Article

International Journal of Market Research

Measurement equivalence in probability and nonprobability online panels

International Journal of Market Research
2022, Vol. 64(4) 484–505

DOI: 10.1177/14707853221085206
journals.sagepub.com/home/mre


Hafsteinn Einarsson
The University of Manchester, Manchester, UK

Joseph W Sakshaug
Institute for Employment Research, Nuremberg, Germany; Ludwig Maximilian University of Munich, Nuremberg, Germany; University of Mannheim, Mannheim, Germany

Alexandru Cernat
The University of Manchester, Manchester, UK

Carina Cornesse
German Institute for Economic Research, Berlin, Germany; Research Institute Social Cohesion, Bremen, Germany; University of Mannheim, Mannheim, Germany

Annelies G Blom
University of Mannheim, Mannheim, Germany; University of Bergen, Bergen, Norway

Abstract

Nonprobability online panels are commonly used in the social sciences as a fast and inexpensive way of collecting data in contrast to more expensive probability-based panels. Given their ubiquitous use in social science research, a great deal of research is being undertaken to assess the properties of nonprobability panels relative to probability ones. Much of this research focuses on selection bias, however, there is considerably less research assessing the comparability (or equivalence) of measurements collected from respondents in nonprobability and probability panels. This article contributes to addressing this research gap by testing whether measurement equivalence holds between multiple probability and nonprobability online panels in Australia and Germany. Using equivalence testing in the Confirmatory Factor Analysis framework, we assessed measurement equivalence in six multi-item scales (three in each country). We found significant measurement differences between probability and nonprobability panels and within them, even after weighting by demographic variables. These results suggest that combining or comparing multi-item scale data

Corresponding author:
Hafsteinn Einarsson, Social Statistics, The University of Manchester, Humanities Bridgeford Street, Manchester M13 9PL, UK.
Email: hafsteinn.einarsson@postgrad.manchester.ac.uk

from different sources should be done with caution. We conclude with a discussion of the possible causes of these findings, their implications for survey research, and some guidance for data users.

Keywords
confirmatory factor analysis, latent variables, multi-item scales, panel vendors, web surveys

Introduction

For over a decade, there has been a prominent rise in the use of nonprobability online panels for survey data collection (Callegaro, Baker et al., 2014). In contrast to traditional probability-based surveys, where units are directly sampled at random from the target population with a known (or knowable) probability of selection, nonprobability online panels are often characterized by an additional layer of pre-selection. The pre-selection step occurs when persons are recruited, usually through mass advertising (e.g., via websites, pop-up ads, and sponsored search results), offering the chance to participate in periodic web surveys, often in exchange for small gifts or monetary rewards. The selection process is therefore dependent on people reacting to passive advertising and self-selecting into the panel. This setup allows panel providers to build massive pools of pre-selected panelists, from which large samples of ready and willing respondents can be drawn and surveyed at short notice, usually at a fraction of the cost of drawing and recruiting traditional probability samples (Callegaro, Villar et al., 2014). Such panels are commonly used in the commercial sector, and increasingly also in the non-profit and academic sectors, to study attitudes, preferences, and behaviors with a growing body of this work making its way into the peer-reviewed literature (e.g., Hitchman et al., 2015; Powell et al., 2011; Skitka & Sargis, 2006; Taichman, 2020).

Despite their widespread use, nonprobability online panels have been criticized on the basis that they—and the samples drawn from them—do not accurately represent the general population or even the population of Internet users (Callegaro, Villar et al., 2014; Cornesse et al., 2020; Lehdonvirta et al., 2020). Indeed, most comparison studies find that samples drawn from nonprobability web panels are less accurate with respect to population benchmarks than probability online and offline samples, even when quota sampling or post-survey adjustments are used (Cornesse et al., 2020; Loosveldt & Sonck, 2008; Pasek, 2016), though the extent of inaccuracies tends to vary across panel providers (Blom, Ackermann-Piek et al., 2017; Kennedy et al., 2016).

These discrepancies are usually attributed to selection bias due to the non-random recruitment of panel members (Cornesse et al., 2020). However, what is largely missing from the literature are studies that investigate differences in survey measurements collected between probability and nonprobability online panels. The concern is that nonprobability respondents who join such panels may take less care in answering the survey items, as their motivation for participating in the survey may differ from those who participate in probability-based surveys (Cornesse & Blom, 2020). Although panels can discourage undesirable response behaviors (e.g., straight-lining) by employing different controlling procedures, the extent of controlling may vary from panel to panel. Such behavior raises the question of whether responses to multi-item scales, and the latent (or "hidden") constructs they aim to measure, are measured in the same way between probability and nonprobability samples, as well as across different nonprobability sample providers.

