Survey Research Methods (2021)
Vol. 16, No. 1, pp. 29-43
doi:10.18148/srm/2021.v16i1.7811
European Survey Research Association

Measurement Equivalence in Sequential Mixed-Mode Surveys

Joseph W. Sakshaug
Ludwig Maximilian University of Munich, Germany, and
Institute for Employment Research,
Nuremburg, Germany

Alexandru Cernat
School of Social Sciences
University of Manchester, United Kingdom

Richard J. Silverwood
Centre for Longitudinal Studies
UCL Social Research Institute
London, United Kingdom

Lisa Calderwood
Centre for Longitudinal Studies
UCL Social Research Institute
London, United Kingdom

George B. Ploubidis
Centre for Longitudinal Studies
UCL Social Research Institute
London, United Kingdom

Many surveys collect data using a mixture of modes administered in sequential order. Although the impacts of mixed-mode designs on measurement quality have been extensively studied, their impacts on the measurement quality of unobservable (or latent) constructs is still an understudied area of research. In particular, it is unclear whether latent constructs derived from multi-item scales are measured equivalently across different sequentially-administered modes—an assumption that is often made by analysts, but rarely tested in practice. In this study, we assess the measurement equivalence of several commonly-used multi-item scales collected in a sequential mixed-mode (Web-telephone-face-to-face) survey: the Age 25 wave of the Next Steps cohort study. After controlling for selection via an extensive data-driven weighting procedure, a multi-group confirmatory factor analysis was performed to assess measurement equivalence across the three modes. We show that cross-mode measurement equivalence is achieved for nearly all scales, with partial equivalence established for the remaining. Although measurement equivalence was achieved, some differences in the latent means were observed between the modes, particularly between the interviewer-administered and self-administered modes. We conclude with a discussion of these findings, their potential causes, and implications for survey practice.

Keywords: confirmatory factor analysis; interviewer-administered survey; measurement invariance; mode effects; scalar equivalence; Web survey

1 Introduction

Rating (or Likert-type) scales are commonly used in survey research to measure latent constructs (or factors) that are not directly observable. Scales consisting of multiple items are typically used by researchers to form an index that relates to, and can be inferred to, the true score of the latent construct. For example, a commonly-used scale in the screening of psychiatric disorders is the General Health Questionnaire (GHQ-12). The GHQ-12 consists of 12 items each answered using a Likert-type response scale. The responses to these items are summed to produce an overall score that corresponds to a respondent's true score on the latent variable of psychological distress (Hamer, Chida, & Molloy, 2009; Jackson, 2007). Other oft-used scales include the Locus of Control, which assesses the extent to which people believe they have control over certain outcomes in their lives (Ashby, Kottman, & Draper, 2002; Shepherd, Owen, Fitch, & Marshall, 2006), and the Alcohol Use Disorders Identification Test-Consumption (AUDIT-C), which is used to identify hazardous alcohol consumption (King et al., 2012; Reinert & Allen, 2007). Given the considerable costs and resources spent on developing, testing, and implementing multi-item scales in surveys, it is important that the correlational relationship (or factor structure) between the observed items re-

Contact information: Joseph W. Sakshaug, Institute for Employment Research, Regensburger Strasse 104, 90478 Nuremberg (Germany) (E-Mail: joe.sakshaug@iab.de)

flects the latent construct of interest across different survey conditions.

A key assumption underlying measurement scales is that individuals (or groups) who possess the same value on the latent variable provide the same answers to the same scale items regardless of the conditions under which the measurements are collected. This assumption is known as measurement equivalence or measurement invariance (Jöreskog, 1971; Meredith, 1993; Vandenberg & Lance, 2000). Establishing measurement equivalence is necessary to form a baseline against which means and relationships of latent variables can be meaningfully compared between groups of individuals on the same measurement scale. The ability to perform specific types of analyses and comparisons on the latent variables depends on the validity of this assumption. However, achieving full measurement equivalence can be problematic in surveys when different groups of respondents, who possess the same underlying latent score, interpret the items differently or answer them in ways that produce different responses to the same items. This is a common concern in cross-national surveys where translation, cultural differences in question meaning, and response behavior can distort measurements and impede valid comparisons between countries (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Davidov, Schmidt, Billiet, & Meuleman, 2018; Rutkowski & Svetina, 2014), but also in mixed-mode surveys where the same items administered in different modes may be interpreted and answered in different ways (Klausch, Hox, & Schouten, 2013).

Mixed-mode designs are common in survey research. In particular, sequential mixed-mode designs, where multiple modes are administered in sequence, are becoming more common as they've been shown to reduce costs and improve the representativeness of the respondent pool (Bianchi, Biffignandi, & Lynn, 2017; De Leeuw, 2005; Roberts, Joye, & Ernst Stähli, 2016; Wagner, Arrieta, Guyer, & Ofstedal, 2014). However, such designs can have adverse effects on measurement quality. For instance, it is well-known that respondents tend to give different answers to the same items when interviewed in different modes, especially in interviewer-administered (e.g. face-to-face) and self-administered (e.g. Web) modes. Various types of measurement mode effects have been cited in the literature, including social desirability bias, acquiescence bias, and primacy/recency effects (Jäckle, Roberts, & Lynn, 2010). Mode effects have the potential to distort answers and introduce systematic bias in multi-item scales, but whether such effects are severe enough to render latent variable measurements nonequivalent or incomparable between sequentially-administered modes is a relatively understudied issue.

In the present study, we address this issue by performing equivalence testing on several commonly-used, multi-item scales administered in a well-known sequential mixed-mode (Web, telephone, face-to-face) survey in the UK: the Next Steps cohort study. Using multi-group confirmatory factor analysis, we investigate whether cross-mode measurement equivalence, or at least partial equivalence, can be established for each scale.

2 Background

2.1 Measurement Mode Effects

It is well documented that mixing data collection modes can influence the quality of survey measurements (Ansolabehere & Schaffner, 2014; DeMaio, 1984; Dillman et al., 2009; Heerwegh & Loosveldt, 2011; Holbrook, Krosnick, Moore, & Tourangeau, 2007; Hope, Campanelli, Nicolaas, Lynn, & Jäckle, 2014; McClendon, 1991; Nicolaas, Campanelli, Hope, Jäckle, & Lynn, 2015; Revilla, 2015; Smyth, Olson, & Kasabian, 2014; Tourangeau, Rips, & Rasinski, 2000; Ye, Fulton, & Tourangeau, 2011). Specifically, there is a tendency for the same respondents to give different answers to the same questions posed in different modes. Given that specific types of measurement effects are more prominent in certain modes, mixing modes with different measurement properties can give rise to differential measurement effects, often referred to as measurement mode effects (De Leeuw, 2005).

