Survey Research Methods (2013)
Vol. 7, No. 3, pp. 157-168
ISSN 1864-3361
http://www.surveymethods.org

Evaluating Relative Mode Effects on Data Quality in Mixed-Mode Surveys

Jorre T. A. Vannieuwenhuyze
Institute for Social & Economic Research, University of Essex, UK

Melanie Revilla
Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra, Spain

In order to compare data-quality of different data-collection modes, multitrait-multimethod (MTMM) experiments have been implemented in a mixed-mode survey parallel to the fourth round of the European Social Survey (ESS, 2008/2009). Special interest lies in measurement effects between the modes, which refer to the raw effect of a data-collection mode on quality. Nevertheless, mere comparison between quality estimates of the different modes does not allow drawing conclusions about measurement effects because they are completely confounded with selection effects. However, by simultaneous analysis of the mixed-mode data and the main ESS data and by treating the dataset of origin as an instrumental variable, some conditional measurement effects and selection effects can be disentangled. This paper provides a preliminary exploratory analysis of this approach. The results generally yield low to fair measurement effects, while the selection effects on some items are rather large. Overall differerences between the modes are thus mainly caused by differences in respondent composition rather than differences in measurement error. The analysis of the ESS data, however, reveals many problems which are caused by design deficiencies. These deficiencies include, among others, the inclusion of too many modes, design-driven violations of the model assumptions, too few methods within the MTMM experiments, and small sample sizes. These deficiencies should be resolved in future studies.

Keywords: Mode effects, Measurement effects, Selection effects, Multitrait-multimethod, Reliability, Validity, Quality, Instrumental variable, European Social Survey

1 Introduction

The European Social Survey (ESS) started in 2002 as a biennial survey about changing social values and attitudes in Europe. In order to encourage equivalence across countries, all waves' main surveys have chiefly been carried out by personal face-to-face interviews so far. However, because of the expensiveness of face-to-face interviews and declining response rates, small *mixed-mode* surveys have been set up parallel to the main single-mode surveys in order to examine the suitability of mixed-mode designs instead of single-mode face-to-face designs in future rounds. Within these mixed-mode surveys, data from different groups of sample members is collected by different data-collection modes like personal face-to-face interviews, telephone interviews, or Web questionnaires.

Using a mixed-mode design instead of a single-mode face-to-face design might result in lower selection error, i.e. the error introduced by only observing a small subset of population members instead of the entire population (de Leeuw, 2005; Dillman, Smyth, & Christian, 2009). Firstly, a mixed-mode survey may reduce systematic selection error (e.g. nonresponse error) relative to a single-mode face-to-face survey because certain population members might not be willing or able to respond face-to-face in the single-mode survey but do respond by telephone or Web in the mixed-mode survey. In this case, the mixed-mode survey may offer greater external validity than the single-mode survey. Secondly, a mixed-mode survey may reduce random selection error (e.g. sampling error) relative to a single-mode face-to-face survey because some respondents may respond by a cheap mode in the mixed-mode survey, while they would respond by the expensive face-to-face mode in the single-mode survey. As a result, with a mixed mode-design, larger samples can be drawn within the same budget constraints. In this case, the mixed-mode survey may offer greater external reliability than the single-mode survey.

However, data-collection modes may affect data-quality in particular ways (de Leeuw, 2005; Voogt & Saris, 2005; Dillman et al., 2009). A difference in data-quality between two modes is called a *measurement effect* because it is caused by differences in measurement error accompanying both modes. Put differently, a measurement effect occurs if two data-collection modes provide data of different quality for the same group of respondents. As such, measurement effects are problematic because, firstly, they threaten the validity of comparisons between the respondents of different data-collection modes and, secondly, they also threaten the comparability of mixed-mode data with other single-mode

Contact information: Jorre T. A. Vannieuwenhuyze, Institute for Social & Economic Research, University of Essex, UK, e-mail: jtavan@essex.ac.uk

data (for example, the past single-mode ESS rounds).

Nonetheless, it is difficult to analyse measurement effects on data quality because the measurement effects may be confounded by selection effects (Revilla, 2010). *Selection effects* occur when the groups of respondents selected for the different modes differ on the target variables. A difference between two data quality estimates obtained from two modes in a mixed-mode dataset might thus not only be caused by the data-collection modes themselves (i.e. a measurement effect) but also by a mere difference between the respondent groups selected for both modes (i.e. a selection effect).

This paper aims to explore the instrumental variable method for examining measurement effects and selection effects on data quality estimates in the ESS obtained from Multitrait-Multimethod (MTMM) experiments. The instrumental variable method provides one possible solution to circumvent confounding measurement and selection effects by merging mixed-mode data with comparable single-mode data, and has already been used for the analysis of the ESS data by Vannieuwenhuyze, Loosveldt, and Molenberghs (2010). MTMM experiments, in turn, provide useful tools to examine and estimate the quality of data obtained by particular data-collection modes by analysing repeated sets of questions using different measurement methods (e.g. response scales)(Revilla, 2010). The application of the instrumental variable method on MTMM quality estimators in the ESS data, however, reveals many problems which are caused by design deficiencies. These deficiencies include, among others, the inclusion of too many modes, design-driven violations of the model assumptions, too few methods within the MTMM experiments, and small sample sizes. A discussion of these deficiencies provides guidelines for design improvement in future mixed-mode research, and forms the main goal of this paper.

The remainder of the paper is structured as follows. Section 2 introduces the Dutch ESS round four mixed-mode survey, which will be used for analysis. Section 3 discusses the use of the instrumental variable method within the analysis of MTMM experiments. Section 4 provides an exploratory analysis of measurement effects and selection effects on MTMM quality estimators in the ESS. Finally, Section 5 discusses the problems and drawbacks of the current analysis, and suggests solutions for future research by using appropriate survey designs.