To address these open issues, we carried out an investigation of measurement equivalence in several multi-item attitudinal scales administered in parallel probability and nonprobability online panel surveys. Measurement equivalence is a necessary assumption to make valid comparisons

between different groups of respondents on multi-item scale measurements and latent constructs. If measurement equivalence holds, then the means of latent variables and relationships between them can be validly compared across different groups (Baumgartner & Steenkamp, 2006; Meredith, 1993). Using Confirmatory Factor Analysis, we formally test for measurement equivalence between probability and nonprobability online surveys, and between nonprobability surveys conducted by different online panel vendors. If measurement equivalence can be established between the groups studied, then data from probability and nonprobability online panels can be validly compared and combined, but if it cannot such comparisons become problematic and data users may need to adjust their analyses accordingly.

Previous research

Several studies have explored differences and similarities between probability and nonprobability sample surveys. These studies have typically focused on assessing how accurately the surveys represent the general population by comparing the survey estimates to "gold standard" external benchmark data (Cornesse et al., 2020). The key finding from this literature is that probability surveys tend to produce more accurate population estimates than nonprobability surveys, even after applying weighting adjustments (e.g., Blom, Ackermann-Piek et al., 2017; Dutwin & Buskirk, 2017; Macinnis et al., 2018; Sturgis et al., 2018).

Some researchers have suggested that differences between probability and nonprobability surveys occur because probability surveys are more likely to conduct interviews offline (e.g., via face-to-face or telephone interviewing) and cover a broader portion of the general population than nonprobability sample surveys, which are mostly conducted online and restricted to the population of Internet users (Ansolabehere & Schaffner, 2014; Berrens et al., 2003). However, studies aiming to disentangle mode effects and sampling designs have found that both offline and online probability surveys are more accurate than nonprobability surveys (Blom, Ackermann-Piek et al., 2017; Brüggen et al., 2016; Chang & Krosnick, 2009; Macinnis et al., 2018; Scherpenzeel & Bethlehem, 2011; Yeager et al., 2011).

While several studies have examined differences in sample accuracy between probability and nonprobability surveys, little research has focused on other aspects, such as measurement differences, which may contribute to the observed discrepancies. Findings from the sparse literature on measurement differences are mixed. Concerning single items and multi-item scales, Chang and Krosnick (2009) identified a range of measurement differences when comparing a random-digit dial (RDD) telephone survey to a probability online panel survey and a nonprobability online panel survey. These included random measurement error (the nonprobability sample produced more reliable estimates of candidate preferences than the probability online panel survey, which in turn was less reliable than the RDD survey), satisficing (the nonprobability online survey produced considerably lower rates of midpoint selections, followed by the probability online survey, while the RDD survey produced the highest rates), and social desirability bias (the nonprobability online survey produced fewer socially undesirable answers than the probability online survey, which provided fewer socially undesirable answers than the RDD survey). The authors concluded that nonprobability online surveys produce higher measurement quality than RDD surveys, but at the cost of lower representativeness, while probability online surveys produce the optimal combination of measurement quality and representativeness. Cornesse and Blom (2020) found that three probability online surveys consistently produced less straight-lining in grid questions than seven nonprobability online surveys. Although two other undesirable survey behaviors, item nonresponse and midpoint selection, were not significantly different across the surveys. Furthermore, Greszki

et al. (2014) found that a probability online survey performed better in terms of minimizing "speeding" (i.e., answering survey questions faster than normal) compared to a nonprobability online survey.

Despite the limited number of studies on measurement differences between probability and nonprobability surveys, it is reasonable to expect that measurement differences between the samples, and not just differential selection bias, influence their overall accuracy. Unlike probability surveys, which rely on a range of established and theoretically proven sampling procedures (Kish, 1965; Lohr, 2019), nonprobability surveys usually recruit their participants from a pool of volunteers on the Internet using online advertisements, pop-up questionnaires on websites, or open invitations via email lists that generally promise monetary incentives to anyone who voluntarily registers to join the panel (see Callegaro, Villar et al., 2014 for an overview of nonprobability online panel recruitment procedures). These findings illustrate the risks of incentive-driven survey recruitment which may affect response behavior by motivating undesirable response styles, such as speeding through the questionnaire.