There are at least two mode-specific features that largely explain the manifestation of specific types of measurement effects in different modes (De Leeuw, 2005). The first feature is whether the questions are communicated visually or orally. Both communication channels seem to affect the cognitive processes and memory capacity of respondents (Krosnick & Alwin, 1987; Schwarz, Strack, Hippler, & Bishop, 1991). Consequently, respondents may be more likely to select the first options presented in a visual list of possible response options without carefully considering the other options in order to reduce their cognitive load, resulting in a primacy effect. In aural modes, the reverse behaviour is more likely to occur—with respondents having a higher propensity to select the last options they hear in a spoken list due to constraints on working memory capacity, resulting in a recency effect (Schwarz et al., 1991). Reducing cognitive load has also been cited as a potential reason why aural modes tend to elicit more extreme responses to items than in visual modes (Christian, Dillman, & Smyth, 2008; De Leeuw, 1992; Dillman et al., 2009). Differences in the frequency of similar (or nondifferentiated) answers given to attitudinal item batteries also tend to be larger between aural and visual modes than within them (Chang & Krosnick, 2009; Fricker, Galesic, Tourangeau, & Yan, 2005; Greene, Speizer, & Wiitala, 2008; Heerwegh & Loosveldt, 2008; Holbrook, Green, & Krosnick, 2003; Kim, Dykema, Stevenson, Black, & Moberg, 2019).

The second mode feature is interviewer presence. The presence of an interviewer is known to affect socially desirable responding. A common research finding is that respondents are more likely to give answers to sensitive items that portray themselves more favourably, in line with social and societal norms, in interviewer-administered (as opposed to self-administered) modes (De Leeuw, 1992, 2005; Tourangeau, Conrad, & Couper, 2013; Tourangeau & Smith, 1996). In general, social desirability bias tends to be greater in telephone interviews, followed by face-to-face, and self-administered (e.g. Web, mail) modes (Bowling, 2005; Cernat, Couper, & Ofstedal, 2016; Heerwegh, 2009; Kreuter, Presser, & Tourangeau, 2008; Tourangeau & Yan, 2007). Measurement mode effects for sensitive items have also been reported in experimental mixed-mode designs involving sequentially-administered self- and interviewer-administered modes (Kappelhof & De Leeuw, 2019; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2012).

Another type of measurement error that has been shown to vary between self- and interviewer-administered modes is acquiescence bias. Acquiescence occurs when respondents haphazardly agree to statements or answer "yes" to questions regardless of their content, which might be done to minimize cognitive burden (Knowles & Condon, 1999; Krosnick, 1991). Although some studies have found that acquiescence is lessened in self-administered modes compared to interviewer-administered modes, the results tend to be modest and not always statistically significant (De Leeuw, 1992; Dillman & Mason, 1984; Fricker et al., 2005; Greene et al., 2008; Heerwegh, 2009; Heerwegh & Loosveldt, 2011; Holbrook et al., 2003; Tarnai & Dillman, 1992).

Measurement mode effects can potentially introduce systematic biases in rating scale items in mixed-mode surveys. To minimize the risk of measurement mode effects, the general recommendation is not to mix modes that differ with respect to interviewer presence or communication channel (aural/visual), or if this is unavoidable, to do so with great care. For example, Klausch et al. (2013) state that "caution is required when combining data from interviewer- and self-administered modes, especially if considerable amounts of attitudinal rating scale questions are to be included." However, the decision to mix self- and interviewer-administered modes in sequential mixed-mode surveys is often done deliberately in order to reduce costs while maximizing response rates and minimizing the risk of noncoverage and nonresponse error (Bianchi et al., 2017; De Leeuw, 2005; Roberts et al., 2016; Wagner et al., 2014), and these considerations may outweigh the risks of potential measurement effects. In this situation, efforts might be taken to mitigate these risks by administering items that are most susceptible to measurement effects (e.g. sensitive items) via computer-assisted self-completion (CASI) or one of its variants (e.g. audio-CASI) in the interviewer-administered mode, in order to minimize the effects of interviewer presence and communication channel across the different mode types.

2.2 Measurement Equivalence in Mixed-Mode Surveys

Given that sequential mixed-mode designs are commonly used in practice but are also susceptible to measurement mode effects, it is important to assess whether responses to multi-item scales used to measure underlying (latent) attributes are comparable across modes and meet the critical assumption of measurement equivalence assumed by researchers. Without testing this assumption, researchers may incorrectly conclude that respondents who answer identically to the multi-item scales (or possess the same composite score) in different modes have the same value on the latent variable of interest, when in fact they do not. That is, meaningful comparisons of factors between individuals interviewed in different modes may be distorted by mode effects.

Measurement equivalence is generally investigated in the framework of multi-group confirmatory factor analysis using multi-group structural equation modelling (SEM) (Jöreskog, 1971; Meredith, 1993; Vandenberg & Lance, 2000). There are at least three levels of measurement equivalence considered in this framework with each level permitting specific types of analyses to be performed on the latent variable. The levels typically follow a hierarchical structure with additional restrictions cumulatively imposed on parameters of the measurement model for the higher levels (Meredith, 1993; Millsap, 2012; Steenkamp & Baumgartner, 1998).

The usual procedure in testing for measurement equivalence across groups (e.g. modes, cultures, etc.) starts with the configural model. Configural equivalence is the least stringent level of measurement equivalence as it does not impose any restrictions on the measurement model between groups, other than the factorial structure is identical between the groups of interest. It implies that the observed items are related to the same latent factor in each group (i.e. the factor structure is restricted to be the same), but the nature of the relationship does not have to be equivalent in each group. When configural equivalence holds, then construct validity is achieved and further levels of measurement equivalence can be tested. The second level is metric equivalence (or weak factorial or loading equivalence). This implies that the factor loadings, representing the strength of the linear relationship between the observed items and the latent factor, are the same across groups (Bollen, 1989). If metric equivalence holds, then the meaning of the latent factor is the same in all groups and correlations or relationships between the latent factor and external variables can be compared across groups.

The third level is scalar equivalence (or strong factorial or intercept equivalence), which implies that the intercepts of the measured variables (or thresholds, in the case of categorical variables), in addition to the factor loadings, are equal across groups. This level of equivalence permits the comparison of latent mean differences across groups and indicates that respondents use the scale in the same way (Ploubidis, McElroy, & Moreira, 2019; Vandenberg & Lance, 2000; Widaman & Reise, 1997). Additional levels of measurement equivalence (e.g. strict factorial equivalence), which impose more restrictive constraints on the model parameters, can be tested. However, these levels of equivalence are highly constrained and rarely necessary in practical applications; thus, we do not consider them further and instead focus only on the three most common forms of measurement equivalence: configural, metric, and scalar.

In cases where full scalar equivalence cannot be achieved, partial equivalence can be established by identifying the specific item parameters causing the nonequivalence and relaxing their constraints (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998). For example, the restriction of equal intercepts across groups might be relaxed for some items and the parameters freely estimated in order to establish partial equivalence and allow for substantive comparisons (e.g. the latent means). Although partial equivalence allows valid conclusions to be drawn about the latent variable, it requires that at least two observed items of the latent factor achieve full scalar equivalence.

