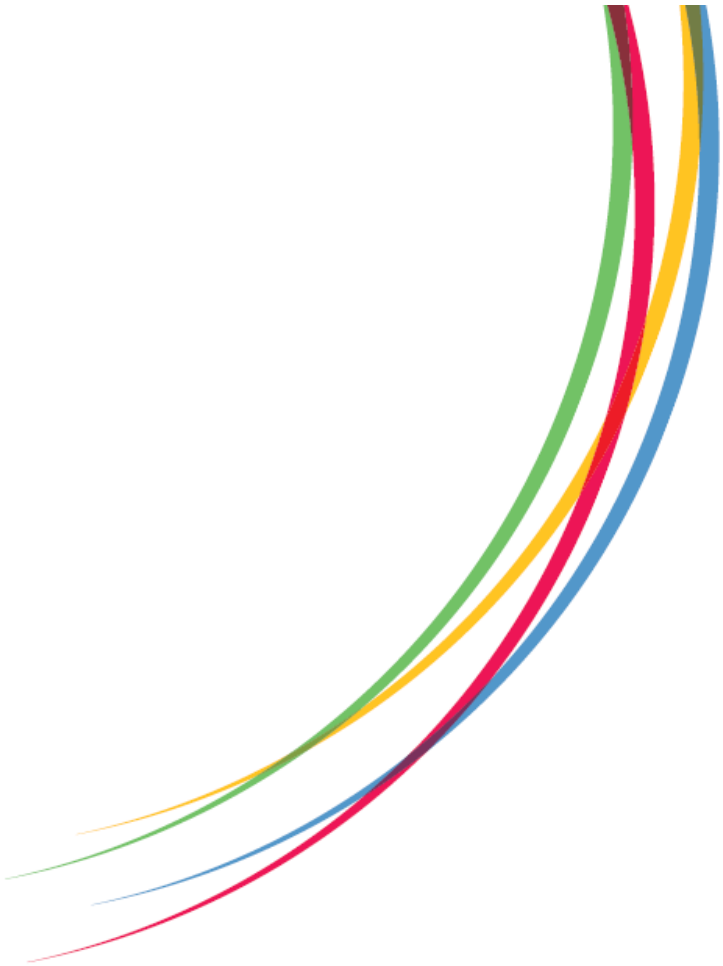


Understanding Society

Working Paper Series

No. 2015 – 01

May 2015

Using equivalence testing to disentangle selection and measurement in mixed modes surveys

Alexandru Cernat

Institute for Social and Economic Research, University of Essex

Non-Technical Summary

Questionnaires can be administered using various mediums, from face-to-face interviews to self-administration on the internet. This opens up the possibility of mixing two or more modes of interview both for the same respondent (e.g., at different points in time) and across individuals (e.g., offering the possibility of answering by web to some respondents) in order to reduce costs and/or decrease non-response bias.

But to evaluate the utility of mixed modes designs we must disentangle the systematic tendency of selecting a mode (e.g., older people preferring face to face interviews) from measurement differences (e.g., people answering by web are more honest than those answering by telephone). In this paper a new approach to separating these two effects is put forward. A small simulation study using the SF12 health scale is conducted in order to show how the method works and what are the possible limitations.

The method proposes using equivalence testing, a statistical way of evaluating how similarly scales are measured across groups, to control for potential measurement differences between modes. This will make it possible to calculate the selection effect (i.e., tendency of selecting a certain mode). The simulation study shows that the method gives unbiased estimates as long as the two main assumptions, isolation and exhaustiveness, hold. The potential of the approach and the possibilities for future development are discussed.

Using equivalence testing to disentangle selection and measurement in mixed modes surveys

Alexandru Cernat*
Institute for Social and Economic Research, University of Essex

*Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK (email: acerna@essex.ac.uk)

Abstract

Mixed modes are becoming increasingly popular in surveys. This approach can decrease costs and non-response bias. But in order to evaluate the utility of this approach we must separate selection and measurement effects of the different modes. In this paper I propose a new way of applying the front-door method to control for measurement differences between modes: equivalence testing with latent measurement models. A small simulation study will show how this approach works and how it can be biased if the assumptions of exhaustiveness and isolation are not true in the observed data.

Key words: mixed modes survey, causality, latent measurement models, equivalence testing.