2 The European Social Survey

The analysis data of this paper stem from a mixed-mode survey which was set up parallel to the fourth wave of the main European Social Survey (ESS) in the Netherlands in 2008-2009 (Eva et al., 2010). The main Dutch ESS survey was completely conducted by Computer-Assisted Personal Interviews (CAPIs) using showcards. In the mixed-mode design, sample members were asked to respond through one of the three data-collection modes, i.e. a Web Self-Administration Questionnaire (WSAQ), a Computer-Assisted Telephone Interview (CATI), or a CAPI. Half of the sample was assigned to a 'concurrent' design and half of the sample to a 'sequential' design. In a concurrent design the data-collection modes are offered simultaneously, while in a sequential design the modes are offered sequentially. However, in both the concurrent and sequential design, all modes were offered from the very first telephonic contact and no significant differences could be noticed between both designs regarding the mode selection of respondents. As a consequence, we decided to further ignore this distinction between the concurrent and the sequential design.

Both the main survey and the mixed-mode survey started from two independent random samples. The mixed-mode survey started from a sample of 1756 people with a matched phone number while the main survey started from a sample of 2674 people with a matched phone number. Sample members without matched phone numbers were also included in both surveys, but, in the mixed-mode survey, almost all these respondents answered by CAPI due to the particular form of the design. As a consequence, these people are hardly useful to evaluate mode effects. For both surveys, a simple random sample of households was drawn from the very same sampling list and one household member older than 15 years was randomly selected within each selected household. To correct for differences in household sizes, normalized design weights proportional to the household size are used in all further analyses. Response frequencies of both datasets can be found in Table 1.

Both the main ESS sample and the mixed-mode ESS sample are separately weighted on a set of socio-demographic variables (a cross-classification of age and gender, urbanization, and household size) by using raking procedures (Deming & Stephan, 1940; Izrael, Hoaglin, & Battaglia, 2000). The marginal population distributions of these variables were obtained from the 'Centraal Bureau voor de Statistiek' (CBS, see http://www.cbs.nl).

3 Methods

This section discusses the use of the instrumental variable method for the estimation of measurement and selection effects on quality estimates from MTMM experiments. Subsection 3.1 provides a formal description of the confounding between measurement effects and selection effects. Subsection 3.2 discusses the instrumental variable method to circumvent this confounding. Subsection 3.3 describes Multitrait-Multimethod (MTMM) experiments, which can be used to estimate data quality and which are implemented in the ESS. Subsection 3.4 discusses the estimation procedures when the instrumental variable method is applied to MTMM experiments.

3.1 The confounding problem

Merely comparing a data quality estimator q^2 across the data-collection modes within the mixed-mode survey data does not allow drawing conclusions about measurement effects on data quality because differences in q^2 between the modes can also be caused by different groups of sample members being selected for the different data-collection

Table 1 The fourth round of the European Social Survey (ESS) includes a single-mode CAPI sample and a mixed-mode WSAQ-CATI-CAPI sample.

	round 4 main ESS		round 4 mixed-mode ESS
Frequencies:			
WSAQ			333
CATI			177
CAPI	1363[a]		215
none-response	1150		862
Non-eligible	161		169
Response rate	0.542		0.457
Difference Response rates		0.086	
Std. error		0.016	
p		< 0.001	

[a] Because the MTMM experiments (see Table 2) were only assigned to approximately one third of the respondents, the actual sample frequencies are much lower: 439 for Media, 452 for Social Trust, 413 for Political Trust, and 437 for Satisfaction.

modes, i.e. a selection effect (Revilla, 2010; Vannieuwenhuyze et al., 2010). This problem of confounded measurement and selection effects overlaps with the central problem of causal inference (e.g., among others, Morgan & Winship, 2009; Pearl, 2009; Weisberg, 2010) and will be discussed in this section.

It should first be noted that two distinct research questions can be put forward. The first research question deals with a comparison between CAPI on the one hand and a combination of CATI and WSAQ on the other hand. This question arises from interest in a possible change in data quality merely because some respondents do not respond by the regular ESS CAPI mode. Whether these respondents answer either by CATI or by a WSAQ is less important here. The main question merely is whether a mixed-mode survey including all three modes can replace the regular ESS single-mode face-to-face design without loss of quality. The second research question deals with a comparison between each of the modes separately. Indeed, measurement and selection effects can also be expected between CATI and WSAQ, and knowledge of these mode effects might provide insight into the suitability of mixed-mode designs including two modes instead of three. However, this paper will only focus on the first research question because the instrumental variable method, which forms the prime analysis model of this paper, in combination with the ESS round 4 data does not allow answering the second research question without additional problematic assumptions.

The occurrence of measurement effects between the modes means that respondents would have responded differently if different data-collection modes had been used. As a consequence, each respondent can theoretically be represented by two data-lines instead of one where each line represents the respondent's answer when a particular data-collection mode had been used. A variable D can then be defined which refers to this *mode of Data-collection*, and takes either values p or tw. The value p refers to the data-lines where the respondents' data had been collected by CAPI. The value tw refers to the data-lines where the respondents' data had been collected by CATI or WSAQ depending on the choice the unit would make when only these two modes were offered. Apart from D, another variable G can be defined which refers to the mode Group to which a respondent is really selected in the mixed-mode survey. Like D, this variable takes either value p if a respondent answers by CAPI or tw if a respondent answers by CATI or WSAQ.

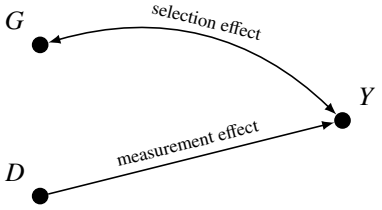
The target variables Y, which are used to calculate data quality, may relate to both D and G (Figure 1). First, by definition, Y is causally affected by the mode of data-collection D because the mode defines the measurement error in the response. The effect of D on Y thus denotes the measurement effect between the modes. Second, Y may relate to the mode group G for which a respondent is selected in the mixed-mode survey. The relation between G and Y reflects a selection effect as it implies differences in respondent compositions between the modes.