Several studies have assessed measurement equivalence across survey modes (Cernat et al., 2016; Cernat & Revilla, 2021; De Leeuw, 1992; De Leeuw, Mellenbergh, & Hox, 1996; Gordoni, Schmidt, & Gordoni, 2012; Heerwegh & Loosveldt, 2011; Hox, De Leeuw, & Zijlmans, 2015; Klausch et al., 2013; Revilla, 2013; Revilla & Saris, 2013; Tomé, 2018). They generally find that full (or partial) scalar equivalence is more common between self-administered modes (namely, mail and Web) and between interviewer-administered modes (e.g. telephone, face-to-face), and less common between self- and interviewer-administered modes. Some studies have found systematic measurement bias for sensitive items. For example, Klausch et al. (2013) find that interviewer-administered modes (i.e. face-to-face, telephone) have higher category thresholds for attitudinal items about police and traffic than the self-administered modes (i.e. mail, Web), which the authors suggest might be due to a stronger tendency for socially desirable responding in the interviewer modes. Heerwegh and Loosveldt (2011) find higher intercepts and larger latent variable means in the telephone mode of a crime victimization survey than in the mail mode, indicating stronger socially desirable responding in the telephone mode. Similar findings are reported by Cernat et al. (2016) with respondents reporting higher levels of depression in the Web mode than in telephone and face-to-face modes.

2.3 Research Gaps and Study Questions

The above assessments primarily come from designed mode experiments or comparisons of different modes administered in parallel surveys, which are rarely implemented in practice or used by substantive researchers. Multi-item scales are commonly administered in non-experimental sequential mixed-mode designs, but assessments of their cross-mode measurement equivalence are currently lacking in the literature. One exception is Hox et al. (2015), who examine measurement equivalence in a (non-experimental) sequential mixed-mode (Web, telephone, face-to-face) survey. Across 14 multi-item scales about family life and health in the third wave of the Netherlands Kinship Panel Study (NKPS), they find that measurement equivalence is achieved for most scales across the three modes, but in most cases only partial equivalence is reached.

In addition to measurement mode effects, another reason why measurement equivalence may not be achieved across modes is due to differential nonresponse (or selection). Mixed-mode surveys are susceptible to differential nonresponse because different types of respondents tend to have different propensities for being interviewed in a given mode (De Leeuw, 2005). Usually selection effects and measurement effects are completely confounded, even in experimental mode comparisons, but especially in sequential mixed-mode designs where the probability of responding in a given follow-up mode is conditional on whether a response was obtained in the previously-offered mode(s). As mixed-mode surveys are already susceptible to measurement mode effects, selection mode effects can further distort the equivalence of multi-item scales (Cernat et al., 2016). Most measurement equivalence studies attempt to control for selection effects by utilizing auxiliary data that are assumed to be mode insensitive, such as demographic variables or register data. However, in most practical applications, the number of control variables is limited to explain the selection mechanism and satisfy the Missing at Random assumption (Little & Rubin, 1989). Hox et al. (2015) use available data from two previous waves of the NKPS to create a propensity score adjustment, which they apply to their measurement equivalence analysis. The authors show that controlling for selection improves measurement equivalence in most cases, but not all.

We contribute to this literature by investigating the validity of the cross-mode measurement equivalence assumption in a sequential mixed-mode (Web, telephone, and face-to-face) survey in the UK: the Next Steps cohort study. Using multi-group confirmatory factor analysis, we examine several sensitive and non-sensitive multi-item scales, including the GHQ-12, Locus of Control, and AUDIT-C scales which, to our knowledge, have not been tested for cross-mode equivalence in a sequential mixed-mode survey. In addition, we apply a novel data-driven nonresponse adjustment procedure utilizing data collected from seven previous waves of Next Steps to control for selection.

The following research questions are addressed:

1. Do the multi-item scales achieve measurement equiva-

lence across sequentially-administered Web, telephone, and face-to-face modes after adjusting for selection?

2. For the scales that do not achieve full measurement equivalence, can partial equivalence be established?

3. To what extent (if any) do the latent variable means vary across the three modes?

3 Data and Methods

3.1 The Next Steps Cohort Study

Next Steps is a national cohort study in the UK which follows a representative sample of people born between 1st September 1989 and 31st August 1990. Cohort members were initially recruited in schools during their adolescence at age 13/14. The target population consists of young people who were enrolled in Year 9 in English state and independent schools and pupil referral units in February 2004. The sample design considered schools as primary sampling units and included an oversampling of deprived schools and minority ethnic groups within schools. The fielded sample at baseline comprised approximately 21,000 young people with a total of 15,770 persons interviewed at baseline. An additional minority supplement was added at the Age 17 wave. From ages 15–20, the fielded sample consisted of cohort members who had participated at the previous wave. For the Age 25 data collection (wave 8), which is the focus of our study, the fielded sample included all cohort members who had ever participated in the study.[1]

In the Age 25 wave, a sequential mixed-mode design was implemented, starting with a request to complete the survey online, followed by telephone, and then face-to-face for the remaining non-respondents. A total of 7,707 (out of 15,108) eligible cohort members participated in the survey (Web: 4,797; telephone: 690; face-to-face: 2,220), for an overall response rate of 51 percent (Response Rate 1; American Association for Public Opinion Research, 2016; Bailey, Breeden, Jessop, and Wood, 2017).

3.2 Multi-Item Scales

We perform measurement equivalence testing on seven multi-item scales which are each assumed to measure a single latent construct. Some scales have been validated to reflect an underlying latent construct while others are basic inventories of formative items that relate to a specific topic but were not specifically designed to reflect a single latent construct (Bollen & Lennox, 1991). Following other literature (e.g. Hox et al., 2015), we treat all formative scales as reflective ones in the forthcoming analyses. A brief description of each scale is presented below. All scale items are presented in the online appendix (Tables A1–A7).

Adult Identity Resolution Scale (Adult). The Adult scale is a 3-item (4-point ordinal) scale. The scale was designed to study identity development from adolescence through early adulthood (Côté, 1996; Côté & Levine, 2002; Côté, Mizokami, Roberts, & Nakama, 2016). The scale is commonly used to assess whether people consider themselves to be a fully matured adult and feel respected as an adult by others.

Alcohol Use Disorders Identification Test-Consumption (AUDIT-C). The AUDIT-C is a 3-item (5-point ordinal) short scale adapted from the longer 10-item AUDIT instrument (Bradley et al., 2003; Bush, Kivlahan, McDonell, Fihn, & Bradley, 1998). It is an internationally recognized and widely used tool to assess how much and how often men and women consume alcohol. It is commonly used to identify problem drinkers or those who have alcohol use disorders, such as alcohol abuse or dependence. The scale was administered via computer-assisted self-interviewing (CASI) in the face-to-face interviews.

Bullying. Bullying is a 7-item (2-point; yes/no) formative scale. It focuses on victimization rather than perpetuation and asks whether respondents have experienced different forms of bullying (e.g. physical or emotional abuse) in the past 12 months. The scale was administered via CASI in the face-to-face interviews.

General Health Questionnaire (GHQ-12). The GHQ-12 is a 12-item (4-point ordinal) scale adapted from the original 60-item instrument (Goldberg & Williams, 1988). It is primarily used as a psychometric screening instrument to identify common mental disorders. Each item indicates the severity of a particular symptom of mental ill health. It is extensively used in cross-cultural settings for evaluating minor psychological disorders. The scale was administered via CASI in the face-to-face interviews.

Leisure A. Leisure A is a 6-item (4-point ordinal) formative scale. The items ask about the frequency of taking part in various types of recreational activities (e.g. sport/exercise, going to cinema).

Leisure B. Leisure B is a 4-item (4-point ordinal) formative scale. It measures the frequency of engaging in different types of volunteering activities (e.g. attending local group meetings, donating to charity).