JEL Codes: C81, C83

Acknowledgements: I would like to thank all the people that helped with this paper: Peter Lynn, Paul Clarke, Daniel Oberski and Jorre Vannieuwenhuyze. This work was supported by a +3 PhD grant awarded by the UK Economic and Social Research Council.

Introduction

Using multiple modes of interview (i.e., face to face, telephone, web) to conduct surveys is increasingly popular as it can potentially lower costs while minimizing non-response bias (De Leeuw, 2005). But despite the increased popularity of mixed mode surveys there is still an acute need for methods to evaluate such designs. In order to gauge their effectiveness it is essential to separate the effects of modes on selection and measurement. Only then is it possible to investigate if the additional modes, usually more expensive, manage to include different types of individuals and make the overall sample more representative. Additionally, identifying the measurement effects of the different modes can inform design decisions.

Most of the literature in mixed modes research has used multiple items to control for different selection propensity in modes in order to estimate measurement effects, also know as the *back-door method* (Morgan and Winship, 2007; Pearl, 2009; Vannieuwenhuyze et al., 2014). Recently, a different approach has been put forward, which aims to control for mode differences in measurement, known as the *front-door method* (Morgan and Winship, 2007; Pearl, 1995, 2009; Vannieuwenhuyze et al., 2014). This approach may prove an important development as situations can be envisaged where good back-door variables are not available but front-door ones are. Furthermore, considerable research and theory has been developed to estimate and explain measurement differences between modes. This knowledge can be fruitfully applied to the front-door method. While this approach has great potential, it hinges on the ability to find new variables that are able to control for measurement differences across modes.

The present paper will propose a new way to separate selection and measurement in mixed mode research by utilizing equivalence testing as a front-door method. While testing for equivalence has been previously used in the mixed mode literature (Cernat, 2014; Gordoni et al., 2011; Heerwegh and Loosveldt, 2011; Hox et al., 2015; Klausch et al., 2013; Révilla, 2013; Vannieuwenhuyze and Révilla, 2013) it has been usually implemented to estimate measurement differences between modes after controlling for selection. The potential of this approach as a front-door method for estimating selection mode effects on a latent variable has been ignored so far. It is this point that this paper will elaborate on.

In order to show the potential of this method and it's assumptions the next two sections will present the main theoretical background of causal models and equivalence testing. Next, a simulation study will exemplify the method and the potential bias when assumptions do not hold. Finally, conclusions and limitations will be discussed.

Causal models and mixed modes

The fundamentals for the current discussion of causal analysis is based on the counterfactual model which stipulates the existence of multiple causal states to which the population of interest could be exposed. In the simple case of a mixed mode design with two modes each individual could answer either in the first mode, m_1, or in the second one, m_2. Using the notation of Vannieuwenhuyze et al. (2014) this will be denoted by D and is called *mode of data collection*. Nevertheless, in a survey each respondent participates only using one mode, the *mode group*, denoted by G_δ (where δ stands for the design used). Figure 1 graphically

presents this situation. In the ideal counterfactual data we would have both D and G_δ and they would not be related (situation a). Unfortunately, most of the real data has only one observation per individual and thus the two variables can't be distinguished (situation b).

Usually, the interest lies with the mean of a variable in the reference mode: $\mu_{m_1} = E(Y|D = m_1)$. Nevertheless, calculating this is not possible with observed data as it requires counterfactual information:

$$\mu_{m_1} = \mu_{m_1 m_1}\tau_{m_1} + \mu_{m_1 m_2}\tau_{m_2} \tag{1}$$

where μ_{dg} is the conditional mean $E(Y|D = d, G_\delta = g)$. In this equation $\mu_{m_1 m_1}$ can be observed in the data as the people that answered using m_1 while $\mu_{m_1 m_2}$ is a counterfactual as it represents what would the respondents from m_2 would have answered had they participated in m_1. Here τ_g represents the propensity to answer in each group: $P(G_\delta = g)$.

Using this notation we can estimate the selection and the measurement effects:

$$S_{m_1}(\mu) = \mu_{m_1 m_1} - \mu_{m_1 m_2} \tag{2}$$

$$M_{m_1}(\mu) = \mu_{m_2 m_2} - \mu_{m_1 m_2} \tag{3}$$

The selection effect, $S_{m_1}(\mu)$, would be different from zero only if the people in the two modes would have different different means had they all answered in m_1. Similarly, the measurement effect, $M_{m_1}(\mu)$, is given by the difference between the respondents in m_2 and those in m_1 if they had answered in the second mode. These formulae highlight the importance of estimating the counterfactual in separating selection and measurement, this being essential for the evaluation of mixed mode designs.