Within the ideal situation where the responses of all respondents are observed in both modes p and tw, there is no relation between D and G (Figure 1(a)). Indeed, D and G are independent because two data-lines can theoretically be defined for each respondent, one for each mode of data-collection, irrespective of the actual mode group for which the respondent is selected in the mixed-mode survey. Of course, some of these datalines are not observed in practice and this non-observation causes estimation problems of selection and measurement effects, as will be discussed below.

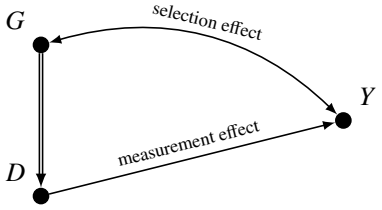
Let $Y_{d,g}$ and $q^2_{d,g}$ further refer to the conditional variable $Y|D = d, G = g$, and the conditional quality estimate $q^2(Y_{d,g})$, i.e. the quality of data obtained by mode d for the respondents who are selected for mode group g within the ESS mixed-mode survey. It is now tempting to calculate the conditional *Measurement Effect*

$$\mathrm{ME}(q^2) = q^2_{p,tw} - q^2_{tw,tw}. \quad (1)$$

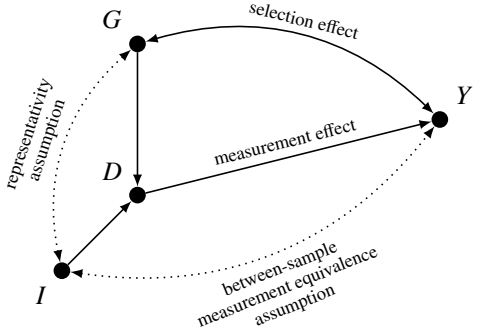
Indeed, this measurement effect reflects the difference in quality for the very same group of CATI and WSAQ respondents as if their data had been measured by CATI or WSAQ and CAPI respectively. Knowledge of this measurement effect would allow making the mixed-mode data consistent with the main ESS single-mode CAPI data by correct-



(a) When all modes are observed for all respondents, the mode group (G) and mode of data-collection (D) are independent, and no confounding between measurement and selection effects occurs.



(b) In a mixed-mode dataset, the mode group (G) and mode of data-collection (D) are equal (double arrow), and measurement and selection effects are completely confounded.



(c) An instrumental variable (I) breaks the prefect relation between the mode group (G) and the mode of data-collection (D), and measurement and selection effects become partly unconfounded. An instrumental variable, however, requires two assumptions about zero correlation (dotted lines).

Figure 1. The relations between variables in mixed-mode data can be represented by causal graphs (Pearl, 1995, 2009).

ing the CATI and WSAQ respondents' data. Additionally, it is tempting to calculate the conditional *Selection Effect*

$$\mathrm{SE}(q^2) = q^2_{p,p} - q^2_{p,tw}. \tag{2}$$

Indeed, this selection effect reflects the difference between CATI and WSAQ respondents and the CAPI respondents as if all data had been measured by CAPI's, the data-collection mode of the main ESS survey. Knowledge of this selection effect provides evidence of the usefulness of a mixed-mode design instead of a single-mode design. Indeed, a zero selection effect would mean that both groups of respondents are similar and, thus, that data of the CAPI group are sufficient for analysis. The sum of the measurement effect and the selection effect is equal to the *Overall mode Effect*, which is the difference between the observed responses of both modes:

$$\mathrm{OE}(q^2) = q^2_{p,p} - q^2_{tw,tw} = \mathrm{ME}(q^2) + \mathrm{SE}(q^2).$$

From (1) and (2) it is clear that estimation of the measurement and selection effects requires three groups of responses. The first group (p, p) contains responses collected by a CAPI from people who would actually select CAPI within the mixed-mode survey. The second group (tw, tw) contains responses collected by a CATI or a WSAQ from people who would actually select CATI or WSAQ within the mixed-mode survey. Both these groups of responses can straightforwardly be estimated from the mixed-mode data. The third group (p, tw) contains responses collected by a CAPI from people who would in fact select CATI or WSAQ within the mixed-mode survey. These responses, however, are not observed because, by definition, all CATI and WSAQ respondents complete the survey by CATIs or WSAQs and not by CAPIs. Indeed, within the observed data, the selected mode group G fully determines the mode of administration D for every respondent (Figure 1(b)). The quality $q^2_{p,tw}$ of the third group of responses is not observed and is called a *counterfactual* (Rubin, 1974; Holland, 1988; Galles & Pearl, 1998). This counterfactual is however required for the estimation of the measurement and selection effects.

3.2 Instrumental variable

The confounding between $\mathrm{ME}(q^2)$ and $\mathrm{SE}(q^2)$ as defined in (1) and (2) can partly be circumvented by simultaneously analysing the main ESS data and the mixed-mode

ESS data (Vannieuwenhuyze et al., 2010; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2012). Such a simultaneous analysis defines a new variable I which refers to the dataset of origin. Put differently, variable I takes either the value 'mixed-mode ESS' if a respondent was sampled for the mixed-mode ESS survey, or the value 'main ESS' if a respondent was sampled for the main ESS survey. In theory, I may relate with D, G, and Y (Figure 1(c)). A relation between I and D is straightforward since membership of the main ESS data implies that data are collected by CAPI. With respect to the other relations, assumptions can be made.

First, a relation between I and Y means that respondents would answer differently in a CAPI when they were sampled for the main ESS or the mixed-mode ESS. These differences are assumed to be zero. In other words, this assumption implies that CAPI measurement error is equivalent in the mixed-mode and the main ESS data. This assumption is further called the *between-sample measurement equivalence assumption* (Note that between-sample measurement equivalence should not be confounded with between-mode measurement equivalence, the latter being the topic of investigation within this paper). However, the validity of this assumption might not be guaranteed within the ESS round 4 data because there is a difference in CAPI administration between the main survey and the mixed-mode survey. In both surveys, the principal part of the ESS questionnaire was administered by using face-to-face interviews, but a supplementary part, including some of the analysis variables of this paper, was administered differently. Indeed, this supplementary part was also administered by face-to-face interviews in the mixed-mode survey, but it was administered by paper self-completion questionnaires in the main ESS survey. These paper questionnaires were given by the interviewer to the respondents at the end of the face-to-face interview but were taken back by the interviewer before he left the respondent's house. This difference between a face-to-face interview and an anonymous paper questionnaire might invalidate the between-sample measurement equivalence assumption, but it could easily have been circumvented by using one and the same data-collection strategy for the CAPI respondents in both surveys (for example by using a face-to-face interview for all questions).