Locus of Control (Locus). Locus is a 4-item (4-point ordinal) short scale adapted from the longer 13-item instrument developed by Rotter (1966). It measures how

[1] The Next Steps data (University College London, 2021) are available through the UK Data Service: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=5545. Replication code is available upon request.

strongly people believe that they themselves (as opposed to external factors) have control over the outcome of events that affect their lives. The scale is widely used internationally in different fields of research, including psychology, health, and marketing. The scale was administered via CASI in the face-to-face interviews.

3.3 Accounting for Selection

As previously noted, measurement effects can be confounded with selection effects when people have different propensities to respond in a given mode. This is especially the case in sequential mixed-mode surveys where the mode of interview completion is largely dictated by the behaviour of the sample member. In order to study measurement effects, it is therefore important to control for selection as much as possible to isolate the measurement mode effects and minimize confounding of selection. Researchers have proposed a variety of methods to control for selection effects in mixed-mode surveys, including regression (Jäckle et al., 2010), matching (Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011), weighting (Hox et al., 2015), or utilizing record data (Sakshaug, Yan, & Tourangeau, 2010; Voogt & Saris, 2005).

To account for selection into each mode of the sequential mixed-mode design used in Next Steps, we employed a data-driven mode-specific unit nonresponse weighting adjustment procedure that used all available variables (e.g. sociodemographic, behavioural, attitudinal) collected from all prior waves of Next Steps. The general procedure, first introduced by Mostafa et al. (2021) followed by Silverwood, Calderwood, Sakshaug, and Ploubidis (2020), consisted of a two-stage analytic approach that was applied separately to each of the three mode phases used in the sequential mixed-mode design.

In the first stage, a series of seven (one for each of the previous Next Steps waves) within-wave multivariable Poisson regressions were fitted with mode-specific response at Wave 8 as the outcome. For the Web mode, response was modelled as Web response(1) vs. non-Web response(0), for the telephone mode response was modelled as telephone response(1) vs. non-telephone response(0), and for the face-to-face mode response was modelled as face-to-face response(1) vs. non-face-to-face response(0). All variables whose association reached a statistical significance level of 5% were retained for the second stage.

In the second stage, each of the retained variables from the first stage was singly-imputed to produce a complete dataset of predictors. After imputation, the retained variables from each subsequent wave (starting with Wave 1) were cumulatively entered into a series of multivariable Poisson regressions predicting mode-specific response at Wave 8. Mode-specific response at Wave 8 was modelled as a function of predictors from a given wave adjusted for all predictors from the prior waves that were retained in the first stage. For example, when considering predictors of mode-specific response at Wave 8 observed at Wave 4, predictors from Waves 1–3 identified in the first stage were controlled for but predictors from Waves 5–7 identified in the first stage were not included in the model. The predictors which remained statistically significant at the 5% level after controlling for predictors from prior waves were then retained. This was done to preserve the temporal sequence of the longitudinal information available in Next Steps.

The retained second-stage predictors were then used to create propensity score adjustment weights to adjust for unit nonresponse in each mode. Five propensity score subgroups were generated for each mode-specific response outcome using quintiles of the estimated propensity scores. The final adjustment weight was then calculated as the inverse of the average propensity score identified in each subgroup multiplied by the Next Steps design weight. A total of three adjustment weights were created, one for each mode, with the weight corresponding to the mode in which a particular respondent participated. We apply these weights in all forthcoming analyses to control for the confounding effects of mode-specific selection and aid in isolating the studied measurement effects. A complete list of all predictors used in the final adjustment model for each mode, which included several sociodemographic (e.g. sex, ethnicity, education) and other background variables, is provided in the online appendix materials (Tables A8–A10).

3.4 Measurement Equivalence Testing Procedure

To test the measurement equivalence of the multi-item scales between the three modes we use multi-group Confirmatory Factor Analysis (MCFA) equivalence testing for complex survey samples. Mplus 8.3 software (Muthén & Muthén, 1998-2017) was used to test the three most common forms of measurement equivalence across the three modes (Meredith, 1993):

Configural equivalence, i.e. the factor structure is the same across the different modes;

Metric equivalence, i.e. configural equivalence holds and the factor loadings are the same across modes; and

Scalar equivalence, i.e. metric equivalence holds and the intercepts are the same in all modes.

If measurement equivalence holds across modes, then it is possible to compare unstandardized relationships (for metric equivalence) and/or latent means (for scalar equivalence) across the modes.

A simple MCFA model with one latent construct was used for each multi-item scale. A visual depiction of the measurement model is given in Figure 1. The latent variable is represented by a circle and the observed variables (items) are represented by squares. Each observed value, y_i, is explained by the true latent value, T, with a loading/slope parameter λ_i, an intercept/threshold (or conditional mean) τ_i, and a residual term ε_t. The first loading was set to 1 for identification purposes. For testing configural equivalence, the loadings and intercepts/thresholds were allowed to be estimated freely without restriction. For testing metric equivalence, the loadings were restricted to be equivalent across modes. For scalar equivalence, we additionally restricted the intercepts/thresholds to be equal across modes.

All scale items are treated as categorical (ordinal) with the exception of the AUDIT-C scale items, which are treated as continuous. The rationale is that the AUDIT-C scale is the only scale with at least five categories and is approximately normally distributed (Rhemtulla, Brosseau-Liard, & Savalei, 2012). We use THETA parametrization together with Weighted Least Squares Means and Variance (WLSMV) estimation for categorical variables and Maximum Likelihood Estimation with Robust Standard Errors (MLR) for the AUDIT-C scale. Full Information Maximum Likelihood (Enders, 2010) is used to handle item missing data in the AUDIT-C scale, which is assumed to be Missing at Random given the model of interest. For the categorical scales, pairwise present data are analyzed. All analyses account for the complex sample design (stratification, clustering, weighting) used in Next Steps.

To assess whether measurement equivalence holds, conventional goodness-of-fit criteria are applied to assess the fit of each measurement model. The fit criteria include the chi-square test statistic (lower is better), Comparative Fit Index (CFI; higher is better), Tucker-Lewis Index (TLI; higher is better), and the Root Mean Square Error of Approximation (RMSEA; lower is better) (see West, Taylor, and & Wu, 2012, for an overview of these fit criteria and their recommended cut-off values). In line with the current SEM literature, we present a mix of fit indices, but in order to make decisions about equivalence we focus on changes in the CFI when adding the constraints to the different modelling steps. A change (Δ) larger than 0.01 in the CFI indicates that the restrictions do not hold (Chen, 2007; Davidov et al., 2018). While we adopt this threshold, we note that other researchers have suggested slightly more liberal cut-off values (e.g. Rutkowski & Svetina, 2014).

We compare the factor loadings and item thresholds (or intercepts, for the AUDIT-C scale) between the three modes. For most scales (AUDIT-C, Bullying, GHQ-12, and Locus), the items were administered via self-completion in the face-to-face mode. Here, we expect to find smaller differences in the loadings and/or thresholds/intercepts between Web and face-to-face modes relative to comparisons involving the telephone mode, which did not use self-completion in any of the scales. In cases where we do not find full scalar equivalence between modes, we investigate the cause of this and try to find a good fitting model that exhibits partial equivalence (Byrne et al., 1989). Identifying the exact variables and coefficients that cause differences between the modes can help us better understand the potential mechanisms that lead to these differences.