Using only the observed data does not enable the estimation of the two types of mode effects. As a result, a series of models have been put forward in order to estimate the counterfactuals. The causal literature has presented three main techniques: instrumental variables, the back-door approach and the front-door approach (Morgan and Winship, 2007; Pearl, 2009; Vannieuwenhuyze et al., 2014). The focus here will be on the latter two.

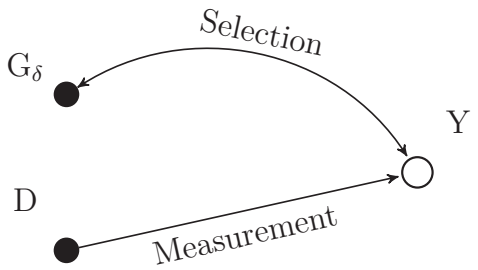
The *back-door method* aims to use a series of covariates (B in Figure 1e) that explain both the variable of interest, Y, and the survey mode (G_δ). It has been shown that by controlling for such variables it will be possible to calculate the counterfactual $\mu_{m_1 m_2}$ and, thus, calculate the measurement effect (Morgan and Winship, 2007; Pearl, 2009; Vannieuwenhuyze et al., 2014).

While this technique has been used repeatedly in the mixed mode field it does have two important assumptions. The first one is the *ignorable mode selection assumption*. This implies that the B variables will capture the entire relationship between mode and the variable of interest Y (i.e., selection effect into survey mode). When this assumption does not hold the estimates of selection and measurement effects of mode on Y will be biased as they will still be confounded with selection on unmeasured B variables. The second assumption is the *mode insensitivity assumption*. This means that there is no relationship between B and D. In practice this implies that the measurement of the controlling variables is not influenced by the mode of measurement.

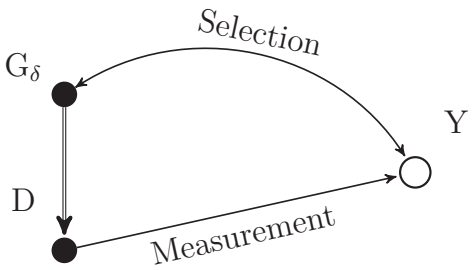
The back door has been applied in the mixed mode literature multiple times using techniques such as regression (e.g. Jäckle et al., 2010), matching (e.g. Lugtig et al., 2011), weighting (e.g. Hox et al., 2015) and controlling for covariates in Structural Equation Modelling (e.g., Heerwegh and Loosveldt, 2011).

Figure 1: Counterfactual models for separating selection and measurement in a mixed mode design. Adapted from Vannieuwenhuyze et al. (2014).