Second, the correlation between I and G means that the mixed-mode ESS data and the main ESS data do not represent the same population. However, this correlation is assumed to be zero which means that both datasets do represent the same population. This assumption is called the *representativity assumption* and can be defended or offended by some arguments (Vannieuwenhuyze et al., 2010).

Firstly, we can theoretically assume that systematic coverage and nonresponse error are equal in both samples. Coverage is equal in both samples since both the main ESS round 4 and the mixed-mode sample are drawn from the same sampling frame. Nonresponse is assumed to be equal in both samples because it is well-known and generally observed that CAPI often results in high response rates relative to the other modes (de Leeuw, 2008). Consequently, a switch from a single-mode CAPI survey to a mixed-mode survey is probably mainly driven by the idea of lowering relative costs and sampling error rather than lowering non-response error. In other words, a switch from a single-mode CAPI survey to a mixed-mode survey theoretically assumes that the WSAQ and CATI respondents of the mixed-mode survey would also accept to participate by a CAPI when they were sampled for the main ESS data-collection.

Secondly, if the samples represent the same population, the response rates and respondent composition should be more or less equal (even though this statement is not necessarily true in the opposite direction). However, the response rate of the mixed-mode survey is remarkably significantly smaller than the response rate of the main ESS survey (see Table 1). This inequality is probably caused by the absence of a CAPI follow-up for all sample members of the mixed-mode survey who chose to participate by WSAQs but did not respond afterwards. This inaccuracy in design implementation might explain the difference in response rates and probably means that the representativity assumption is not completely valid, but could easily have been circumvented by an appropriate design. The respondent composition on several socio-demographic variables (a cross-classification of age and gender, urbanization, household size, and education), in contrast, did not yield any significant differences between both samples (see Eva et al., 2010). These insignificant differences can be used as an argument enforcing the representativity assumption. Nonetheless, the small differences between both datasets are corrected by using normalized inverse propensity score weights derived from the complete set of variables mentioned above (Rosenbaum & Rubin, 1983; Sato & Matsuyama, 2003). As a consequence, respondent composition on these socio-demographical characteristics is equal in both datasets.

If both the between-sample measurement equivalence assumption and the representativity assumption hold, variable I is called an instrumental variable (Bowden & Turkington, 1990; Angrist, Imbens, & Rubin, 1996; Heckman, 1996, 1997), and allows estimating the measurement and selection effects as defined in (1) and (2) (Vannieuwenhuyze et al., 2010). Indeed, under both assumptions, the main ESS data includes two groups of responses, i.e. the responses of the CAPI respondents group and the responses of the CATI and WSAQ respondents group if all data had actually been collected by CAPIs. The data of the main ESS, denoted by $Y_{p,.} = (Y|D = p)$, thus follows a mixture distribution of both groups of responses, i.e. $Y_{p,p}$ for the CAPI group and $Y_{p,tw}$ for the CATI and WSAQ group:

$$\mathrm{P}(Y_{p,.}) = p_p \ \mathrm{P}(Y_{p,p}) + p_{tw} \ \mathrm{P}(Y_{p,tw}), \tag{3}$$

where $p_p = \mathrm{P}(G = p)$ and $p_{tw} = \mathrm{P}(G = tw)$, the proportions of respondents selected for mode group p and tw respectively within the mixed-mode design. Four of the five quantities in (3) can directly be estimated from the available data. The distribution $\mathrm{P}(Y_{p,.})$ can be estimated from the main ESS data while the distribution $\mathrm{P}(Y_{p,p})$ can be estimated from the CAPI respondents of the mixed-mode ESS data. The proportions p_p and p_{tw}, in turn, can be estimated from the entire mixed-mode ESS data. As a result, the last remaining distribution

$P(Y_{p,tw})$ can be estimated as well. This distribution, however, allows the calculation of the counterfactual quality $q^2_{p,tw}$ in (1) and (2).

3.3 The multitrait-multimethod model to estimate quality

One way to examine and estimate the quality of data obtained by a particular data-collection mode is the analysis of multitrait-multimethod (MTMM) experiments (Campbell & Fiske, 1959; Andrews, 1984; Scherpenzeel, 1995; Scherpenzeel & Saris, 1997; Revilla, 2010). An MTMM experiment starts from the repetition of particular questions or *traits* within the survey questionnaire, but using a different *method* each time (e.g. a different response scale).

The mixed-mode ESS round 4 survey includes four MTMM experiments which are also included in the main ESS round 4 survey (Saris & Gallhofer, 2002). These experiments each contain two or three questions or traits measured by two distinct response scales or methods (Table 2). The first set of traits includes questions about media usage and was measured by an 8-point scale and in hours and minutes respectively. The next two sets of traits include questions about social trust and political trust and were measured by an 11-point and a 6-point scale, respectively. The last set of traits includes questions about satisfaction and was measured by 11-point scales using extreme and normal labels on the end points of the scales. In order to minimize memory effects, similar questions were asked at the beginning and the end of the questionnaire which guarantees a time gap of at least 20 minutes (Van Meurs & Saris, 1990).

Two additional remarks are required. First, as already noted in Section 3.2, there is a difference between the mixed-mode and the main ESS survey with respect to the CAPI implementation. Within the main ESS the supplementary part of the questionnaire was administered by paper self-completion questionnaires while the regular face-to-face interview was used in the mixed-mode survey. In practice, the supplementary part included all second methods M_2 of the MTMM experiments. Second, within the main ESS survey, the MTMM experiments were randomly assigned to approximately one third of the respondents in order to reduce the questionnaire length. This means that the actual response frequencies used within the analyses are much lower than the total response frequency of the main ESS (see the footnote of Table 1).