4 Results

4.1 Descriptive Statistics and Evaluation of Weighting Procedure

Before proceeding with the measurement equivalence analysis, descriptive statistics for respondents answering in each mode in Wave 8 of Next Steps are shown in the online appendix materials (Figure A1). The statistics are shown before and after applying the weighting adjustment to account for selection into mode with reference to the Wave 1 (baseline) respondents. Before accounting for selection, there are 20 (out of 54) mode-specific estimates with 95% confidence intervals that do not overlap with those of the corresponding Wave 1 estimates. The majority of these differences are observed for Web (9 out of 18) and telephone (9 out of 18), indicating selection bias for these modes. However, after accounting for selection via weighting, the descriptive estimates in each mode group are much more in line with the Wave 1 reference estimates—only three estimates (all in telephone) do not overlap with the Wave 1 reference estimates. Furthermore, there are no strong differences between the mode groups in Wave 8 after applying the weighting adjustment, with the exception of household tenure (ownership) which is overrepresented in the telephone group. Taken together, these results are evidence that the aforementioned weighting procedure was effective at reducing selection bias in the mode groups, and this enables us to focus our attention on measurement effects in the equivalence analysis below.

4.2 Full Scalar Equivalence (RQ1)

We start the equivalence analysis by investigating the fit indices for all seven scales and all three types of measurement equivalence (Table 1). Overall, the models display good fit based on CFI, TLI, and RMSEA. If ΔCFI > 0.01 is used as the cut-off value for making decisions regarding the equivalence testing restrictions, then almost all scales satisfy the requirements of scalar equivalence with the lone exception being the Leisure A scale, which attains only metric equivalence. We note that the results are slightly sensitive to the use of selection weights. If selection were not accounted for (see online appendix, Table A11), then two additional scales (Leisure B and AUDIT-C) would not achieve scalar equivalence. Thus, the use of the weighting procedure improves the equivalence of these scales.

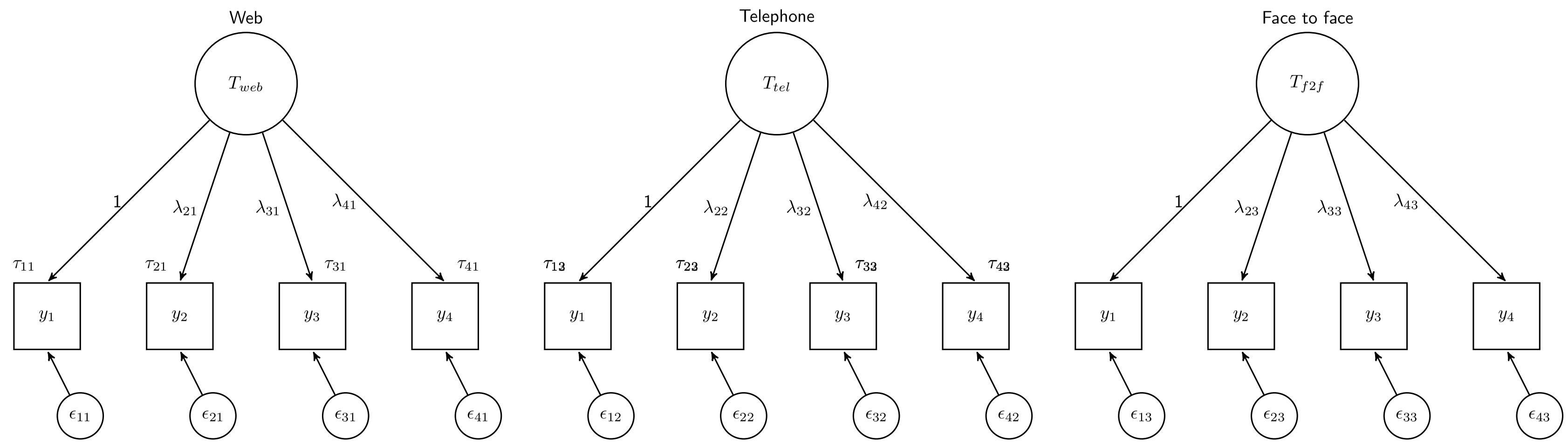

Figure 1. Visual representation of a measurement model tested for a 4-item scale across modes.

Table 1
Goodness of fit by scale and model. Equivalence testing is performed by adding cumulative restrictions.

Scale	Model	X^2	df	CFI	TLI	RMSEA
Adult	Configural	24.6	8	0.998	1.00	0.03
Adult	Metric	47.1	12	0.996	1.00	0.04
Adult	Scalar	90.4	20	0.993	1.00	0.04
GHQ	Configural	3189.7	188	0.950	0.95	0.08
GHQ	Metric	2180.4	210	0.967	0.97	0.06
GHQ	Scalar	2244.8	254	0.967	0.97	0.06
Leisure A	Configural	412.7	41	0.919	0.91	0.06
Leisure A	Metric	395.0	51	0.925	0.93	0.05
Leisure A	Scalar	611.3	71	0.882[*]	0.93	0.06
Leisure B	Configural	89.5	16	0.985	0.98	0.04
Leisure B	Metric	92.4	22	0.985	0.99	0.04
Leisure B	Scalar	149.7	34	0.976	0.99	0.04
Locus	Configural	109.7	16	0.945	0.94	0.05
Locus	Metric	118.8	22	0.943	0.95	0.04
Locus	Scalar	134.2	34	0.941	0.97	0.04
Bullying	Configural	160.1	47	0.977	0.97	0.03
Bullying	Metric	160.7	59	0.979	0.98	0.03
Bullying	Scalar	182.8	71	0.977	0.98	0.03
AUDIT-C	Configural[a]	210.8	3	0.954	0.86	0.17
AUDIT-C	Metric[a]	231.0	7	0.950	0.94	0.11
AUDIT-C	Scalar[a]	223.2	13	0.953	0.97	0.08

[a] Models show estimation issues because of negative variances.
[*] ΔCFI > 0.01

4.3 Partial Equivalence (RQ2)

We next investigate the causes for the lack of scalar equivalence in the Leisure A scale. In addition to the previous models, we investigate the constraints that lead to a decrease in fit when applied to the scalar equivalence model. Using modification indices, we cumulatively free coefficients until the fit of the model is not significantly worse than the metric model based on the ΔCFI > 0.01 criterion.

Based on this procedure, three thresholds, one for each of three items (cinema/live performances, leisure activity groups, and sport/exercise) are significantly different across modes. In all the cases, the Web responses are significantly different from the other two modes (Table 2).

To better understand the sizes and potential causes of these differences, we show the estimated probabilities of endorsing each category for each of the three items based on the final partial equivalence model. Based on Figure 2, it is apparent that Web respondents are significantly more likely to engage in recreational activities like going to cinema or group activities than telephone and face-to-face respondents, a response pattern that is indicative of less recency. In contrast, Web respondents engage in sport or exercise less frequently than telephone and face-to-face respondents, which is more in line with a social desirability effect.