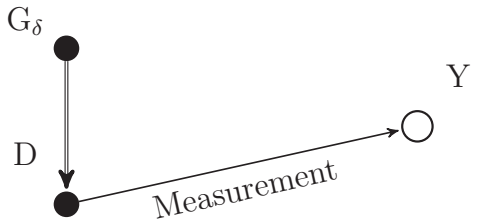
a. Counterfactual data



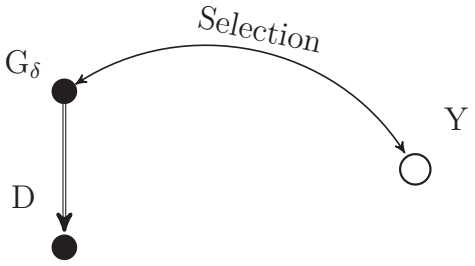
b. Mixed mode data



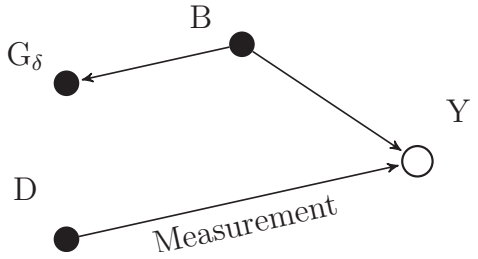
c. No selection



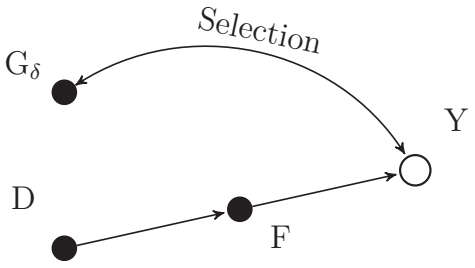
d. No measurement



e. Back door



f. Front door



Another approach to separating selection and measurement in mixed mode designs is the *front-door method* (Morgan and Winship, 2007; Pearl, 1995, 2009; Vannieuwenhuyze et al., 2014). Here the aim is to find a set of variables F (Figure 1f) that explain the measurement effect of the mode on the variable of interest.

As with the previous approach the front-door also makes a number of assumptions. The first one is the *exhaustiveness assumption*. This implies that the F-variables capture the entire causal effect of D on Y. If this is not true, part of the estimated selection differences will include differential measurement. Then, the *isolation assumption* requires that F is independent of G_δ; if it does not hold, then F will also include part of the selection effect.

The front-door approach is relatively new in the causal literature and has been rarely used in the mixed mode field (Vannieuwenhuyze et al., 2014). Although the assumptions of the method are similar to those of the back-door the variables used in the two procedures to separate selection and measurement are very different. Increasing the use of the front-door will hinge on finding appropriate variables to control for measurement differences. Raising awareness of this procedure and developing new ways to implement it in the field of mixed modes will provide researchers with new tools to evaluate surveys that combine multiple modes. Next, we turn to latent models and how they can be used to estimate and correct for relative bias in measurement.

Equivalence testing and measurement

The use of latent variables in psychology, sociology or education has developed considerably in the last half a century in an aim to control for the inevitable fallibility of observed items and in order to get closer to substantial concepts used in theory. This development has been based on the Classical Test Theory (Lord and Novick, 1968) and has been extended with the use of latent variables in Structural Equation Modeling, Latent Class and Item Response Theory. These approaches assume that there is an underlying, unobserved, concept of interest that is measured with error by observed variables.

One such general model is the Confirmatory Factor Analysis (Bollen, 1989). Here we assume that a vector p of observed items, y, are explained by an m set of underlying continuous latent variables, ξ:

$$y^{(g)} = \upsilon^{(g)} + \Lambda^{(g)}\xi^{(g)} + \epsilon^{(g)} \tag{4}$$

where Λ is a $p * m$ matrix of factor loadings, υ is a vector of intercepts or thresholds and ϵ is a p vector of residuals (variances) independent of ξ and with a mean of zero (Bollen, 1989). The superscript g indicates that the coefficients may vary across g groups. Let μ_ξ and ϕ_ξ be the mean and the variance of ξ.

In this framework the loadings, Λ, and residuals, ϵ, can be considered to estimate the reliability of the items (Bollen, 1989). The intercept, or the threshold when the observed variables are categorical, is linked to the systematic part of the model. Variations of these quantities are also know in the Item Response Theory as discrimination and difficulty.

This measurement approach has been further extended to estimate relative bias by comparing these models across groups (Meredith, 1993; Millsap, 2012; Steenkamp and Baumgartner, 1998). Because researchers are usually interested in ξ it is essential that this is measured similarly (i.e., be equivalent or invariant) in each group of

interest. If this is not the case, then any use of the latent variable may confound differences in measurement with substantive differences.