Data quality can be examined by MTMM experiments through the estimation of the relative reliability and validity of the different trait and method combinations using structural equation models (SEM's) (Werts & Linn, 1970; Jöreskog, 1970; Alwin, 1974; Andrews, 1984). The relationships between the different traits measured by the different methods are modelled on the *True Score model* (Saris & Andrews, 1991; Saris & Gallhofer, 2007):

$$\begin{aligned} Y_{ij} &= r_{ij}T_{ij} + e_{ij}, \text{ and} \\ T_{ij} &= v_{ij}F_i + m_{ij}M_j, \end{aligned} \tag{4}$$

where Y_{ij} refers to the observed variable from trait i and method j, T_{ij} refers to the intended True score of Y_{ij}, F_i refers to trait i (also called Factor i), and M_j refers to Method j (See Figure 2 for an example). The Y_{ij}'s are observed or manifest variables, while the T_{ij}'s, F_i's, and M_j's are unobserved or latent variables.

The true scores T_{ij} correspond to the systematic components of the observed variables Y_{ij}, i.e. after correction for a random observational error e_{ij}. The square of the standardized effect of the true score T_{ij} on the observed variable Y_{ij} is called the *reliability* r^2_{ij} of question Y_{ij}. The true scores themselves depend on both the traits and the methods. The square of the standardized effect of the trait F_i on the true score T_{ij} is called the *validity* v^2_{ij} of Y_{ij}. The effect of the method M_j on the true score T_{ij} is called the method effect coefficient m_{ij}. It is further assumed that the random errors e_{ij} and the methods M_j are not correlated with each other or with the traits, and that the effects of the methods on the different traits are equal (i.e. $m_{ij} = m_{i'j}$ for all i, i', and j).

The quality q^2_{ij} of Y_{ij} is now defined as $q^2_{ij} = r^2_{ij}v^2_{ij}$. Comparing the quality of each trait and method combination across the modes allows examining the relative effect of a data-collection mode on data quality.

3.4 Quality estimation with an instrumental variable

Let $\boldsymbol{Y}$ represent all observed variables Y_{ij} from a particular MTMM experiment and let $\boldsymbol{q^2}$ represent the vector of all quality indicators q^2_{ij} of this MTMM experiment calculated on $\boldsymbol{Y}$. Under the regular assumption of multivariate normality, a sufficient statistic to calculate $\boldsymbol{q^2}$ is the covariance matrix S of $\boldsymbol{Y}$. A nice feature of the mixture distribution in (3) is that the covariance matrix of the counterfactual (p, tw) group of responses can easily be calculated (Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000):

$$\mathsf{S}_{p,tw} = \frac{1}{p_{tw}}\mathsf{S}_{p,\cdot} - \frac{p_p}{p_{tw}}\mathsf{S}_{p,p} - \frac{p_p}{p^2_{tw}}(\bar{\boldsymbol{Y}}_{\boldsymbol{p,p}} - \bar{\boldsymbol{Y}}_{\boldsymbol{p,\cdot}})(\bar{\boldsymbol{Y}}_{\boldsymbol{p,p}} - \bar{\boldsymbol{Y}}_{\boldsymbol{p,\cdot}})', \tag{5}$$

where $\bar{\boldsymbol{Y}}_{\boldsymbol{p,\cdot}}$ and $\bar{\boldsymbol{Y}}_{\boldsymbol{p,p}}$ denote the mean vectors of $\boldsymbol{Y}_{p,\cdot}$ and $\boldsymbol{Y}_{p,p}$. All quantities at the right-hand side of (5) can directly be estimated from the data as discussed in Section 3.2.

The analysis of the MTMM experiments further requires the sample size of each group of responses. The relative sample size of $\mathsf{S}_{p,tw}$ is, of course, not readily available but is calculated as the weighted sum

$$n_{p,tw} = \frac{n^2_{p,\cdot} + n^2_{p,p}}{n_{p,\cdot} + n_{p,p}},$$

where $n_{p,\cdot}$ and $n_{p,p}$ represent the sample sizes of the main ESS survey and the CAPI group in the mixed-mode ESS survey respectively.

Starting from the covariance matrices, the reliability and validity coefficients are obtained by LISREL starting from the covariance matrices obtained for all three groups of responses, i.e. group (p, p), (tw, tw), and (p, tw). However, model (4) usually requires at least three methods in order to

Table 2 The ESS includes MTMM experiments about four topics, each including two methods and two or three traits.

Topic	Traits	M_1[a] Method 1	M_2[b] Method 2
Media	On an average weekday, how much time, in total: F1: do you spend watching television? F2: do you spend listening to the radio? F3: do you spend reading the newspapers?	8 points	Hours and min.
Social trust	F1: Generally speaking would you say that most people can be trusted or that you can't be too careful in dealing with people? F2: 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?	11 points	6 points
Political trust	How much do you personally trust each of the institutions: F1: Dutch parliament? F2: The legal system? F3: The police?	11 points	6 points
Satisfaction	How satisfied are you with: F1: the present state of the economy in NL? F2: the way the government is doing its job? F3: the way democracy works?	extreme labels	normal labels

[a] The first method M_1 is part of the principal questionnaire which is always administered by a face-to-face interview.
[b] The second method M_2 is part of the supplementary questionnaire which is administered by a face-to-face interview in the mixed-mode ESS survey but by a paper self-completion questionnaire in the main ESS survey.

