0.45	(0.15)	3.0	0.30	(0.19)	1.6
β_{42}	0.27	(0.09)	3.1	0.51	(0.16)	3.2
β_{31}	-0.30	(0.10)	-3.1	-0.68	(0.22)	-3.0

Table 4.2: Unstandardized estimates without and with correction for measurement error. Only the regression coefficients in the model and not the variance parameters are shown for the sake of brevity.

corresponding estimate of the error variance, with standard errors[2]. It can immediately be seen that the reliabilities and error variances are not perfectly estimated but contain some uncertainty.

The model shown in figure 4.3 can easily be specified in standard structural equation modeling software[3]. The latent variables "fear of crime", "political efficacy", "social trust", and "political trust" are each specified to have a composite (sum) score as a single indicator. The error variance ψ_{ii} of each single indicator is fixed to the corresponding value found in the last column of table 4.1. The relationships between the latent variables are then estimated correcting for measurement error in the composite scores.

Using this procedure we can obtain estimates of the "structural" parameters in the model corrected for measurement error. If the model is specified without latent variables or by fixing the error variances to zero, the "naive" parameter estimates without correction for measurement error are obtained. Table 4.2 shows both the naive and measurement error-corrected unstandardized estimates with standard errors and t-values.

Table 4.2 shows that in complex models such as the one analyzed, measurement error biases the estimates. This bias is not necessarily downwards. The corrected effects of the instruments efficacy and fear of crime are indeed increased after correction for measurement error, but the reciprocal effects between social and political trust are lower than without correction for measurement error. The standard errors for the corrected coefficients are larger, but those of the effects of efficacy and fear of crime are much more increased than the standard errors of the reciprocal effects of social and political trust.

The parameters of most interest to substantive researchers are the direct effects of social

[2]The error variances and reliabilities were obtained by first fitting a confirmatory factor model to the indicators of these constructs. The error variance and reliability of the simple sum score was then obtained by adding "ghost variables" to the model. The standard errors of these quantities were obtained by bootstrapping (Raykov, 2009).

[3]We have used the `OpenMx` package in R (Boker et al., 2010; R Development Core Team, 2010), and LISREL (Jöreskog & Sörbom, 1996) to double-check the results.

trust on political trust and vice versa, as well as the so-called 'total effects'. The direct effect of social trust on political trust is stronger than the converse effect. The total effects of social trust and political trust can be calculated as 1.0 and 0.40, respectively. This implies that for a given amount of change in social trust, political trust, which was measured on the same scale, is expected to increase by the same amount. The effect is also statistically significant at the 0.05 level ($t = 2.7$). The reverse is not the case, as social trust can be expected to increase by only 40% of the change in political trust, but this effect is not statistically significant at the 0.05 level ($t = 1.2$). This is largely in correspondence with suggestions made in the literature on social capital (Newton, 2007).

If the same model were analyzed *without* correction for measurement error, the conclusions would be rather different. Without correction, both direct effects are significant at the .05 level. The direct effect of social trust on political trust and vice versa are 11% and 50% too large respectively, and the total effects are respectively 40% and 50% overestimated.

It is clear that in the example, correction for measurement error has effects that do not just bias the results in a "conservative" direction, but that affect the conclusions in an way that is unpredictable without knowledge of the variable's reliabilities. For this reason the estimation of reliability and correction for the errors is essential.

This example assumed that the error variances were known precisely, rather than estimated as was the reality. We will now show that this false assumption may cause inference to be affected, and provide a solution to this problem.

4.3 Correction of the standard errors for uncertainty about fixed error variances

Structural equation models are linear simultaneous equations, and therefore can be expressed as hypotheses about the population covariance matrix Σ of some vector of observed variables. A structural equation model can always be formulated by the equation

$$\Sigma = \Sigma(\theta), \tag{4.3}$$

where θ is a vector of model parameters (Bollen, 1989). Among these parameters are the structural regression coefficients and variance parameters, as well as the measurement parameters. We will denote the variance of the measurement error variables by the matrix Ψ.

Given an observed covariance matrix S and a model $\Sigma(\theta)$, the parameters θ of the model can be consistently estimated by minimizing the minimum-distance function

$$F = (s - \sigma(\theta))' V (s - \sigma(\theta)), \tag{4.4}$$

where V is a possibly stochastic weight matrix that converges in probability to a positive definite matrix. Here $s = \operatorname{vech} S$, the unique observed variances and covariances, and $\sigma(\theta) = \operatorname{vech} \Sigma(\theta)$, the unique model-implied variances and covariances. It can be shown that different choices for the matrix V yield different estimators, including maximum likelihood (Satorra, 1989).

If there is no uncertainty about Ψ, a standard formula for the variance of the parameter estimates applies (Satorra & Bentler, 1990, 239). We will denote this variance as

$\text{var}_{\text{standard}}(\hat{\theta})$. This is the standard error output given by SEM software. The precise form of this equation will depend on distributional assumptions as well as possible complex sampling and other issues, which are not the topic of this discussion. All of these expressions have in common, however, that it is assumed that any fixed measurement error variances are exactly known with no uncertainty.

When the measurement error variance is not known exactly, but only fixed to a consistent estimate, the estimation procedure still provides consistent estimates corrected for measurement error. As will now be proved, standard errors obtained from $\text{var}_{\text{standard}}(\hat{\theta})$ must be adjusted: the measurement error variances in Ψ affect the free parameters of the model but are themselves fixed in the analysis.

The error (co)variance matrix is denoted Ψ. Let its variance matrix due to a previous estimation be called Σ_Ψ. The problem we are now faced with is how to take this variance matrix of Ψ into account in standard error calculations after the model corrected for measurement error has been estimated.

Consider the vector of parameter estimates $\hat{\theta}$, which is a function of the data (the observed covariance matrix S, and of the now random matrix Ψ.

$$\hat{\theta} = \hat{\theta}(S, \psi), \tag{4.5}$$

where $\psi := \text{vech}\,\Psi$.