4.4 Means of Latent Variables (RQ3)

Lastly, we investigate the means of the latent variables to understand the extent to which other mode differences are present in the models (Figure 3). To estimate them, we use the Web mode as the reference group and calculate how different are the means in the other two modes for each latent variable. 95% confidence intervals are shown to better assess whether there are differences between the modes.

Overall, there are some differences in the means of the latent variables between modes for some scales, with the exceptions of the AUDIT-C and Bullying scales, which are comparable between all three modes (i.e. their confidence intervals overlap). Face-to-face and Web differ in three out of seven scales as do telephone and Web. Face-to-face and telephone are slightly more similar to each other with differences appearing only for the two leisure scales. For the leisure scale items, the telephone mode shows a greater tendency for social desirability effects than Web and face-to-face.

Both telephone and face-to-face differ from Web for only one scale, the GHQ scale. Given the sensitive nature of the GHQ items, the observed pattern could be due to socially desirable responding as telephone and face-to-face respondents were more likely to endorse the socially desirable responses

Table 2
Identification of partial equivalence across modes for leisure scales A and B. Coefficients were cumulatively freed based on the highest modification indice.

Scale	Model	Cumulatively freed coefficient	X^2	df	CFI	TLI	RMSEA
Leisure A	Configural		412.7	41	0.919	0.91	0.06
Leisure A	Metric		395.0	51	0.925	0.93	0.05
Leisure A	Scalar		611.3	71	0.882	0.93	0.06
Leisure A	Scalar partial	F; T 3; Web[a]	508.5	70	0.904	0.94	0.05
Leisure A	Scalar partial	B; T 3; Web[a]	470.7	69	0.912	0.94	0.05
Leisure A	Scalar partial	A; T 1; Web[a]	439.7	68	0.919	0.95	0.05

[a] The notation refers to: the scale item; the threshold; and the mode in which it was freed.

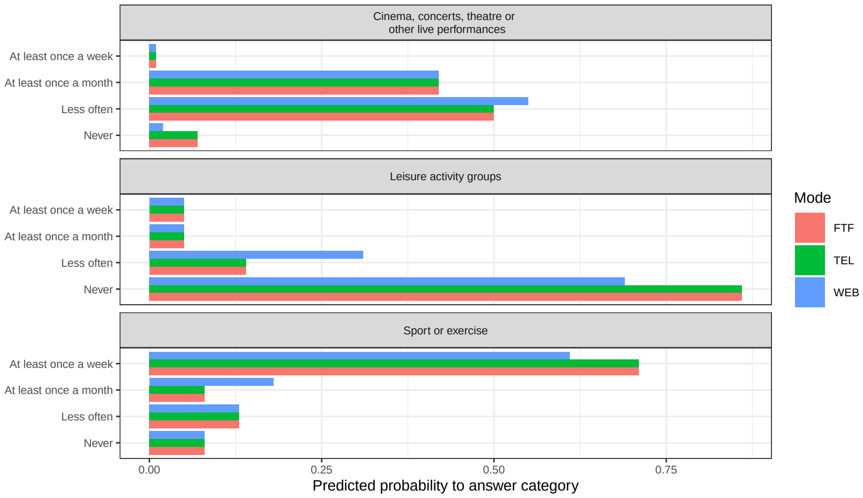

Figure 2. Predicted probabilities for endorsing a category conditional on the latent factor score, by mode, for selected Leisure Scale A items. Coefficients are based on the thresholds from the final partial equivalence model.

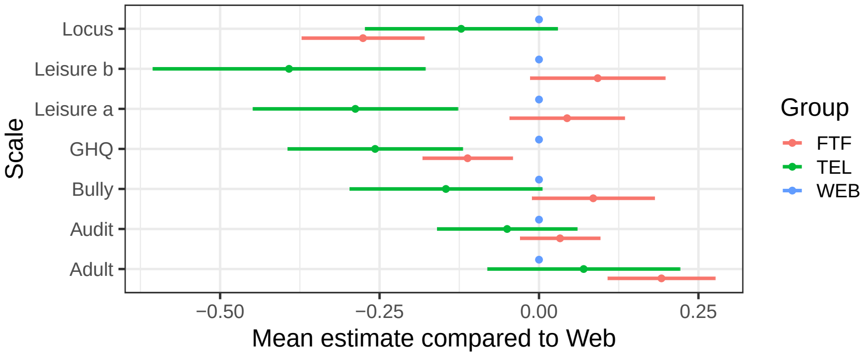

Figure 3. Latent variable means by scale and mode. The mean of the latent variable in the Web group is used as the reference and fixed at 0. Results based on the full scalar model.

(first few categories).

For the Adult scale, the latent mean for face-to-face is larger than for Web. This suggests that face-to-face respondents were more likely to choose the last few categories, indicating stronger adulthood development, than Web respondents, which could be driven either by social desirability or recency effects. For the Locus of Control scale, face-to-face respondents were more likely than Web respondents to choose the first few categories. A potential explanation for this pattern is unclear as the scale was self-administered in both modes.

To summarize, we find some differences in the latent means between the three modes. These differences could be driven by residual systematic measurement error (e.g. social desirability) that is not severe enough to preclude scalar equivalence.[2] Another possible cause of the differences could be due to residual selection (e.g. healthier people might answer in a particular mode) that is not accounted for in the weighting procedure. To partially investigate these two potential causes, the latent means were re-estimated with and without the selection weights (Figure A2 in the online appendix) and with and without fitting the partial equivalence model for the Leisure A scale (results not shown). While these approaches shifted the latent means slightly, they did not manage to explain the differences in the means between the modes. Given that the selection weights were previously shown to be effective in reducing selection bias in the mode groups, but have little effect on the equivalence results, we suspect that the latent mean differences are most likely caused by residual systematic measurement error that is not accounted for in the measurement models. Though, our research design does not allow us to explicitly test this claim.

5 Discussion

This study evaluated the assumption of cross-mode measurement equivalence for seven multi-item scales collected in a sequential mixed-mode (Web-to-telephone-to-face-to-face) survey: the UK Next Steps wave 8/Age 25 cohort study. Several well-known scales were evaluated, including AUDIT-C, Locus of Control, GHQ-12, and Adult Identity Resolution as well as a formative Bullying scale and two formative leisure scales. To our knowledge, no study has assessed cross-mode measurement equivalence for these scales in a sequential mixed-mode setting.

The evaluation yielded two principal findings. First, after controlling for selection through an extensive data-driven weighting procedure, we found that nearly all scales achieved full scalar measurement equivalence across the three modes. The only exception was a formative leisure scale, which showed nonequivalence in the conditional means (thresholds) for the Web mode. A closer analysis revealed that Web respondents were more likely to engage in going to the cinema or group activities compared to respondents in the other modes, but engaged in sport or exercise less frequently than their telephone and face-to-face counterparts. The latter pattern is in line with socially desirable responding, whereas the former is more consistent with less recency. However, after freeing the problematic threshold restrictions, this leisure scale achieved partial scalar equivalence.