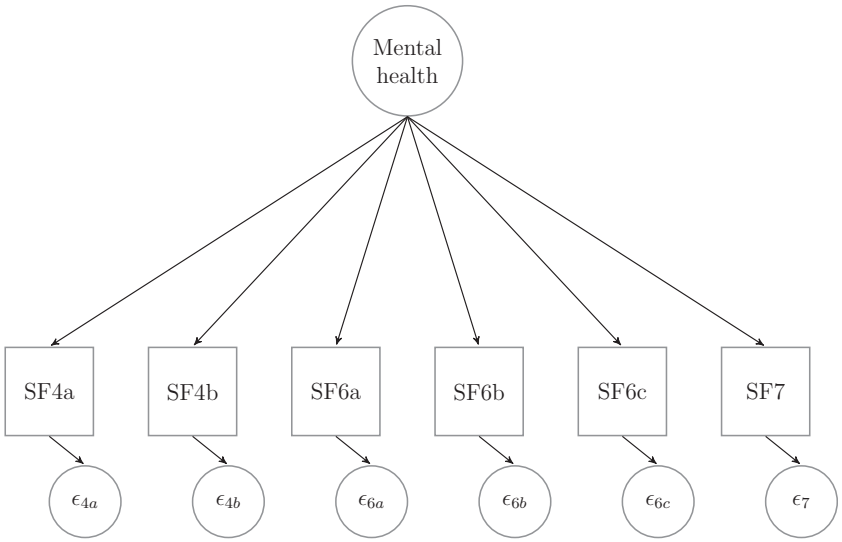
In order to evaluate whether the measurement model is equivalent across groups, and relative measurement error is the same, a series of nested models are tested. In each group different levels of equality restrictions are added across groups. Usually, the procedure starts with a general model, called the *configural model*, which assumes that the same structure is found across groups, but no equality constraints are imposed on the coefficients. If this model is found to fit the data, then a set of restrictions can be imposed on the Λ coefficients. If this model also holds (i.e., if it's not significantly worse than the configural model) the model is considered *metric equivalent* across groups (Steenkamp and Baumgartner, 1998). Next, a new set of restrictions can be added on the intercepts/thresholds, υ. If this model is accepted (i.e., fits the data well) then it is considered *scalar equivalent* (Steenkamp and Baumgartner, 1998), *strong factorial equivalent* (Meredith, 1993) or *first-order measurement invariant* (Millsap, 2012). The model can further be restricted to *strict factorial invariance* (Meredith, 1993) or *second-order invariance* (Millsap, 2012) by imposing equal random errors, ϵ. It should be noted that in order to compare means of the latent variable(s), μ_ξ, scalar equivalence needs to be found while in order to compare variances, ϕ_ξ, strict factorial invariance must be accepted.

The different levels of cross-group equality presented above are relatively strict and are hard to find in real-life data. As such, the concept of *partial equivalence* has been put forward (Byrne et al., 1989; Steenkamp and Baumgartner, 1998). This implies that even if not all the coefficients are equal across groups unbiased coefficients of ξ can be estimated if at least two items are equivalent and if the differences found on the other items are controlled for. This compromise has been found valuable as real world data has shown this to be quite common (e.g., Davidov, 2008).

While equivalence testing has become very popular due to the methodological and substantive insights it brings it nevertheless has a number of limitations. One of them refers to the fact that it can be implemented only when multiple items (preferably more than two for each ξ) of the same dimension are measured (Alwin, 2007). Secondly, the procedure estimates only relative bias. The measurement model may be the same across groups but may lack validity. Thirdly, the usual procedure for ascertaining the level of equivalence is exploratory and may capitalize on chance. Finally, the procedure cannot deal with certain types of systematic errors. For example, if primacy (i.e., tendency of selecting the first category irregardless of the question) is higher in all the items of one group then the difference will be included in the mean of ξ, thus confounding substantive and measurement differences across groups. This can be ameliorated by including the systematic errors in the model as has been done with acquiescence (Billiet and Davidov, 2008; Billiet and McClendon, 2000), method (Andrews, 1984; Campbell and Fiske, 1959; Saris and Gallhofer, 2007) or extreme response style (Kankaraš et al., 2011).

As mentioned at the beginning of the article, the equivalence testing procedure has been used a number of times in the mixed mode literature. The most typical use is to estimate measurement differences between modes after controlling for selection using a set of back-door variables, usually socio-demographics items (Gordoni et al., 2011; Hox et al., 2015; Klausch et al., 2013). Alternatively, it has been used to compare measurement differences of mode designs (Biemer, 2001) when these were randomly allocated (e.g., Cernat, 2014).

Figure 2: Measurement model to be tested for equivalence.



Previous research has also considered one of the limitations presented previously and have included other systematic errors in the model when comparing modes, such as acquiescence (Heerwegh and Loosveldt, 2011) or method (Révilla, 2013).

Equivalence testing as front-door approach

Given the the discussion so far, a natural question arises: is it possible to use equivalence testing, which was developed to estimate and control for differences in measurement, as a front-door to separate mode effects on selection and measurement? Because we do not expect mode to have a causal impact on the latent variable any differences found on this dimension can be due to selection, measurement or a combination of the two. Using equivalence testing we should be able to control for measurement differences, if these appear in the form of partial equivalence.