Conditioning on ψ the variance of the estimate $\hat{\theta}$ equals:

$$\text{var}(\hat{\theta}) = E_\psi[\text{var}(\hat{\theta}(S|\psi)] + \text{var}_\psi[E(\hat{\theta}(S|\psi)], \tag{4.6}$$

where the first term will be close to the "standard" variance formula,

$$E_\psi[\text{var}(\hat{\theta}(S|\psi)] \approx \text{var}[\hat{\theta}(S|\psi)] = \text{var}_{\text{standard}}(\hat{\theta}), \tag{4.7}$$

and, by a Taylor expansion, the second term equals

$$\text{var}_\psi[E(\hat{\theta}(S|\psi)] \approx \left(\frac{\partial\hat{\theta}}{\partial\psi'}\right) \text{var}(\psi) \left(\frac{\partial\hat{\theta}}{\partial\psi'}\right)', \tag{4.8}$$

which is clearly non-negative definite. Thus the correct asymptotic variance of the estimated free parameter vector $\hat{\theta}$ under the null hypothesis is a simple sum of two terms: the 'standard' variance and the added variance due to the estimation of the measurement error variance matrix Ψ:

$$\text{var}(\hat{\theta}) = \text{'Standard' variance} + \text{variance due to estimation of } \Psi. \tag{4.9}$$

This finding is highly important because it shows clearly that under two conditions the variance of the estimates is always increased by uncertainty about the fixed measurement error parameters ψ. The standard errors without correction will necessarily be biased downward under these conditions:

1. The variance of ψ is positive, i.e. there is some uncertainty about the measurement error variance, *and*

2. The parameter in question is related to the measurement error variance ψ. Parameters that are independent of ψ will be unaffected.

We will now derive an explicit formula for the variance of parameters of structural equation models with uncertainty about ψ. This provides a solution to the problem of downward biased standard errors when the measurement error variances ψ have been fixed.

Let $f(\hat{\theta}, \psi) := \partial F / \partial \theta$. Then $\hat{\theta}$ and ψ are related by the fact that $\hat{\theta}$ equals the solution to the equation $f(\hat{\theta}, \psi) = 0$. We can therefore invoke the implicit function theorem to find $\partial \hat{\theta} / \partial \psi'$:

$$\frac{\partial \hat{\theta}}{\partial \psi'} = -\left(\frac{\partial f}{\partial \hat{\theta}}\right)^{-1} \frac{\partial f}{\partial \psi} = -\left(\frac{\partial^2 F}{\partial \theta \partial \theta'}\right)^{-1} \left(\frac{\partial^2 F}{\partial \theta \partial \psi'}\right). \tag{4.10}$$

From Satorra (1989) and Neudecker and Satorra (1991) it can be shown that

$$-E\left(\frac{\partial^2 F}{\partial \theta \partial \theta'}\right) = \Delta_\theta' V \Delta_\theta \equiv J \tag{4.11}$$

and

$$E\left(\frac{\partial^2 F}{\partial \theta \partial \psi'}\right) = \Delta_\theta' V \Delta_\psi \equiv D. \tag{4.12}$$

We can conclude that the 'extra' variance due to the estimation of Ψ under the null hypothesis equals

$$\begin{aligned} \text{var}_\psi[E(\hat{\theta}(S|\psi)] &\approx \frac{\partial \hat{\theta}}{\partial \psi'} \text{var}(\psi) \left(\frac{\partial \hat{\theta}}{\partial \psi'}\right)' \\ &= J^{-1} D \Sigma_\psi D' J^{-T} \qquad (4.13) \\ &= (\Delta_\theta' V \Delta_\theta)^{-1} (\Delta_\theta' V \Delta_\psi) \Sigma_\psi (\Delta_\psi' V \Delta_\theta)(\Delta_\theta' V \Delta_\theta)^{-T}. \qquad (4.14) \end{aligned}$$

It is interesting to note that this formula does not depend on the sample size n. The amount of variance added due to the uncertainty of Ψ is independent of the sample size used to estimate $\hat{\theta}$.

The Δ_θ and Δ_ψ matrices are a function only of the model parameters in θ (Neudecker & Satorra, 1991). Estimates of these derivative matrices can be obtained by replacing the model parameters by their estimates. Thus, we have provided in equation 4.14 the expression that provides a one-step solution for the correction of the standard errors for the estimation of measurement error variances.

4.4 Application to a multiple regression model with uncorrelated regressors

In this section we give an analytical and numerical example of our solution by applying our method to multiple regression model with uncorrelated regressors.

Suppose the regression of a dependent latent variable η on two independent latent variables ξ_1 and ξ_2 is of interest. The dependent variable η is measured with an error-prone single indicator y, the independent variable ξ_1 with an error-prone single indicator x_1. The independent variable ξ_2 is measured perfectly by the error-free observed variable x_2, so that we may just as well write x_2 instead of ξ_2. The two independent variables ξ_1 and x_2

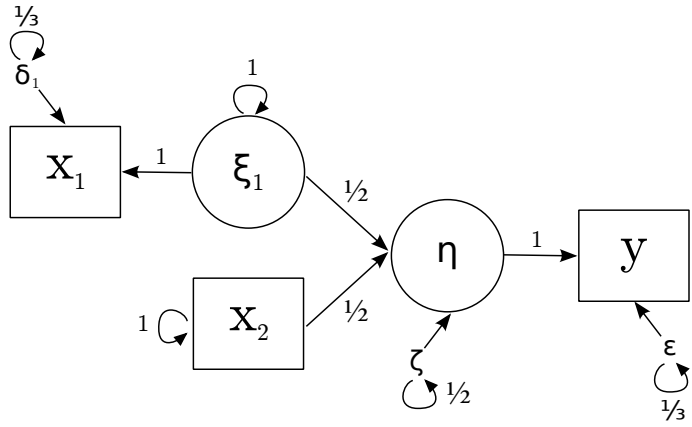


Figure 4.4: Path diagram for a multiple regression model. The independent variables ξ_1 and x_2 are uncorrelated. Circular arrows denote variance parameters. Unstandardized population parameter values are given. The reliabilities of x_1, x_2, and y equal 0.75, 1, and 0.75, respectively.

are, moreover, uncorrelated with each other. This model can be formulated as a structural equation model as shown in figure 4.4.

The figure provides fictional population parameter values for the unstandardized coefficients. It can be seen that x_2 has been perfectly measured while x_1 and y both have a reliability of 0.75. The regression coefficients of interest β_1 and β_2 both equal 0.5, and the R^2 also equals 0.5. The independent variables are uncorrelated.