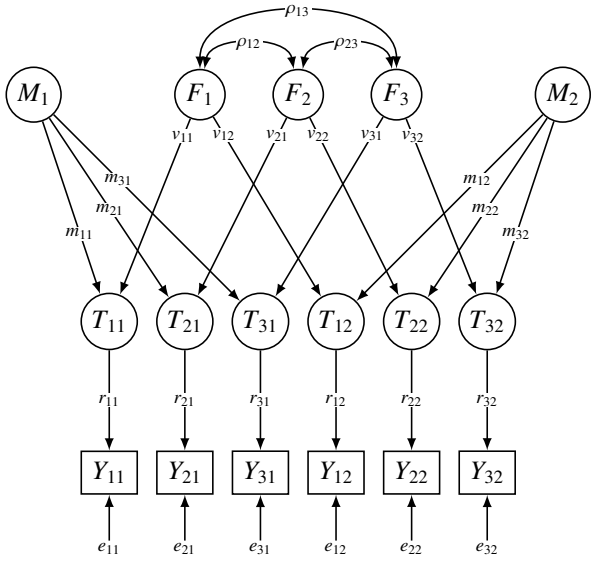


Figure 2. An MTMM experiment can be represented by causal graphs with six observed outcome variables Y, six latent true scores T, two latent methods M, and three latent traits F.

avoid identification problems while the ESS MTMM experiments only include two methods for each trait. Nevertheless, a multi-group analysis with parameter equality constraints across the different groups allows identification of the model (Saris, Satorra, & Coenders, 2004).

All parameters are first constrained to be equal across the groups. Subsequently, these constraints are step-wise removed by using the JRule software (which is described in Saris, Satorra, & Van der Veld, 2009) until an acceptable model is obtained. The JRule software uses modification indices (MI), the Expected Parameter Chance (EPC), and statistical power to detect misspecified parameters. This software thus has the advantage of detecting misspecified parameters while taking both type 1 and type 2 errors into account. Based on the JRule output, constraints on misspecified parameters can be relaxed within LISREL step by step until the obtained model sufficiently fits the data. As a result, differences in quality estimates across the modes of data-collections (i.e. measurement effects) and the mode groups (i.e. selection effects) can be observed. The appendix provides a list of all the corrections made to the final model.

4 Results

For most items, differences between the response groups seem to be minor for the reliability coefficients, the validity coefficients, and the quality estimates, as well as for the average quality over the different traits for a given method (i.e. $\bar{q}^2 = E_i(q_{ij}^2)$) (Table 3). These minor differences indicate the absence of mode effects. For other items, differences are remarkably large. Most large differences can be found among the reliability and validity coefficients of media usage. However, these large differences may be caused by a difficult identification of the media usage model, which may lead to estimates being very sensitive to model corrections.

The overall mode effects on quality OE(q^2) are generally rather small or fair, except for the Media items and one of the Political Trust items (Table 4). The most extreme overall mode effect is found on the question about trust in the Dutch parliament on an 11-point scale. The quality of this item is 0.30 higher for the CATI and WSAQ respondents' answers compared to the CAPI respondents' answers. Nevertheless, the average overall mode effects within the different traits are fair and in line with previous research (Revilla, 2010; Revilla & Saris, 2012). Remarkably, the signs of the overall mode effects are inconsistent within the methods of media usage. These inconsistencies mean that the quality of the CAPI respondents' answers scores better for some items but worse for others. The reason for these inconsistencies is not clear though.

The central question is whether the overall mode effects are caused by a real effect of the data-collection mode on data quality, i.e. a measurement effect, or by a different respondent composition, i.e. a selection effect. The results show fairly low measurement and selection effects for Social trust and Satisfaction, but large effects occur for some Media and Political trust items (Table 4).

With respect to measurement effects (ME(q^2)), most of the effects are negative but some are positive. Positive measurement effects mean that CAPI provides better quality than the CATI and WSAQ combination for people who are actually selected for CATI or WSAQ in the mixed-mode experiment, while the opposite is true for negative measurement effects. However, the effects are generally low, except for the fairly large differences in time spent watching television in hours and minutes and in trust in the Dutch parliament on a 6-point scale. It can thus be concluded that the data-collection modes in general do not have a large effect on data quality but special attention might be given to some questions and response scales. This conclusion is an argument in favour of further use of mixed-mode survey data-collections even though special care for particular survey contents might be required.

Most of the selection effects (SE(q^2)) are negative as well, which suggests better quality for the CATI and WSAQ respondents relative to CAPI respondents when all data are actually measured by CAPI. However, the effects are also generally low, except for the large differences in time spent watching television and reading newspapers in hours and minutes, trust in the Dutch parliament on an 11-point scale, and trust in the legal system on a 6-point scale. Large selection effects might principally explain the large overall mode effect on some items, which means that previous research may have overestimated the real impact of a mode on data quality by merely considering the overall mode effect instead of the measurement effect. The occurrence of selection effects may further suggest an advantage of using a mixed-mode design instead of a single-mode CAPI design because different groups of respondents are selected for the different modes.

5 Discussion

This paper used an instrumental variable to disentangle measurement effects from selection effects on MTMM quality estimates in the ESS round 4 mixed-mode survey data. The results of the analysis show low or fair measurement effects while some selection effects are large. In general, overall mode effects are thus mainly caused by differences in respondent composition across the modes. These results, however, are preliminary because many problems were encountered during the analyses. Nonetheless, this section will argue that many of these problems are caused by design deficiencies which can easily be avoided in future studies. The main aim of this section and, by extension, this paper is to provide guidelines for future mixed-mode survey research to allow more accurate research on mode effects in mixed-mode data.

First, using an instrumental variable only allows comparing CAPI with a combination of CATI and WSAQ within the ESS round four mixed-mode experiment. However, the CATI and WSAQ combination is somewhat strange as large mode effects can be expected between these two modes as well. Especially the fact that CATI is an interview mode while WSAQ is a self-administration mode may cause considerable differences. Moreover, previous research already

Table 3 For most items, minor differences are observed between the response groups for the standardized reliability coefficients (r), the standardized validity coefficients (v), and quality estimates (q) .