To see if this is the case and understand how results may be biased if assumptions don't hold a small simulation study will be presented below. Let's assume we want to measure mental health and we want to know whether people with different levels of health select into modes. One possible way to measure this is with items from the SF12 scale (Ware et al., 2007). SF12 is a scale developed to measure both physical and mental health. As such, we will choose only those items that measure the latter sub-dimension (Figure 2).

In order to have plausible values for the population model we will use results from the Understanding Society Innovation Panel wave 5 (Cernat, 2014; McFall et al., 2013). Applying the model in Figure 2 to these data we retrieve the following values that will be used as the true/population scores in the simulation study for the first group, m_1 (we will call these *Coef. 1*):

$$y = \upsilon + \Lambda\xi + \epsilon \tag{5}$$

$$\begin{bmatrix} SF_{4a} \\ SF_{4b} \\ SF_{6a} \\ SF_{6b} \\ SF_{6c} \\ SF_{7} \end{bmatrix} = \begin{bmatrix} 4.4 \\ 4.6 \\ 2.5 \\ 2.8 \\ 4 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 0.9 \\ -0.45 \\ 0.5 \\ 0.7 \\ 0.8 \end{bmatrix} \begin{bmatrix} \xi \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \\ 0.6 \\ 0.7 \\ 0.6 \\ 0.6 \end{bmatrix}$$

Let us further assume that the mean and variance for the mode of interest are $\mu_{\xi}^{m1} = 0$ and $\phi_{\xi}^{m1} = 1$. Furthermore, selection effects on the latent variable for the second mode will be added: $\mu_{\xi}^{m2} = 1.5$ and $\phi_{\xi}^{m2} = 1.5$. We know from the literature on equivalence testing that estimating a Multi-Group Confirmatory Factor Analysis assuming strict factorial invariance when only selection on the latent variable is present will lead to unbiased estimates (Hox et al., 2015; Meredith, 1964). Now lets assume that the second mode also has a measurement effect. This can be included in the model by imposing different intercepts, loadings and random errors in m_2 (which we will call *Coef. 2*):

$$y = \upsilon + \Lambda\xi + \epsilon \tag{6}$$

$$\begin{bmatrix} SF_{4a} \\ SF_{4b} \\ SF_{6a} \\ SF_{6b} \\ SF_{6c} \\ SF_{7} \end{bmatrix} = \begin{bmatrix} 5 \\ 4.6 \\ 2.5 \\ 2.8 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 0.5 \\ -0.85 \\ 0.5 \\ 0.7 \\ 0.8 \end{bmatrix} \begin{bmatrix} \xi \end{bmatrix} + \begin{bmatrix} 0.4 \\ 0.2 \\ 0.6 \\ 0.95 \\ 0.6 \\ 0.6 \end{bmatrix}$$

We expect that ignoring the confounding of selection and measurement in the two modes would lead to the biased estimation of the former. This can be clearly seen in the case 1 of Table 1 [1]. The mean and the variance of the selection in the second group is biased when we ignore measurement differences: a bias of 31 for the mean and 20 for the variance for the selection on the latent variable in the second mode. From the previous section we expect that if we are able to find partial equivalence between the modes then we can control for differences in measurement and estimate unbiased mode selection effects on the latent variable. By calculating the same model but freeing the coefficients that are different in the two groups we estimate the correct values for the mode selection effects (case 2 in Table 1). This exemplifies how partial equivalence testing can be used as a front door method for estimating selection on a latent variable of interest.

While this is very encouraging we also know that this approach has two important assumptions. The first one, exhaustiveness, implies that the partial equivalence captures all the measurement differences between the two modes. If this is not true then the selection effect will be biased. To test this let us imagine that in addition to the selection and measurement differences already included in the model, there is also a type of systematic error

[1] The simulations have been run in Mplus 7.2. Sizes of 2000 respondents were assumed for each group. A 1000 repetitions were used. Please contact author for the syntax used in the simulation.

in the second mode. This can take different forms such as acquiescence, social desirability, extreme response styles or recency/primacy. Here we will assume that acquiescence or primacy increases the chances of choosing the first category in the second mode. This is implemented in the model by adding a latent variable in the second group. This has loading of 1 on all the observed variables and a mean and variance of 1 (Billiet and Davidov, 2008; Billiet and McClendon, 2000). As expected, if this type of mode difference in measurement is ignored, then the estimate of selection will be biased (case 3 in Table 1): Means Square Error for $\mu_\xi^{m_2}$ and for $\phi_\xi^{m_2}$ are approximately 1. If appropriate measures are in the data, for example if balanced items are used for the items, then the response style can be modeled. When this is included (case 4 in Table 1) selection effects will not be biased. This highlights both a limitation of the model but also it's flexibility in including multiple types of systematic errors.