Since x_2 is error-free, there are two estimated error variances that have been fixed in the matrix Ψ. This measurement error variance matrix is not exactly known, but has been estimated with its own variance matrix Σ_Ψ. The diagonal elements of this matrix, which denote the variance of the fixed error variance parameters, are named v_1 and v_2. The uncertainty about the error variance of the dependent variable y is called v_1 while v_2 is the uncertainty about the error variance of x_1.

We can now calculate, using equation 4.14 from the previous section, the term that should be added to the variance-covariance matrix of the parameter estimates. The "extra" term to be added to the standard variance equation can be expressed as a function of v_1 and v_2:

$$J^{-1}D\Sigma_\Psi D'J^{-T} = \begin{array}{c} \\ \text{var}(\zeta) \\ \text{var}(\xi) \\ \text{var}(x_2) \\ \beta_1 \\ \beta_2 \end{array} \begin{array}{c} \begin{array}{ccccc} \text{var}(\zeta) & \text{var}(\xi) & \text{var}(x_2) & \beta_1 & \beta_2 \end{array} \\ \begin{pmatrix} v_1 + \frac{v_2}{16} & & & & \\ \frac{v_2}{4} & v_2 & & & \\ 0 & 0 & 0 & & \\ -\frac{v_2}{8} & -\frac{v_2}{2} & 0 & \frac{v_2}{4} & \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{4.15}$$

The diagonal of this matrix is the expected added variance in the parameter estimates due to the uncertainty about Ψ. Each element in the matrix corresponds to a variance of a parameter or a covariance between two parameters. The parameters are shown in the row and column headers.

It can be seen that the standard error of the residual variance $\text{var}(\zeta)$ is the most affected.

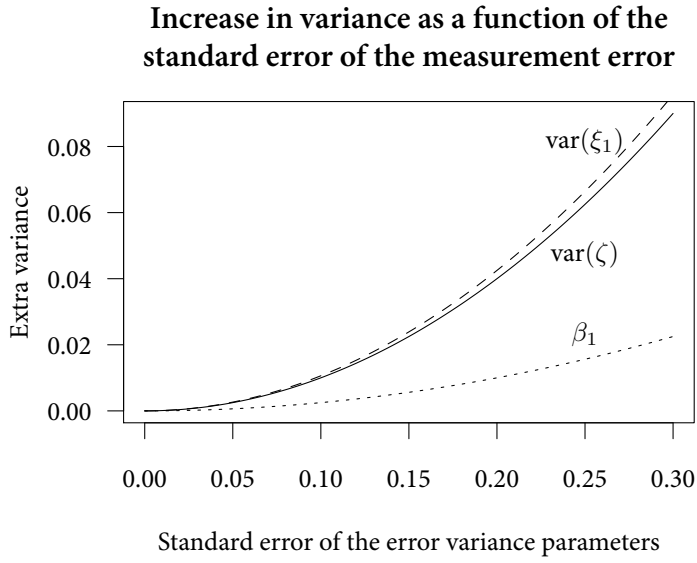


Figure 4.5: Increase in the variance of the three free parameter estimates as a function of the standard error of the measurement error variance (fixed) parameters.

The addition to the variance parameter of the independent variable ξ_1 is exactly equal to the variance of the error variance of its indicator x_1. The variance of the unstandardized regression coefficient β_1 is much (four times) less affected by uncertainty about the amount of measurement error than the other two parameters.

Notably, the variance of the regression coefficient β_2 and the variance parameter var(x_2), both parameters of the error-free variable x_2, are completely unaffected by the uncertainty about the error variances of y and x_1. Because the error-prone x_1 and error-free x_2 variables are uncorrelated, the derivative of the parameters β_2 and var(x_2) with respect to the two error variances equals zero. Since β_2 and var(x_2) are independent of the error variances of x_1 and y, their standard errors are immune to uncertainty about the error variances. While the situation shown is not often encountered in practice, it demonstrates how the effect of uncertainty about fixed parameters depends on the model structure.

The diagonal of the matrix as a function of the square root of the non-zero elements of Σ_Ψ is shown in figure 4.5. The increase in standard error of the parameter estimates will depend on the actual value of the "standard" variance of the parameter estimates.

This section illustrated our results with respect to the effect of uncertainty about fixed error variance parameters on standard errors. It was shown what effects occur in a simple model. The example illustrated that the effect on standard errors depends not only on the amount of uncertainty (v_1 and v_2 in this example), but also on the form of the model (uncorrelated regressors, one of which is error-free).

In the following section a series of Monte Carlo experiments evaluate to what extent the effect on standard errors in a more complex structural equation model is reflected in the coverage of confidence intervals. To validate our suggested correction, the section will

compare the performance of confidence intervals constructed using the "standard" variance with those constructed using our correction.

4.5 Monte Carlo evaluation of the new approach

In the previous sections it was shown that standard errors are underestimated when only the naive "standard" variance formula is used in the two-step method. This underestimation in the presence of uncertainty about fixed error variance parameters should lead to an undercoverage of nominal 95% confidence intervals when using only the "standard" variance formulas. Confidence intervals using our correction should always provide the nominal coverage rate. This section will use a series of Monte Carlo experiments to evaluate by simulation how well naive and corrected standard errors fare in the case of the more complex structural equation model discussed in section 4.2. The simulation is more realistic than the demonstration given in section 4.1, as in that section the population parameter values were assumed known: here we will also simulate sampling variation in the parameter estimates.

As our starting point for this study of the performance of the correction in structural equation models we set up the model of the example shown in figure 4.3. We took the parameter values obtained from the model estimation on the Danish example presented earlier as true population values. Measurement error variance parameters were set to equal those of table 4.1 in the population. This yielded a population covariance matrix of the four observed variables.

We then performed seven Monte Carlo experiments with increasing amounts of uncertainty about the fixed error variance parameters. In each experiment the following steps were followed:

1. The estimation of the error variances in a separate analysis was first simulated. For each of the four error variances, 2000 random draws were taken from normal distributions with means equal to the true population error variances. The standard deviations of these distributions equaled the standard error found in the example of section 4.2 multiplied by the scale factor for the experiment.

2. Next, 2000 samples of size 1500 were taken from a multivariate normal distribution with the population covariance matrix.

3. For each of the 2000 samples and estimated error variances the error variances of the model were fixed to the estimated error variances (step one of the two-step method).