Secondly, while many of the latent means were found to be similar between different modes, there were some differences for some mode comparisons. Most of these differences were observed between Web and face-to-face or between Web and telephone, with fewer differences between face-to-face and telephone, which is consistent with findings from the mixed-mode literature (Cernat et al., 2016; Heerwegh & Loosveldt, 2011; Klausch et al., 2013). Some differences pointed to socially desirable responding: for example, the GHQ scale, which is known to be affected by socially desirable responding (Ormel, Koeter, & van den Brink, 1989; Parkes, 1980; Roustaei, Jafari, Sadeghi, & Jamali, 2015), yielded fewer socially desirable responses in the self-administered Web mode than in telephone and face-to-face. A similar pattern was found for the two leisure scales, which suggested more socially desirable responding in telephone than in the Web mode. Other differences were unclear as the latent means suggested multiple possible mechanisms, including social desirability or recency effects. These results showed that, despite achieving scalar equivalence, there may be other sources of systematic measurement error that are not accounted for in the measurement models.

Overall, it is encouraging that we find full (or at least partial) scalar measurement equivalence for all scales administered in a well-known, population-based sequential mixed-mode survey. While previous studies have shown that mixing self- and interviewer-administered modes in sequential mixed-mode surveys can give rise to both measurement and selection effects (Klausch et al., 2013; Sakshaug, Cernat, & Raghunathan, 2019; Schouten, van den Brakel, Buelens, van der Laan, & Klausch, 2013), it is reassuring that such mixed-mode designs do not necessarily preclude the collection of equivalent or comparable latent variable measurements derived from multi-item scales. Despite recommendations against mixing aural/visual modes or self-/interviewer-administered modes, practical reasons (e.g. costs) often dictate the use of such designs in survey research, particularly designs which exploit relatively inexpensive online data collection. The finding that mixing a Web mode to an otherwise interviewer-administered survey did not compromise on measurement equivalence is therefore advantageous from

[2] It is entirely plausible for the latent means to differ despite claiming scalar equivalence as the mean of the latent variable represents the overall average for the concept given the measurement model, whereas the intercept/threshold represents the conditional average of the observed variables when the latent mean is zero (Byrne et al., 1989).

both a practical and methodological perspective. However, we caution researchers that mixing Web and interviewer-administered modes may still produce unwanted measurement effects that affect the comparability of the latent means in each mode. This interesting finding merits further research to identify the specific factors (e.g. systematic measurement error components) that explain differences in the latent means between different modes, even when scalar equivalence is attained. Lastly, it is reassuring that we find strong equivalence in CASI items and also for some sensitive items (e.g. GHQ-12), as this suggests that implementing CASI in face-to-face interviews may minimize measurement differences with Web.

As with all studies, this one has limitations which could be addressed in future work. For instance, it would be prudent to attempt replication of these results in other sequential mixed-mode studies including those based on cross-sectional samples and other target populations, such as older populations which may not be as Web-savvy as the younger population studied here. Determining whether measurement equivalence can be established for other commonly-used multi-item scales in a sequential mixed-mode setting would also be beneficial. Lastly, despite employing an extensive data-driven variable selection and weighting procedure drawing from the entire pool of variables collected in all prior waves of Next Steps to adjust for mode-specific nonresponse, it is still possible that unobserved variables influenced the selection process and that residual selection effects remained even after applying the weighting adjustment. Although we cannot test for this possibility, we encourage researchers to make full use of all observed information (which, in the case of longitudinal studies, can be immense, but even for cross-sectional surveys, basic demographic and background information seem to be important), in order to make the Missing at Random assumption more plausible when analyzing measurement effects. The data-driven weighting strategy that we employed is one possible strategy.

6 Acknowledgement

This project was funded by a grant from the UK Economic and Social Research Council (ESRC) [ES/T00116X/1].

References

American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys.* (9th ed.). American Association for Public Opinion Research.

Ansolabehere, S., & Schaffner, B. F. (2014). Does survey mode still matter? Findings from a 2010 multi-mode comparison. *Political Analysis*, *22*, 285–303.

Ashby, J., Kottman, T., & Draper, K. (2002). Social interest and locus of control: Relationships and implications. *The Journal of Individual Psychology*, *58*, 52–61.

Bailey, J., Breeden, J., Jessop, C., & Wood, M. (2017). *Next steps age 25 survey: Technical report*. London, UK: NatCen Social Research.

Bianchi, A., Biffignandi, S., & Lynn, P. (2017). Web-face-to-face mixed-mode design in a longitudinal survey: Effects on participation rates, sample composition, and costs. *Journal of Official Statistics*, *33*, 385–408.

Bollen, K. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*, 303–316.

Bollen, K., & Lennox, R. (1991). Conventional wisdom in measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*, 281–291.

Bradley, K., Bush, K., Epler, A., Dobie, D., Davis, T., Sporleder, J., … Kivlahan, D. R. (2003). Two brief alcohol-screening tests from the alcohol use disorders identification test (AUDIT): Validation in a female veterans affairs patient population. *Archives of Internal Medicine*, *163*, 821–829.

Bush, K., Kivlahan, D., McDonell, M., Fihn, S., & Bradley, K. A. (1998). The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. *Archives of Internal Medicine*, *158*, 1789–1795.

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466.

Cernat, A., Couper, M., & Ofstedal, M. B. (2016). Estimation of mode effects in the health and retirement study using measurement models. *Journal of Survey Statistics and Methodology*, *4*, 501–524.

Cernat, A., & Revilla, M. (2021). Moving from face-to-face to a web panel: Impacts on measurement quality. *Journal of Survey Statistics and Methodology*, *9*, 745–763.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*, 641–678.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464–504.

Christian, L., Dillman, D., & Smyth, J. D. (2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. In J. Lepkowski, C. Tucker, J. Brick, E. de Leeuw, L. Japec, P. Lavrakas, … R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 250–275). Hoboken, NJ: Wiley.

Côté, J. E. (1996). Sociological perspectives on identity formation: The culture—identity link and identity capital. *Journal of Adolescence*, *19*, 417–428.

Côté, J. E., & Levine, C. G. (2002). *Identity formation, agency, and culture: A social psychological synthesis*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Côté, J. E., Mizokami, S., Roberts, S., & Nakama, R. (2016). An examination of the cross-cultural validity of the identity capital model: American and Japanese students compared. *Journal of Adolescence*, *46*, 76–85.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75.

Davidov, E., Schmidt, P., Billiet, J., & Meuleman, B. ((2018). *Cross-cultural analysis: Methods and applications*. New York: Routledge.

De Leeuw, E. (1992). *Data quality in mail, telephone and face to face surveys*. Amsterdam: TT Publications.

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

De Leeuw, E., Mellenbergh, G., & Hox, J. J. (1996). The influence of data collection method on structural models: A comparison of a mail, a telephone, and a face-to-face survey. *Sociological Methods & Research*, *24*, 443–472.

DeMaio, T. J. (1984). Social desirability and survey measurement: A review. In C. Turner & E. Martin (Eds.), *Surveying subjective phenomena: Volume 2* (pp. 257–282). New York: Russell Sage Foundation.

Dillman, D., & Mason, R. G. (1984). *The influence of survey method on question response*. Paper Presented at the Annual Meeting of the American Association for Public Opinion Research. Delavan, WI.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Social Science Research*, *38*, 1–18.

Enders, C. E. (2010). *Applied missing data analysis*. New York: Guilford.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*, 370–392.