A second assumption of the front-door method is isolation. This implies that there are no other unobserved variables that have an impact both on measurement and selection. We can think of multiple theoretical situations when this may not be plausible. For example, people with lower working memory may have more measurement error in an auditory mode than a visual one and may also auto-select in one of them. To model such a situation let us imagine we have four groups: the reference mode with high working memory (m_{1a}) and with low working memory (m_{1b}), and the second mode with high working memory (m_{2a}) and with low working memory (m_{2b}). If isolation is not true in the population then working memory will have a differential effect on measurement and selection in the two modes. We can model this by imposing *Coef. 1* in the first three groups: m_{1a}, m_{1b} and m_{2a}. To estimate measurement differences for the fourth group we will impose *Coef.2* on m_{2b}. To simulate different selection we will impose the same mean and variance for the first mode $\mu_\xi^{m_{1a}} = \mu_\xi^{m_{1b}} = 0$ and $\phi_\xi^{m_{1a}} = \phi_\xi^{m_{1b}} = 1$ and differential selection within mode 2: $\mu_\xi^{m_{2a}} = 1$, $\phi_\xi^{m_{2a}} = 1$, $\mu_\xi^{m_{2b}} = 2$ and $\phi_\xi^{m_{2b}} = 2$.

In the real data, if we do not measure working memory then we assume that everything within each mode is equal (i.e., coefficients of $m_{1a} = m_{1b}$ and $m_{2a} = m_{2b}$). The theoretical expectation is that this indeed will bias the estimate of selection in the latent variable. This is obvious in case 5 of Table 1, where the coefficients for selection in the two subgroups of the second mode are equal but both coefficients have systematic error with bias ranging from 7 for $\phi_\xi^{m_{2b}}$ to 114 for $\phi_\xi^{m_{2a}}$. The last case of the simulation study shows once again that this assumption can be freed if we measure working memory in the data and if we include it in our model. The estimation of selection on the latent variable is unbiased and the model controls for differential measurement and selection.

Conclusions and discussion

This paper has shown how it is possible to conceptualize equivalence testing as a front-door method to estimate selection on a latent variable. While this technique has been used multiple times in the field of mixed modes it has yet to be considered on its own terms as a method to deal with the confounding of selection and measurement. The simulation study has shown that the method will work and give unbiased estimates.

That being said, the model does make two important assumption: isolation and exhaus-

Table 1: Simulation results

Nr.	Model: Population	Model: Estimation	Coefficient	Population	Model	Bias*	M.S.E**
1	Selection + partial equivalence	Selection	$\mu_\xi^{m_2}$	1.5	1.97	31.33	0.22
			$\phi_\xi^{m_2}$	1.5	1.8	20.00	0.09
2	Selection + partial equivalence	Selection + partial equivalence	$\mu_\xi^{m_2}$	1.5	1.5	0.00	0.00
			$\phi_\xi^{m_2}$	1.5	1.5	0.00	0.00
3	Selection + partial equivalence + response style	Selection + partial equivalence	$\mu_\xi^{m_2}$	1.5	2.48	65.33	0.96
			$\phi_\xi^{m_2}$	1.5	2.56	70.67	1.14
4	Selection + partial equivalence + response style	Selection + partial equivalence + response style	$\mu_\xi^{m_2}$	1.5	1.5	0.00	0.00
			$\phi_\xi^{m_2}$	1.5	1.5	0.00	0.00
5	Selection + partial equivalence + non-isolation	Selection + partial equivalence	$\mu_\xi^{m_{2a}}$	1	1.75	75.00	0.07
			$\phi_\xi^{m_{2a}}$	1	2.14	114.00	0.02
			$\mu_\xi^{m_{2b}}$	2	1.75	-12.50	0.07
			$\phi_\xi^{m_{2b}}$	2	2.14	7.00	0.02
6	Selection + partial equivalence + non-isolation	Selection + partial equivalence + non-isolation	$\mu_\xi^{m_{2a}}$	1	1	0.00	0.00
			$\phi_\xi^{m_{2a}}$	1	1	0.00	0.00
			$\mu_\xi^{m_{2b}}$	2	2	0.00	0.00
			$\phi_\xi^{m_{2b}}$	2	2	0.00	0.00