4. The model shown in figure 4.3 was estimated, correcting for measurement error using the fixed error variance parameters (step two of the two-step method).

5. Both naive and corrected standard errors for the models were calculated for each of the 2000 samples, as well as 95% confidence intervals.

6. The percentage of samples out of 2000 for which the 95% confidence interval of the estimates contained the true population parameter values was calculated.

This process was repeated seven times in total, each time with a different amount of uncertainty. To vary the uncertainty we took the original estimates of the variances of

				Resulting s.e. of reliability coef, when variance of error var. scaled by					
	$\hat{\psi}$	s.e.	$\hat{\text{Rel}}$..	1	2	4	8	16	32
socialTrust	6.00	(0.22)	0.85	0.01	0.01	0.02	0.05	0.09	0.19
systemTrust	6.29	(0.24)	0.88	0.01	0.01	0.02	0.05	0.10	0.19
fearcrime	1.33	(0.04)	0.75	0.01	0.02	0.04	0.07	0.15	0.25
efficacy	1.17	(0.07)	0.80	0.02	0.03	0.07	0.14	0.24	0.33

Table 4.3: Error variances of the four observed variables of the models, and their basic standard errors (columns two and three). Corresponding reliability coefficients and their basic standard errors are also shown (columns four and five). Each Monte Carlo experiment multiplied the original variance of the error variance by the scale value. Corresponding standard errors of the reliability coefficients are shown in the last columns.

	$\bar{\theta}$	Rel. bias	sd($\hat{\theta}$)	$\overline{\text{s.e.}}_{\text{naive}}$	$\overline{\text{s.e.}}_{\text{cor}}$	C.I.$_{\text{naive}}$	C.I.$_{\text{cor}}$
β_{43}	0.77	-0.00	0.16	0.17	0.17	95.9%	95.9%
β_{42}	0.52	0.03	0.15	0.16	0.16	97.0%	97.0%
β_{34}	0.28	-0.08	0.20	0.21	0.21	98.5%	98.5%
β_{31}	-0.70	0.04	0.23	0.24	0.24	98.0%	98.0%
ϕ_{33}	12.19	0.06	3.87	1.96	1.96	97.0%	97.0%
ϕ_{22}	1.64	-0.00	0.10	0.10	0.10	94.3%	94.3%
ϕ_{34}	-6.22	-0.05	3.12	3.38	3.38	96.9%	96.9%
ϕ_{44}	12.42	0.03	1.60	1.68	1.68	94.4%	94.4%
ϕ_{12}	-0.60	0.00	0.07	0.08	0.08	95.4%	95.4%
ϕ_{11}	1.71	-0.00	0.11	0.11	0.11	95.0%	95.0%

Table 4.4: Simulation results without uncertainty in the estimates of measurement error variance.

the error variance parameters (the squares of the standard errors) and multiplied them by different "scale" values. A scale value of 0 indicates no uncertainty, i.e. perfect estimates of the error variances, a scale value of 1 indicates the same amount of uncertainty as was found in the example, and higher scale values indicate x times as much variation in the error variance estimate as found in our example. One consequence of this choice is that the different relative sizes of the standard errors will cause different effects on the standard errors of the model parameter estimates. This has been done in an attempt to introduce realistic differences in the relative amounts of uncertainty into the experiments. To provide a large range of different amounts of uncertainties, powers of two were taken as the scale values, yielding seven separate Monte Carlo experiments.

The scale values are essential in our simulations as they represent the actual amount of uncertainty present in the fixed parameters of the model for each experiment. The true population values of the error variances are shown in table 4.3 along with their standard errors in the example (scale = 1). These values are difficult to interpret in absolute terms. Therefore table 4.3 also provides the corresponding true population reliability coefficients (i.e. the square roots of the reliabilities) and the standard errors of these reliability coefficients for different scale values[4]. Standard errors for scale = 0 are not shown as they all

[4]In calculating these standard errors the estimation of a covariance matrix from a sample of size 1500 was also taken into account.

equal zero. This provides an insight into the amount of uncertainty corresponding to each scale value. It is clear from the table and the preceding discussion that we can expect that the naive standard errors will be more biased downwards as the scale factor increases, so that the coverage of 95% confidence intervals will worsen as the scale value becomes larger.

Without uncertainty in the measurement error estimates, the average of the simulations should equal the true parameter values and the coverage of 95% confidence intervals using both the uncorrected and the corrected standard errors should approach 95% as the number of simulations increases.

Table 4.4 shows the results of a monte carlo simulation without any uncertainty in the measurement error variances: in this experiment the assumption of perfect certainty about the error variances is correct. The table shows the average over samples of the parameter estimates $\bar{\theta}$, as well as the relative bias $(\bar{\theta} - \theta)/\theta$. The fourth column shows the standard deviation over samples $\text{sd}(\hat{\theta})$ of the estimates. The subsequent two columns show the average of the "naive" and "corrected" standard errors. As might be expected, these two are exactly equal when the assumption of perfect certainty is correct. The last two columns show the percentage of samples for which the true parameter value is included in nominal 95% confidence intervals.

In general the results shown in table 4.4 suggest that the parameter estimates when there is no uncertainty about the fixed coefficients are approximately unbiased and that 95% confidence interval provide a coverage close to this nominal rate.

The real question, however, is what results are obtained when the assumption of perfect certainty about the error variances does not hold. We will now briefly discuss the simulation results shown in figures 4.6 and 4.7 which show the properties of the naive and corrected standard errors at increasing levels of uncertainty.

Uncertainty about the measurement error variance does not influence the unbiasedness of the estimates. In all simulations the estimates obtained are indistinguishable from those shown in the first column of table 4.4, while also the relative bias is not affected. A slight difference starts to occur at very high levels of uncertainty due to an increased number of improper solutions. When these improper solutions are excluded from the analysis the effect disappears.

The standard deviation of the estimates across samples is, however, clearly affected by increasing uncertainty about the error variance parameters. Figure 4.6 shows, for each parameter of the model except ϕ_{21}, the standard deviation of that parameter's estimate across samples (the solid line marked "True"). It can clearly be seen that for all parameters the variability of the parameter estimate increases as the amount of uncertainty is increased.