Research on online access panels indicates that the promised monetary incentive is the most important motivator for people to join the panel and the strongest predictor of subsequent survey participation (Keusch et al., 2014; Sparrow, 2006). This is supported by Sparrow (2006), who showed that 52% of respondents to the ICM online panel primarily joined because "they felt it would be an enjoyable way to earn money or enter prize draws." Similar results were reported by Zhang et al. (2019), who found that "professional respondents", defined as those who were registered in at least seven panels, were more likely to report "for money" as their main reason for joining the panel, compared to less-experienced respondents. Interestingly, the authors found that the professional respondents produced higher-quality responses than their more novice counterparts, suggesting that they may take the survey response task more seriously. Furthermore, repeated survey participation by professional respondents may produce panel conditioning, where the respondent's response behavior changes as they become more familiar with the survey and questionnaire (Buck et al., 1977; Hillygus et al., 2014; Struminskaya & Bosnjak, 2021).

Given that respondents likely differ in their motivations for participating in probability and nonprobability online surveys as well as the associated risk of more undesirable response behaviors (e.g., straight-lining) in the latter, it is conceivable that multi-item scales and latent constructs may not be measured in the same way in nonprobability and probability surveys. Furthermore, given the multitude of nonprobability panel providers using different recruitment protocols, incentives, and methodologies, it is plausible that the measurement structure of latent variables may vary across different online panel providers. Currently, no research has been published on this issue. Of course, differences in the measurement structure might be attributed to the composition of the respondents who self-select themselves into the panels. That is, respondents in one panel may have different characteristics to respondents in a different panel, which may drive measurement differences if these characteristics correlate with different response behaviors. Thus, balancing the composition of the samples drawn from different panel providers is an important consideration. While full balancing is unlikely to be achieved in practice, weighting adjustments can help standardize the sample with respect to some observable characteristics and this may improve the comparability of the group measurements (Hox et al., 2015). We consider this issue in our investigation.

Research questions

Given the proliferation of nonprobability online panel surveys, the reviewed literature highlights a research gap in exploring differences in measurement between these surveys and probability surveys. On one hand, if measurement equivalence is attained in multi-item scales, then researchers

can be more confident that latent constructs are measured in the same way in both probability and nonprobability surveys. On the other hand, if measurement equivalence is unattainable, then this could raise serious issues when combining or comparing results from both types of surveys. Similar issues could arise within a particular survey type, for instance, when scale data collected from multiple nonprobability panel providers are not equivalently measured. Using demographic weights to account for differential selection may help to improve measurement equivalence if the likelihood of survey participation varies among demographic groups and is related to response behaviors; thus, we also assess the effects of weighting on measurement equivalence. We address these research gaps by analyzing measurement equivalence in probability and nonprobability panel surveys in Australia and Germany. Specifically, we address the following research questions:

Q1: Are multi-item scale measurements equivalent between probability and nonprobability panel surveys?
Q2: Are multi-item scale measurements equivalent between different nonprobability panel providers?
Q3: Does weighting by demographics improve measurement equivalence between probability and nonprobability surveys, and between different nonprobability panel providers?

Data and methods

Probability surveys: Australia

We utilize two probability-based surveys from the Social Research Centre's 2015 Online Panels Benchmarking Study in Australia (Pennay et al., 2016). The first is an address-based sample (ABS) survey. The sample was drawn from a national address index for Australia (the Geocoded National Address File) using a stratified sampling design. The ABS survey allowed for multiple modes of completion. Printed questionnaires were mailed to all households, and a link to the online version was provided for those who preferred to complete the survey online. The cover letter invited the household member (aged 18 years or over) with the next or most recent birthday to complete the survey. Telephone follow-ups were conducted with those who did not respond via hard copy or online. Data collection took place between 6th November and 23rd December 2015. A total of 2,050 households were contacted and 538 persons completed the survey, which resulted in a response rate of 26.2% (based on Response Rate 3; AAPOR, 2016). A total of 208 persons completed the survey online, 202 completed the hard copy version, and 128 were interviewed via telephone.