** Bias = 100*(Population - Sample)/Population; ** Mean Square Error = variance of sample estimation + Bias^2.*

tiveness. The simulation has shown that indeed when these do not hold in the population the sample estimates of selection will be biased. Nevertheless, the method is flexible enough to give users the opportunity to include any potential biasing factors as covariates. This makes for a very versatile method for disentangling selection and measurement.

Equivalence testing has it's own limitations as a statistical method, such as the need for multiple items or capitalization on chance. This may lead to other types of biases when the method is applied to the real world data. The paper has not tackled this issue directly but there is considerable ongoing research that should reduce these issues in the future (e.g., Asparouhov and Muthén, 2014).

The paper has only highlighted the utility of the approach and possible limitations. In order to make it more attractive for real world applications further research is needed. For example, a thorough study that simulates multiple types of models with varying degrees of miss-specification (e.g., multiple types of errors, multiple types of unobserved covariates) may indicate to users the degree of bias they can expect when applying this method. Similarly, developing methods to utilize the information estimated using this approach for other purposes, such as creating weights or correcting substantive models, should be pursued.

References

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

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

Asparouhov, T. and Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4):495–508.

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.

Billiet, J. and Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research*, 36(4):542–562.

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

Bollen, K. (1989). *Structural equations with latent variables*. Wiley-Interscience Publication, New York.

Byrne, B., Shavelson, R., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3):456.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105.

Cernat, A. (2014). Impact of mixed modes on measurement errors and estimates of change in panel data. *Understanding Society Working Paper Series*, (05):1–21.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the european social survey. *Survey Research Methods*, 2(1):33–46.

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, 21(5):233–255.

Gordoni, G., Schmidt, P., and Gordoni, Y. (2011). Measurement invariance across Face-to-Face and telephone modes: The case of Minority-Status collectivistic-oriented groups*. *International Journal of Public Opinion Research*.

Heerwegh, D. and Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, 27(1):49–63.

Hox, J. J., De Leeuw, E. D., and Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, 6.

Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1):3–20.

Kankaraš, M., Vermunt, J., and Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods and Research*, (20):1–31.

Klausch, T., Hox, J. J., and Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, 42(3):227–263.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company, Inc.

Lugtig, P. J., Lensvelt-Mulders, G. J., Frerichs, R., and Greven, F. (2011). Estimating nonresponse bias and mode effects in a mixed mode survey. *International Journal of Market Research*, 53(5):669–686.

McFall, S., Burton, J., Jäckle, A., Lynn, P., and Uhrig, N. (2013). Understanding society – the UK household longitudinal study, innovation panel, waves 1-5, user manual. *University of Essex, Colchester*, pages 1–66.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, 29(2):177–185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543.

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. Routledge Academic, 1 edition edition.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, 1 edition edition.

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

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2Rev e. edition edition.

Révilla, M. A. (2013). Measurement invariance and quality of composite scores in a face-to-face and a web survey. *Survey Research Methods*, 7(1):17–28.

Saris, W. and Gallhofer, I. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Wiley-Interscience, 1 edition.

Steenkamp, J. E. M. and Baumgartner, H. (1998). Assessing measurement invariance in Cross-National consumer research. *Journal of Consumer Research*, 25(1):78–107.

Vannieuwenhuyze, J. T., Loosveldt, G., and Molenberghs, G. (2014). Evaluating mode effects in Mixed-Mode survey data using covariate adjustment models. *Journal of Official Statistics*, 30(1):1–21.

Vannieuwenhuyze, J. T. A. and Révilla, M. (2013). Relative mode effects on data quality in Mixed-Mode surveys by an instrumental variable. *Survey Research Methods*, 7(3):157–168.

Ware, J., Kosinski, M., Turner-Bowker, D. M., and Gandek, B. (2007). *User's Manual for the SF-12v2 Health Survey*. QualityMetric, Incorporated.