The same figure also shows, in each graph, the average of the "naive" standard errors and our "corrected" standard errors. The regular standard errors calculated under the assumption of perfect certainty about the error variances do not change appreciably as the uncertainty about the error variances increases. Since the "true" standard deviations across samples does increase for all parameters, this indicates an underestimation of the standard error. Our corrected standard errors do increase with the uncertainty, and follow the "true" standard deviation closely.

Figure 4.7 summarizes the main coverage results for all parameters. The graph on the left-hand side in figure 4.7 shows the proportion of 95% confidence intervals constructed using "naive" standard errors (i.e. standard program output) that contain the true population parameter value. It can be seen that while some parameter estimates are not affected by the amount of uncertainty in this model, for other parameters the coverage properties

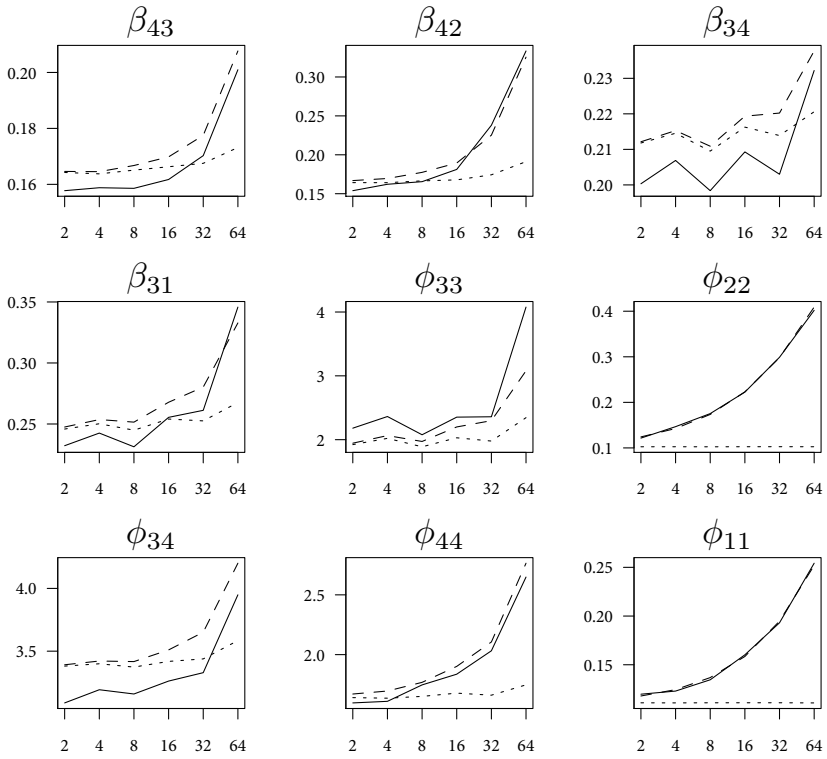


Figure 4.6: True standard deviations across the simulations (solid lines), corrected standard errors (striped lines), and "naive" standard errors (dotted lines) for different uncertainty scale values. Each graph represents these relationships for one model parameter.

worsen considerably. At the highest amount of uncertainty studied, the 95% confidence interval for the variance of "efficacy", has deteriorated to a clearly unacceptable 40% coverage. Other variance parameters are also affected, while unstandardized regression coefficients appear impervious in this model.

The right hand side of figure 4.7 shows the coverage of 95% confidence intervals constructed using our corrected standard errors. The graphs clearly shows that our correction succeeds in correcting the standard errors of the parameter estimates for uncertainty in the fixed error variance parameters.

4.6 Discussion and conclusion

Measurement error biases parameter estimates in structural equation models, making it impossible to ignore in the analysis of such models. SEM allows the researcher to subsume measurement error directly into the model, simultaneously estimating and correcting for measurement error.

However, this method requires that the study design is adequate both for the analysis

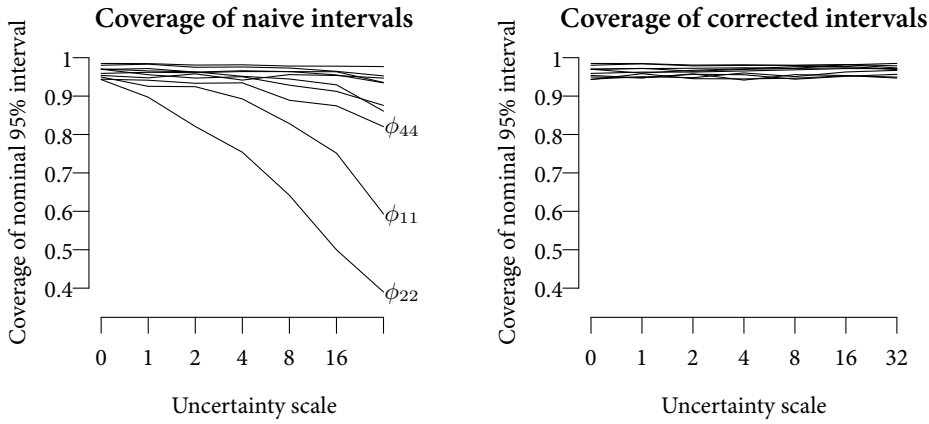


Figure 4.7: Coverage of nominal 95% confidence intervals for the different parameters of the model, as a function of the amount of uncertainty in the measurement error variance estimates. Left the coverage using the naive standard errors is shown, while the right graph shows the coverage when using the newly proposed corrected standard errors.

of the substantive model and the evaluation of measurement error. It also requires that the researcher is knowledgeable about both measurement and substantive issues. A popular alternative that circumvents these disadvantages is to use estimates of measurement error variances from external studies to correct analyses. Error variance parameters in such analyses are fixed to the estimates that were obtained separately, followed by an analysis of the substantive model with correction for measurement error. We have called this method the "two-step" approach.

The two-step approach does not take into account that the fixed error variances are, in reality, estimates with a certain amount of uncertainty. We have shown, first by a simple simulation and subsequently by analytics, that the uncertainty about fixed error variance parameters in the two-step approach will bias standard errors downwards.

The effect of uncertainty about the error variances and therefore the downwards bias depends on two factors: the *amount of uncertainty*, and the *relationship between the fixed and free parameters*. When there is no uncertainty (zero variance) about the fixed parameters, the corrected standard errors will equal the uncorrected standard errors.