Goldberg, D., & Williams, P. (1988). *A user's guide to the general health questionnaire*. Windsor: NFER-Nelson.

Gordoni, G., Schmidt, P., & Gordoni, Y. (2012). Measurement invariance across face-to-face and telephone modes: The case of minority-status collectivistic-oriented groups. *International Journal of Public Opinion Research*, *24*, 185–207.

Greene, J., Speizer, H., & Wiitala, W. (2008). Telephone and web: Mixed-mode challenge. *Health Services Research*, *43*, 230–248.

Hamer, M., Chida, Y., & Molloy, G. J. (2009). Psychological distress and cancer mortality. *Journal of Psychosomatic Research*, *66*, 255–258.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*, 111–121.

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*, 836–846.

Heerwegh, D., & 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*, 49–63.

Holbrook, A., Green, M., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*, 79–125.

Holbrook, A., Krosnick, J., Moore, D., & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly*, *71*, 325–348.

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., & Jäckle, A. (2014). *The role of the interviewer in producing mode effects: Results from a mixed modes experiment comparing face-to-face, telephone and web administration*. ISER Working Paper Series (No. 2014-20), University of Essex.

Hox, J., De Leeuw, E., & Zijlmans, E. A. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, *6*, 1–11.

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

Jackson, C. (2007). The general health questionnaire. *Occupational Medicine*, *57*, 79.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Kappelhof, J., & De Leeuw, E. (2019). Estimating the impact of measurement differences introduced by efforts to reach a balanced response among non-western minorities. *Sociological Methods & Research*, *48*, 116–155.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail-web mixed-mode surveys. *Social Science Computer Review*, *37*, 214–233.

King, W., Chen, J., Mitchell, J., Kalarchian, M., Steffen, K., Engel, S., … Yanovski, S. Z. (2012). Prevalence of alcohol use disorders before and after bariatric surgery. *Journal of the American Medical Association*, *307*, 2516–2525.

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

Knowles, E., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*, 379–386.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*, 847–865.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201–219.

Little, R. J. A., & Rubin, D. B. (1989). The analysis of social-science data with missing values. *Sociological Methods & Research*, *18*, 292–326.

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

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research*, *20*, 60–103.

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

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York: Routledge.

Mostafa, T., Narayanan, M., Pongiglione, B., Dodgeon, B., Goodman, A., Silverwood, R., & Ploubidis, G. (2021). Missing at random assumptions made more plausible: Evidence from the 1958 british birth cohort. *Journal of Clinical Epidemiology*, *136*, 44–54.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide.* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A., & Lynn, P. (2015). Revisiting "yes/no" versus "check all that apply": Results from a mixed modes experiment. *Survey Research Methods*, *9*, 189–204.

Ormel, J., Koeter, M. W. J., & van den Brink, W. (1989). Measuring change with the general health questionnaire (GHQ): The problem of retest effects. *Social Psychiatry and Psychiatric Epidemiology*, *24*, 227–232.

Parkes, K. R. (1980). Social desirability, defensiveness and self-report psychiatric inventory scores. *Psychological Medicine*, *10*, 735–742.

Ploubidis, G., McElroy, E., & Moreira, H. C. (2019). A longitudinal examination of the measurement equivalence of mental health assessments in two British birth cohorts. *Longitudinal and Life Course Studies*, *10*, 471–489.

Reinert, D., & Allen, J. P. (2007). The alcohol use disorders identification test: An update of research findings. *Alcoholism: Clinical and Experimental Research*, *31*, 185–199.

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

Revilla, M. (2015). Comparison of the quality estimates in a mixed-mode and a unimode design: An experiment from the European Social Survey. *Quality & Quantity*, *49*, 1219–1238.

Revilla, M., & Saris, W. E. (2013). A comparison of the quality of questions in a face-to-face and a web survey. *International Journal of Public Opinion Research*, *25*, 242–253.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373.

Roberts, C., Joye, D., & Ernst Stähli, M. (2016). Mixing modes of data collection in swiss social surveys: Methodological report of the LIVES-FORS mixed mode experiment. *LIVES Working Papers*, *48*, 1–42.

Roustaei, N., Jafari, P., Sadeghi, E., & Jamali, J. (2015). Evaluation of the relationship between social desirability and minor psychiatric disorders among nurses in Southern Iran: A robust regression approach. *International Journal of Community Based Nursing & Midwifery*, *3*, 301–308.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31–57.

Sakshaug, J., Cernat, A., & Raghunathan, T. (2019). Do sequential mixed-mode surveys decrease nonresponse bias, measurement error bias, and total bias? An experimental study. *Journal of Survey Statistics and Methodology*, *7*, 545–571.

Sakshaug, J., Yan, T., & Tourangeau, R. (2010). Nonresponse error, measurement error, and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Public Opinion Quarterly*, *74*, 907–933.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, *42*, 1555–1570.

Schwarz, N., Strack, F., Hippler, H., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, *5*, 193–212.

Shepherd, S., Owen, D., Fitch, T., & Marshall, J. L. (2006). Locus of control and academic achievement in high school students. *Psychological Reports*, *98*, 318–322.

Silverwood, R., Calderwood, L., Sakshaug, J., & Ploubidis, G. (2020). A data driven approach to understanding and handling non-response in the next steps cohort. CLS Working Paper 2020/5. London: UCL Centre for Longitudinal Studies. Retrieved from https://cls.ucl.ac.uk/wp-content/uploads/2020/04/CLS-working-paper-2020-5-A-data-driven-approach-to-understanding-and-handling-non-response-in-the-Next-Steps-cohort.pdf

Smyth, J., Olson, K., & Kasabian, A. S. (2014). The effect of answering in a preferred versus a non-preferred survey mode on measurement. *Survey Research Methods*, *8*, 137–152.

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Tarnai, J., & Dillman, D. (1992). Questionnaire context as a source of response differences in mail and telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research*. New York: Springer.

Tomé, R. (2018). The impact of mode of data collection on measures of subjective wellbeing. PhD Thesis, University of Lausanne. Retrieved from https://serval.unil.ch/fr/notice/serval:BIB%5C_F89D8660FBE7

Tourangeau, R., Conrad, F., & Couper, M. P. (2013). *The science of web surveys*. New York: Oxford University Press.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*, 275–304.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859–883.

University College London, C. f. L. S., UCL Institute of Education. (2021). Next steps: Sweeps 1-8, 2004–2016. [Data Collection]. 16th Edition. UK Data Service. SN: 5545. Retrieved from http://doi.org/10.5255/UKDA-SN-5545-8

Vandenberg, R., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Vannieuwenhuyze, J., 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*, 306–322.

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

Wagner, J., Arrieta, J., Guyer, H., & Ofstedal, M. (2014). Does sequence matter in multimode surveys: Results from an experiment. *Field Methods*, *26*, 141–155.

West, S., Taylor, A., & & Wu, W. (2012). Model fit and model selection in structural equation modelling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (1st ed., pp. 209–231). New York: Guilford Press.

Widaman, K., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Ye, C., Fulton, J., & Tourangeau, R. (2011). More positive or more extreme? A meta-analysis of mode differences in response choice. *Public Opinion Quarterly*, *75*, 349–365.