		r_{1j}	r_{2j}	r_{3j}	v_{1j}	v_{2j}	v_{3j}	q^2_{1j}	q^2_{2j}	q^2_{3j}	$\bar{q}^2$
Media:											
p, p	M_1	.94	.80	1.0	.97	.98	.94	.83	.61	.88	.78
	M_2	.84	.91	.82	.93	.99	.78	.61	.81	.41	.61
p, tw	M_1	.99	.78	1.0	.97	.98	.89	.92	.58	.79	.77
	M_2	.52	.91	.46	.93	.99	.55	.23	.81	.06	.37
tw, tw	M_1	1.0	.82	.92	.97	.98	.89	.94	.65	.67	.75
	M_2	.71	1.0	.84	.90	.95	.49	.41	.90	.17	.49
Social trust:											
p, p	M_1	.84	.82		1.0	1.0		.71	.67		.69
	M_2	.90	.77		.88	.80		.63	.38		.50
p, tw	M_1	.90	.83		.97	.96		.76	.63		.70
	M_2	.91	.77		.81	.82		.54	.40		.47
tw, tw	M_1	.90	.82		1.0	1.0		.81	.67		.74
	M_2	.90	.87		.88	.80		.63	.48		.56
Political trust:											
p, p	M_1	.76	.92	.96	.86	.89	.95	.43	.67	.83	.64
	M_2	.88	.99	.86	.91	.94	.91	.64	.87	.61	.71
p, tw	M_1	.93	.95	1.0	.99	.89	.97	.85	.71	.94	.83
	M_2	.95	.92	.88	.92	.88	.82	.76	.66	.52	.65
tw, tw	M_1	.94	.95	1.0	.91	.89	.97	.73	.71	.94	.80
	M_2	.86	.91	.87	.91	.94	.88	.61	.73	.59	.64
Satisfaction:											
p, p	M_1	.81	.92	.97	.98	.99	.97	.63	.83	.89	.78
	M_2	.97	.94	.91	.90	.90	.89	.76	.72	.66	.71
p, tw	M_1	.81	.97	.97	.98	.99	.97	.63	.92	.89	.81
	M_2	.97	.94	.91	.85	.90	.89	.68	.72	.66	.68
tw, tw	M_1	.83	.97	.97	.98	.99	.97	.66	.92	.89	.82
	M_2	.97	.94	.91	.90	.90	.89	.76	.72	.66	.71

For an overview of the Traits, Methods, and coefficients, refer to Table 2 and Figure 2. For an overview of the model constraints, refer to the appendix. $\bar{q}^2$ represents the average total quality over the different traits.

Table 4 The overall mode effects (OE), measurement effects (ME), and selection effects (SE) on quality are generally low to moderate.

		OE(q^2)				ME(q^2)				SE(q^2)			
		q^2_{1j}	q^2_{2j}	q^2_{3j}	$\bar{q}^2$	q^2_{1j}	q^2_{2j}	q^2_{3j}	$\bar{q}^2$	q^2_{1j}	q^2_{2j}	q^2_{3j}	$\bar{q}^2$
Media:													
	M_1	−.11	−.04	.21	.03	−.02	−.07	.12	.02	−.09	.03	.09	.01
	M_2	.20	−.09	.24	.12	−.18	−.09	−.11	−.12	.38	.00	.35	.24
Social trust:													
	M_1	.10	.00		−.05	−.05	−.04		−.04	−.05	.04		−.01
	M_2	.00	−.10		−.06	−.09	−.08		−.09	.09	−.02		.03
Political trust:													
	M_1	−.30	−.04	−.11	−.15	.12	.00	.00	.04	−.42	−.04	−.11	−.19
	M_2	.03	.13	.03	.06	.15	−.08	−.07	.00	−.12	.21	.09	.06
Satisfaction:													
	M_1	−.03	−.09	.00	−.04	−.03	.00	.00	−.01	.00	−.09	.00	−.03
	M_2	.00	.00	.00	.00	−.08	.00	.00	−.03	.08	.00	.00	.03

suggested some differences between CATI on the one hand, and CAPI and WSAQ on the other hand (Revilla, 2010). This forced CATI and WSAQ combination can only be avoided by using mixed-mode designs only including one single additional mode apart from CAPI.

Second, the validity of the analysis largely depends on the validity of the assumptions. The between-sample measurement equivalence assumption, on the one hand, is probably violated because CAPI has been implemented differently within the main ESS and the mixed-mode ESS survey for the administration of the supplementary questionnaire. This problem can easily be avoided by exactly the same data-collection strategy for CAPI in both surveys. The representativity assumption, on the other hand, can be doubted as well, because it requires the mixed-mode experiment data and the main ESS data to represent the same population. Some control on this assumption can be achieved by proper implementation of the survey design. This, for example, requires that, in contrast to the ESS round four mixed-mode survey, nonresponding WSAQ choosers are followed up by CAPI.

Third, the small sample size of both the main ESS and the mixed-mode ESS survey might invalidate the results because of sampling errors. Especially the media experiment results might be problematic because its model was hardly identified and the estimates were very sensitive to model corrections. As a result, small corrections might have led to large differences in the final outcomes.

Fourth, because the ESS MTMM experiments include too few methods, multi-group structural equation models are used in order to identify the parameters from the true score MTMM model. Such multi-group models, however, assume that all groups are independent, but this requirement is not met because the covariance matrices $\mathsf{S}_{p,p}$ and $\mathsf{S}_{p,tw}$ both depend on the answers of the CAPI respondents in the mixed-mode ESS data. This defect might result in wrongly specified constraints in the MTMM models across the different groups, but its impact is probably not severe since the JRule software also takes power and parameter sizes into account when dropping constraints. Nevertheless, this problem will probably be avoided when more methods are included in the MTMM experiments.

Further, three additional comments can be made. First, the analysis data stem from the Netherlands, an adequate country for WSAQs because of high Internet coverage (around 85%, see www.internetworldstats.com). In other countries with low Internet coverage, the use of mixed-mode surveys including WSAQ might be more problematic. Repetitions of the current study in other countries might thus be required before general statements about the usefulness of mixed-mode survey designs are made. Second, the analysis data stem from four particular MTMM experiments. These MTMM topics may hardly represent all possible survey topics. Research on other topics may be required in future studies. Third, besides the instrumental variable method, alternative methods exist which allow disentangling mode effects. These methods include the back-door method, which starts from covariates explaining the selection effects, and the front-door method, which starts from covariates explaining the measurement effects (Pearl, 2009). Nevertheless, both the back-door and front-door method require appropriate control variables which are probably not present in the ESS round 4 datasets.