However, even when there is uncertainty about the fixed parameters, there are models in which certain parameters' variance will remain completely unaffected. This was demonstrated by the example of a multiple regression where the regressors are uncorrelated and one of the regressors is error-free. In such a case the parameters related to the error-free variable are completely independent of the measurement error variance of the error-prone variable, and uncertainty about the fixed error variance will not affect the unstandardized regression coefficient of the error-free variable. The independence in this model is created by the lack of correlation between the regressors.

Section 4.3 presented an analytical solution to the problem of downwards biased standard errors in the two-step method in the form of a term that should be added to the "naive" covariance matrix of the parameter estimates. An estimate of this term can be

readily calculated from the parameter estimates of the model[5].

The subsequent section evaluated our newly developed method by Monte Carlo simulation. Simulations at different levels of uncertainty showed that the variability of the estimates can increase considerably due to uncertainty about the measurement error variance. It was shown that only our corrected standard errors provided the nominal confidence interval coverage, while "naive" confidence intervals can deteriorate substantially in the presence of increasing uncertainty about the fixed parameters.

Comparison with other approaches There are three alternative approaches to the analytical correction we have discussed.

The first alternative to our procedure is the analytical solution for multivariate regression models estimated by the method of moments given by Amemiya and Fuller (1984), and further discussed by Fuller (1987). This solution has several disadvantages. First, it is formulated only for multivariate regression models, and therefore could not be used for more complex models such as the one given as an example in section 4.2. Second, it assumes that the error variance estimates follow a chi square distribution with degrees of freedom, as is the case for certain simple measurement models. This makes the use of measurement error estimates from more complex measurement models much more difficult. Last, extensions developed for SEM such as categorical, count, or censored dependent variables and complex sampling designs are not available. To our knowledge this solution has only been implemented in the software EV CARP (D. Schnell et al., 1987).

A second plausible alternative to the analytical solution of equation 4.14 is a parametric bootstrap procedure (Efron & Tibshirani, 1997). The model or models used to estimate the measurement error variances yield distribution for these parameters. A typical example would be error variances from a factor model, which are asymptotically normally distributed.

Sampling k times from this distribution, and re-applying the estimation procedure correcting for measurement error each time, we obtain a consistent estimate of the variance over replications of $\hat{\theta}$ that is purely due to the variation in Ψ. This quantity can be used as an estimate of the right-hand term of 4.9. Therefore the corrected variance can be obtained by summing the variance from the analysis correcting for measurement error in the previous section and the parametric bootstrap variance.

The bootstrap procedure has the advantage that one does not need to perform the calculations. However, it also has important disadvantages. First, if the measurement error variance is too low or too high relative to the corresponding observed covariances, the bootstrap procedure will yield many inadmissible or non-convergent solutions. This may, in turn, yield inconsistent estimates of the variance to be added. Often measurement error estimates are obtained from multiple sources, and to the above problem are added concerns about the second-order probabilities of inadmissible solutions. Finally, the procedure is conceptually simple but can be tedious to implement, adding another step to the estimation procedure, and requires non-standard software.

The third and final alternative to our solution is MCMC estimation of the model as performed in Bayesian structural equation modeling (Lee, 2007). This requires that in the specification of the model, the distribution of the error variances or reliabilities is explicitly specified along with the measurement and structural part of the model. An advantage is

[5]An implementation for the SEM package `OpenMx` in `R` is available upon request from the first author.

that non-normally distributed measurement error parameters can be taken into account. A disadvantage is that the entire estimation procedure is changed and must be reformulated to take the uncertainty about the amount of measurement error into account, whereas our method is simply an extension of standard methods. Thus this option may be attractive when the researcher already planned for other reasons to estimate the model by MCMC.

In conclusion, different alternatives to our solution exist but each introduces additional complexities and in some cases limits the models that can be analyzed. Considering the disadvantages of the alternatives and the fact that equation 4.14 can be hidden from the user by simple function calls which are provided for free by the authors, we recommend the analytic approach.

Future studies The method presented was evaluated by Monte Carlo simulation using a specific model. However, a key aspect of the effect of uncertainty on a parameter as shown in equation 4.8 is that the parameter must be influenced by changes in the error variance. This was demonstrated explicitly by the numerical example of a multiple regression with uncorrelated regressors. This example could be constructed in such a way that some parameters are affected by the uncertainty about the error variances, while others are immune to this effect.

The amount of uncertainty was taken from a particular example and varied from that point. It was not shown that such degrees of uncertainty occur in practical research. However, there are indications that they do.

For example, an often-cited reliability study by Ware et al. (1978, pp. 40–42) on quality of life and subjective health indicators reports test-retest reliability coefficients between 0.4 and 0.7 based on a sample size of 138 persons. This suggests the standard errors for the reliabilities in this study are between approximately 0.04 and 0.06. The simulation shows that such standard errors are quite high and can have serious implications for inference. Other studies may employ a larger sample, but report separate coefficients for men and women of different age groups (e.g. in Lundberg & Manderbacka, 1996, n=204 and n=409). This will also cause the uncertainty in the reliability estimates to increase substantially.

These examples do not give a systematic study of the literature on reliability and as such cannot be generalized. However, it cannot be excluded that there are many cases in which the amount of uncertainty in reliability coefficients or error variance parameters, combined with the form of the model and the parameters of interest (e.g. standardized or unstandardized regression coefficients), cause the correction we propose to be necessary for correct inference.

In short, the strength or existence of an effect depends on the model used as well as the amount of uncertainty. The evaluation provided in this paper is therefore limited in that we do not know whether the characteristics of the model chosen for the evaluation are typical of applied work. This is clearly an important topic for further research, as it determines under which conditions the effect of uncertainty about error variance is a problem for practitioners.

Without more detailed knowledge about these conditions one should not exclude the possibility that inference is affected. Therefore whenever measurement error variances have been fixed to an estimate about which uncertainty exists, the correction to standard errors presented here should be applied.

References

References

Agresti, A. (2002). *Categorical data analysis*. Wiley-Interscience.

Alwin, D. F. (2007). *Margins of error: a study of reliability in survey measurement*. Wiley-Interscience.

Amemiya, Y., & Fuller, W. A. (1984). Estimation for the multivariate errors-in-variables model with estimated error covariance matrix. *The Annals of Statistics*, *12*(2), 497–509.