The second probability-based sample consisted of persons who previously participated in the Australian National University Poll (ANU Poll), a dual-frame RDD survey. Respondents of the ANU Poll conducted in October 2015 were invited to take part in a "future study about health and wellbeing." Those who agreed were asked for contact details which, depending on their preferences, were used to either email a link to complete the survey online or send a hard copy questionnaire to be returned by mail. The October 2015 ANU Poll used a stratified sample design and the "next birthday method" to select the target respondent. Among the 1,200 respondents in the October 2015 ANU Poll who were invited to take part in the "future study," 693 (58%) agreed to participate and provided an email or postal address for distribution of the questionnaire. Telephone interviews were available if sample members had not responded online or via hard copy. Data collection took place between 19th October 2015 and 11th December 2015. A total of 560 persons completed the survey for a response rate of 80.8% (AAPOR RR1) based on the initial ANU Poll, with 292 online responses, 40 hard copy responses, and 228 telephone responses.

For both Australian probability surveys, the original survey weights were constructed in two steps. First, design weights were calculated to account for respondents having different selection probabilities. In the second step, the design weight was combined with raking weights using known distributions of key sociodemographic characteristics (telephone status, education by age, region, gender, country of birth, age group, and state) based on official statistics published by the Australian Bureau of Statistics. As our study is focused on measurement equivalence in online surveys, all hard copy or telephone respondents are excluded from the analysis to remove mode effects, limiting the analysis to those who competed the web surveys. We further adjust the survey weights accordingly using a standard propensity score weighting procedure to account for selection into the web mode (for more details, see the "Accounting for Selection into Web" section below).

Nonprobability surveys: Australia

Eight nonprobability panel providers were approached to conduct a nationally representative survey of ~600 respondents from their respective panels. Five panel providers responded within the deadline and met the study criteria. The study team did not provide instructions on how "representativeness" should be achieved. All providers implemented quota sampling using age, sex, and geographic information. Panelists were invited via email to participate in the survey (see Pennay et al., 2016, for further details of the recruitment process). Online data collection for the various nonprobability surveys took place in late November and early December 2015. Design weights were neither provided by the nonprobability panel providers, nor were any generated specifically for the purposes of the study as the probabilities of selection were unknown. Thus, each nonprobability record is assigned a design weight of 1. Raking weights were generated for the Australian nonprobability samples using the same benchmark data as were used for the probability surveys. Information on the number of respondents, quota variables, and fieldwork periods for each of the five nonprobability surveys and the two probability surveys are provided in Table 1.

Probability surveys: Germany

The probability surveys from Germany consist of the German Internet Panel (GIP) and the GESIS Panel, both population-based panel surveys representative of the general population. The GIP is an ongoing longitudinal household panel survey of persons residing in Germany, aged 16–75. A multistage stratified area probability design was used to select the sample of households. Each sampled household was approached by face-to-face interviewers for an initial recruitment survey. The recruitment interview was conducted with a non-randomly selected member of the household, who

Table 1. List of Australian probability and nonprobability surveys.

Survey	No. respondents	Quota variables	Fieldwork period
ABS	538	N/A	6th November–18th December 2015
ANU Poll	560	N/A	19th October–11th December 2015
NP Panel 1	601	State, region, age, gender	11th–18th December 2015
NP Panel 2	600	State, region, age, gender	30th November–7th December 2015
NP Panel 3	626	State, region, age, gender	30th November–6th December 2015
NP Panel 4	630	State, region, age, gender	2nd–7th December 2015
NP Panel 5	601	State, region, age, gender	14th–17th December 2015

provided information on all other household members during the interview. After the recruitment interview, all household members within the GIP age range (i.e., between 16 and 75) were invited to register to the panel online (Blom et al., 2015).

To facilitate coverage of the whole population, offline households were provided with Internet access and/or an Internet-capable browsing device if they did not already have one or both (Blom, Herzing et al., 2017). Participants were recruited in two independent recruitment rounds, initially in 2012 with a response rate of 18.5% (AAPOR RR2), followed by a second recruitment round in 2014 with a response rate of 20.5% (AAPOR RR 2). Every two months panel members are invited to login and complete a web survey containing a range of question modules on social and political issues, typically completed within 20–25 minutes. We use data from the March 2015 wave of the GIP, in which 68.7% of panelists (or 3,426 out of 4,989) completed the web survey and include only those respondents who matched the age range of the GESIS Panel and nonprobability samples (18–70 years, as discussed below).