To conclude, the results of this study are preliminary and can be validated by the development of appropriate mixed-mode designs, mixed-mode survey implementations, and inferential techniques. Future mixed-mode design implementations should, for example, control for necessary analysis assumptions, and future inferential techniques should focus on dependencies across groups. The results of this study should therefore not stop researchers from moving to new survey data-collection approaches. It is first recommended to repeat and validate this study in improved ways on other topics, with sensitive and complex questions, and within populations other than the Netherlands.

References

Alwin, D. F. (1974). Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (Ed.), *Sociological methodology.* San Francisco: Jossey-Bass.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409–442.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444-455.

Bowden, R. J., & Turkington, D. A. (1990). *Instrumental variables*. Cambridge: Cambridge University press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *6*, 81–105.

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2), 233–255.

de Leeuw, E. D. (2008). Choosing the method of data collection. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 113–135). New York: Erlbaum.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, 427–444.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail and mixed-mode surveys: the tailored design method* (3rd ed.). Hoboken: Wiley.

Eva, G., Loosveldt, G., Lynn, P., Martin, P., Revilla, M., Saris, W. E., & Vannieuwenhuyze, J. T. A. (2010). *ESS prep WP6 – Mixed mode experiment. delivarable 21 final mode report.* London: City University.

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*. New York (N.Y.): Springer.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *1*, 151–182.

Heckman, J. J. (1996). Randomization as an instrumental variable. *The Reviews of Economics and Statistics*, *78*(2), 336–341.

Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources*, *32*(3), 441–462.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, *18*, 449–484.

Izrael, D., Hoaglin, D. C., & Battaglia, M. P. (2000). A sas macro for balancing a weighted sample. In *Proceedings of the twenty-fifth annual sas users group international conference.* (Paper 275)

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239–251.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York (N.Y.): Wiley.

Morgan, S. L., & Winship, C. (2009). *Counterfactuals and causal inference: methods and principles for social research*. New York, (N.Y.): Cambridge university press.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). New York (N.Y.): Cambridge University Press.

Revilla, M. (2010). Quality in unimode and mixed-mode designs: A multitrait-multimethod approach. *Survey Research Methods*, *4*(3), 151–164.

Revilla, M., & Saris, W. E. (2012). A comparison of the quality of questions in a face-to-face and a web survey. *International Journal of Public Opinion Research*. Retrieved 10 May 2013, from http://ijpor.oxfordjournals.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, *66*(5), 688–701.

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modelling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (1st ed., pp. 575–597). New York (N.Y.): Wiley-Interscience publications.

Saris, W. E., & Gallhofer, I. (2002). *Report on the MTMM experiments in the pilot studies and proposals for round 1 of the ESS.* Retrieved 26/11/2013, from http://www.europeansocialsurvey.org/docs/methodology/ESS1_quality_measurement.pdf

Saris, W. E., & Gallhofer, I. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. New York: Wiley-Interscience publications.

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*(1), 311–347.

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.

Sato, T., & Matsuyama, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology*, *14*, 680–686.

Scherpenzeel, A. (1995). Meta analysis of a European comparative study. In W. E. Saris & A. Münnich (Eds.), *The multitrait-multimethod approach to evaluate measurement instruments* (pp. 225–242). Budapest: Eötvös University Press.

Scherpenzeel, A., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods and Research*, *25*(3), 341–383.

Van Meurs, L., & Saris, W. E. (1990). Memory effects in MTMM studies. In W. E. Saris & L. Van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies* (p. 134-146). Amsterdam: North Holland.

Vannieuwenhuyze, J. T. A., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed mode surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045.

Vannieuwenhuyze, J. T. A., Loosveldt, G., & Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. *International Statistical Review*, *80*(2), 306–322.

Voogt, R. J., & Saris, W. E. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics*, *21*(3), 367-387.

Weisberg, H. F. (2010). *Bias and causation: Models and judgment for valid comparisons*. Hoboken, New Jersey: Wiley.

Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, *74*, 194–212.

Appendix

This appendix provides a list of corrections made to the initial constraint model for all four experiments in LISREL. The groups are ordered as in the LISREL input. To free a parameter in group (p, p), it is first set free in (p, tw) and, additionally, an equality constraint between groups (p, tw) and (tw, tw) is added.

- Media:
  - group (p, p): set e_{22} free, set ρ_{M_1,M_2} free, fix e_{31} to 0
  - group (p, tw): set e_{11}, e_{21}, e_{12}, and e_{32} free, set v_{31} and v_{32} free, fix ρ_{M_1,M_2} to 0
  - group (tw, tw): set e_{21}, e_{31}, e_{12}, and e_{32} free, set φ_{M_2}, ρ_{M_1,M_2}, and m_{12} free, fix e_{11} and e_{22} to 0, equalize v_{31} to group (p, tw)
- Social trust:
  - group (p, p): fix φ_{M_1} to 0
  - group (p, tw): set φ_{M_1}, φ_{M_2}, m_{12}, and e_{11} free
  - group (tw, tw): set e_{22} free, equalize e_{11} to group (p, tw)
- Political trust:
  - group (p, p): set m_{11} and m_{31} free
  - group (p, tw): set m_{11}, m_{12}, m_{31}, v_{11}, v_{21}, e_{11}, e_{21}, e_{12}, e_{22}, and φ_{M_2} free, fix e_{31} to 0
  - group (tw, tw): set m_{32}, e_{11}, e_{21}, and e_{12} free, fix e_{31} to 0, equalize e_{22}, m_{11} and m_{31} to group (p, tw)
- Satisfaction:
  - group (p, p): set m_{31} and $\rho_{e_{12},e_{11}}$ free
  - group (p, tw): set v_{11}, m_{12}, and e_{21} free, fix $\rho_{e_{12},e_{11}}$ to 0
  - group (tw, tw): set e_{11} free, fix $\rho_{e_{12},e_{11}}$ to 0, equalize e_{21} and v_{11} to group (p, tw)