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

Bassi, F., & Fabbris, L. (1997). Estimators of nonsampling errors in interview-reinterview supervised surveys with interpenetrated assignments. *Survey Measurement and Process Quality*, 733–751.

Beckie, T., & Hayduk, L. (1997). Measuring quality of life. *Social Indicators Research*, *42*(1), 21–39.

Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, *45*(3), 289–308.

Biemer, P. (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, P., & Bushery, J. M. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, *26*(2), 139–152.

Biemer, P., Groves, R. M., & Lyberg, L. E. (2004). *Measurement errors in surveys*. Wiley-IEEE.

Biemer, P., & Trewin, D. (1991). A review of measurement error effects on the analysis of survey data. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. D. andN. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (chap. 27). New York: John Wiley & Sons, Inc.

Billiet, J. B., & McClendon, M. K. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 608–628.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

Boker, S., Neale, M., Maes, H., Metah, P., Kenny, S., Bates, T., et al. (2010). Openmx: The openmx statistical modeling package [Computer software manual]. Available from `http://openmx.psyc.virginia.edu` (R package version 0.2.10-1172)

Bollen, K. (1989). *Structural equations with latent variables*. Wiley.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*, *56*, 81–105.

Cannell, C., Miller, P., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological methodology*, *12*, 389–437.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). Nonlinear measurement error models. *Monographs on Statistics and Applied Probability.(Chapman and Hall, New York) Volume*, *63*.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective.* Monographs on Statistics and Applied Probability 105.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5–32.

Cochran, W. G. (1977). *Sampling techniques*. Wiley-India.

Coenders, G. (1996). *Structural equation modeling of ordinally measured survey data*. Unpublished doctoral dissertation, Universitat Ramon Llull.

Coromina, L., Saris, W. E., & Oberski, D. L. (2008). The quality of the measurement of interest in the political issues in the media in the ESS. *ASK. Spo ecze stwo. Badania. Metody*(17), 7.

Corten, I. W., Saris, W. E., Coenders, G., van der Veld, W., Aalberts, C. E., & Kornelis, C. (2002). Fit of different models for multitrait-multimethod experiments. *Structural Equation Modeling*, *9*, 213–232.

Crowne, D. P., & Marlowe, D. (1960, August). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354.

De Leeuw, E., & Van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis. *Telephone survey methodology*, 283–299.

Dijkstra, W., & Van der Zouwen, J. (1982). *Response behaviour in the survey-interview*. Academic Pr.

Dillman, D. (2007). *Mail and internet surveys: The tailored design method*. John Wiley & Sons Inc.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, 295–311.

Donsbach, W., & Traugott, M. (2007). *The SAGE handbook of public opinion research*. Sage Publications Ltd.

Droitcour, J., Caspar, R., Hubbard, M., Parsley, T., Visscher, W., & Ezzati, T. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. *Measurement errors in surveys*, *11*, 185–210.

Efron, B., & Tibshirani, R. (1997). *An introduction to the bootstrap*. Chapman & Hall.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Fowler, F., & Mangione, T. (1990). *Standardized survey interviewing; minimizing interviewer-related error*. Thousand Oaks, CA: Sage Publications, Inc.

Fuller, W. A. (1987). *Measurement error models*. Wiley New York.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215.

Goudy, W. (1976). Nonresponse effects on relationships between variables. *Public Opinion Quarterly*, *40*(3), 360.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R. M. (2002). *Survey nonresponse*. Wiley-Interscience.

Groves, R. M., & Couper, M. (1998). *Nonresponse in household interview surveys*. Wiley-Interscience.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, *72*(2), 167.

Haberman, S. (1979). *Analysis of qualitative data*. New York: Academic Press.

Häder, S., & Lynn, P. (2007). How representative can a multi-nation survey be? In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the european social survey*. SAGE.

Hagenaars, J. A. P. (2002). Directed loglinear modeling with latent variables. In J. A. P. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis*. Cambridge University Press.

Hagenaars, J. A. P., & Heinen, T. G. (1982). Effects of role-independent interviewer characteristics on responses. *Response Behaviour in the Survey-Interview—Dijkstra W, van der Zouwen J, eds*, 91–130.

Hagenaars, J. A. P., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, United Kingdom: Cambridge University Press.

Hambleton, R. K., Rogers, H. J., & Swaminathan, H. (1995). *Fundamentals of item response theory*. Sage Publ.

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. P. (2002). *Cross-cultural survey methods*. Wiley-Interscience.

Hayduk, L. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Johns Hopkins Univ Pr.

Heinen, T. G. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage Thousand Oaks.

Heise, D. R. (1970). Causal inference from panel data. *Sociological Methodology*, *2*, 3–27.

Hui, H. C., & Triandis, H. C. (1989, September). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*(4), 243–252.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, *94*, 401–419.

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, *48*, 398–407.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*, 443–477.

Jöreskog, K. G. (1990). New developments in LISREL: analysis of ordinal variables us-

ing polychoric correlations and weighted least squares. *Quality and Quantity*, *24*, 387–404.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Scientific Software.

Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. (2007). *Measuring attitudes cross-nationally: Lessons from the european social survey*. SAGE.

Kieruj, N. D., & Moors, G. B. (frth). Response scales' vulnerability to acquiescence and extreme response style behavior.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, *57*(297), 92–115.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(1), 1–37.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), pp213–236.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton, Mifflin.

Lee, S. (2007). *Structural equation modeling: A Bayesian approach*. Wiley Chichester, UK:.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley-Interscience.

Little, R. J. A., & Vartivarian, S. (2006). Does weighting for nonresponse increase the variance of survey means? *Quality Control and Applied Statistics*, *51*(6), 623.

Lord, F. M. (1952). *A theory of test scores*. New York: Psychometric Society.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental scores. *Reading, Addison–Wesley*.

Lundberg, O., & Manderbacka, K. (1996). Assessing reliability of a measure of self-rated health. *Scandinavian Journal of Public Health*, *24*(3), 218.

Lyberg, L. (1997). *Survey measurement and process quality*. New York, NY: Wiley-Interscience.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, 223–264.

Magnus, J. R., & Neudecker, H. (2002). *Matrix differential calculus with applications in statistics and econometrics, third edition*. John Wiley, Chichester.

Mahalanobis, P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society A*, *109*.

McArdle, J., & McDonald, R. (1984). Some algebraic properties of the Reticular Action Model for moment structures. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 234–251.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*(4), 351.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551–560.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in mplus. *Mplus Web Notes*.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407–419.

Muthén, B. O., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*, 133–142.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological methodology*, *25*, 267–316.

Muthén, L. K., & Muthén, B. O. (1998). Mplus user's guide. *Los Angeles: Muthén & Muthén*, *2004*.

Neudecker, H., & Satorra, A. (1991). Linear structural relations: Gradient and Hessian of the fitting function. *Statistics and Probability Letters*, *11*, 57–61.

Newton, K. (2007). Social and political trust. *Oxford handbook of political behavior*, 342.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Oberski, D., Saris, W. E., & Hagenaars, J. A. P. (2007). Why are there differences in measurement quality across countries? In G. Loosveldt, M. Swyngedouw, & B. Cambré (Eds.), *Measuring meaningful data in social research.* Leuven: Acco.

Oberski, D., Saris, W. E., & Kuipers, S. (2004). *SQP: survey quality predictor.* Available from `http://www.sqp.nl/`

Olson, K., & Kennedy, C. (2006). Examination of the relationship between nonresponse and measurement error in a validation study of alumni. *Proceedings of the Survey Research Methods Section*.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, *14*, p485–500.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *195*, 1–405.

Putnam, R. (2001). *Bowling alone: The collapse and revival of American community*. Simon and Schuster.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Raykov, T. (2009). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 195–212.

Révilla, M., & Saris, W. (frth). A comparison of surveys using different modes of data collection: European Social Survey versus LISS panel. *RECSM Universitat Pompeu Fabra working papers*. (http://upf.edu/survey/)

Rhodes, R., Macdonald, H., & McKay, H. (2006). Predicting physical activity intention and behaviour among children in a longitudinal sample. *Social Science & Medicine*, *62*(12), 3146–3156.

Roscino, A., & Pollice, A. (2006). A generalization of the polychoric correlation coefficient. In A. C. M. R. Sergio Zani & M. Vichi (Eds.), *Data analysis, classification and the forward search.* Berlin Heidelberg: Springer.

Rost, J., & Walter, O. (2006). Multimethod item response theory. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology.* Washington, DC: American Psychological Association.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*.

Saris, W. E. (1988). *Variation in response functions: A source of measurement error in attitude research*. Sociometric Research Foundation.

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 575–599). New York: John Wiley & Sons.

Saris, W. E., & Gallhofer, I. N. (2007a). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience.

Saris, W. E., & Gallhofer, I. N. (2007b). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1*.

Saris, W. E., Révilla, M., Krosnick, J. A., & Schaeffer, E. M. (frth). Comparing questions with agree/disagree response options to questions with item-specific response options.

Saris, W. E., Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological Methodology*, *34*, 311–347.

Saris, W. E., Satorra, A., & Sorbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, *17*, 105–129.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582.

Saris, W. E., van der Veld, W., & Gallhofer, I. N. (2004). Development and improvement of questionnaires using predictions of reliability and validity. In S. Presser et al. (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 275–299). Hoboken: Wiley.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*(1), 131–151.

Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments. *Quality and Quantity*, *24*, 367–386.

Satorra, A., & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*, *10*(3), 235–249.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, *25*, 341.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of Multitrait-Multimethod matrices. *Applied Psychological Measurement*, *10*(1), 1–22.

Schnell, D., Park, H., & Fuller, W. (1987). *EV CARP.* Ames, Iowa: Statistical Laboratory, Iowa State University. (http://cssm.iastate.edu/software/evcarp.html)

Schnell, R., & Kreuter, F. (2003). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, *21*(3), 389–410.

Schumacker, R., & Lomax, R. (2004). *A beginner's guide to structural equation modeling.* Lawrence Erlbaum.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology, vol, 25*.

Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, *77*(380), 848–854.

Small, B., Hertzog, C., Hultsch, D., & Dixon, R. (2003). Stability and change in adult personality over 6 years: findings from the Victoria Longitudinal Study. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *58*(3), P166.

Smit, J., Dijkstra, W., & Van der Zouwen, J. (1997). Suggestive interviewer behaviour in surveys: An experimental study. *Journal of Official Statistics*, *13*, 19–28.

Stoop, I. A. L. (2005). *The hunt for the last respondent: Nonresponse in sample surveys*. Aksant Academic Pub.

Stoop, I. A. L., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving Survey Response: Lessons Learned from the European Social Survey*. Wiley.

Takane, Y., Young, F., & de Leeuw, J. (1977, March). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, *42*(1), 7–67.

Tångdahl, S. (2005). The variance of some common estimators and its components under nonresponse. *Working Papers*. Available from `http://ideas.repec.org/p/hhs/oruesi/2005_009.html`

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. ASA.

Tourangeau, R., & Smith, T. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*(2), 275.

Uebersax, J. S., & Grove, W. M. (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, *49*, 823–835.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications, Inc.

Varki, S., & Colgate, M. (2001). The role of price perceptions in an integrated model of behavioral intentions. *Journal of Service Research*, *3*(3), 232.

Veld, W. van der. (2006). *The survey response dissected: A new theory about the survey response proces*. Amsterdam: University of Amsterdam.

Vermunt, J. K., & Magidson, J. (2004a). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 41–63). Mahwah: Erblaum.

Vermunt, J. K., & Magidson, J. (2004b). Latent class models. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 175–198). Thousand Oaks, California: Sage Publications, Inc.

Vermunt, J. K., & Magidson, J. (2005a). Latent gold 4.5 user's guide. *Belmont, MA: Statistical Innovations Inc*.

Vermunt, J. K., & Magidson, J. (2005b). Technical guide for latent GOLD 4.0: Basic and advanced. *Belmont Massachusetts: Statistical Innovations Inc*.

Voogt, R. (2004). *I'm not interested: Nonresponse bias, response bias and stimulus effects in election research*. Amsterdam: University of Amsterdam.

Ware, J., Davies-Avery, A., & Donald, C. (1978). Conceptualization and measurement of health for adults in the health insurance study: Vol. V, general health perceptions. *Santa Monica: The Rand Corporation*.

Warner, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research*. University of Chicago Press.

Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, *74*(3), 193–212.