The GESIS Panel is a mixed-mode (web and paper) panel survey of adults (18–70 years) residing in Germany. We analyze data from panel members initially recruited in 2013. A multi-stage stratified probability sampling design using municipal population registers was used to select the initial sample. Face-to-face recruitment interviews were conducted in which all participants were asked to join the GESIS panel. Of the initial sample of 21,870 individuals, 6,210 (28.4%) agreed to join the panel (AAPOR RR1). Panel members are invited to complete a web survey (with a mailed questionnaire option available for those who are unable or unwilling to respond online, see Cornesse and Schaurer (2021) for more information) every two months, with estimated average completion times of 20 minutes for each survey. The core questionnaire modules contain items on values, political behavior, well-being, among others. The questionnaire module used in our analyses was approved by the GESIS Panel team and fielded from 8th February to 14th April 2015. The completion rate among panelists for this module was 61.5% (3,822 out of 6,210). A detailed description of the methodology for the GESIS Panel can be found in (Bosnjak et al., 2018) and a comparison of the GESIS Panel recruitment design to the GIP and other European probability-based online panels can be found in (Blom et al., 2016). As for the Australian mixed-mode surveys, we exclude paper responses and focus only on the web responses in the forthcoming analysis of the GESIS Panel data and adjust the survey weights for selection into the web mode (for more details, see the "Accounting for Selection into Web" section below).

Nonprobability surveys: Germany

Eight nonprobability panel vendors (out of 17 total bids) were contracted by the GIP team for a methodological assessment of the accuracy of such panels (Blom, Ackermann-Piek et al., 2017). Each panel vendor fielded a survey including the same items as those asked in the GIP and GESIS panels. The only technical criteria communicated to the vendors was to recruit a sample of ~1,000 respondents representative of the general population of Germany aged 18–70 years. The study team did not provide explicit instructions on how to achieve representativeness, leaving the task to the individual vendors. One vendor joined the project for free upon learning about the methodological aims of the study. Information on the number of respondents, quota variables, and fieldwork periods for each of the eight nonprobability surveys, and the two probability surveys are provided in Table 2.

Weights were created for the GIP and GESIS Panel surveys and the eight nonprobability surveys based on the standard raking procedure used in the GIP. The raking weights were based on the following benchmark variables taken from the German micro-census: marital status, household size, age, and education.

Table 2. List of German probability and nonprobability surveys.

Survey	No. respondents	Quota variables	Fieldwork period
GIP	3426	N/A	1st–31st March 2015
GESIS Panel	3822	N/A	18th February–14th April 2015
NP Panel 1	1012	Age, gender, region, education	1st–31st March 2015
NP Panel 2	1000	Age, gender, region	5th–18th March 2015
NP Panel 3	999	Age, gender, region	2nd–11th March 2015
NP Panel 4	1000	Age, region	1st–18th March 2015
NP Panel 5	994	Age, gender, region	2nd–16th March 2015
NP Panel 6	1002	Age, gender, region, education	25th March–1st April 2015
NP Panel 7	1000	Age, gender, region	3rd–9th March 2015
NP Panel 8	1038	Age, gender, region	5th–11th March 2015

Accounting for selection into web

Because our focus is on online respondents only, we apply a further adjustment to the original survey weights to account for selection into the web mode for the probability-based mixed-mode surveys, that is, the GESIS Panel, the ABS survey, and the ANU Poll survey. A propensity-score adjustment method (Rosenbaum & Rubin, 1983) was implemented by modeling the likelihood of respondents answering in the web mode versus the non-web mode(s).

The logistic regression model for mode of response for the Australian data included as covariates: age, sex, education, employment status, citizenship, internet usage, number of surveys completed in the past 4 weeks, mobile phone usage, general health status, household status, moving in past 5 years, and home ownership. For the German data, the following covariates were included: age, sex, education, employment status, citizenship, internet usage, general health status, household size, marital status, and home ownership. The fitted model was used to estimate the probability of answering in the web mode for each respondent. These probabilities were sorted from lowest to highest and quintiles were formed. The average propensity score in each quintile was then calculated and the inverse of this average was used to produce the adjustment factor. This factor was then multiplied with the original survey weight to produce the overall weight used in the forthcoming analyses.

Measures

We analyze all multi-item scales measured in the German and Australian surveys. In both countries, this included three multi-item scales. In Australia, the items included questions on New technology, Internet use, and Psychological distress (Kessler et al., 2002). The New technology scale dealt with how willing respondents were to adapt new brands and technologies, the Internet use scale items dealt with how often various Internet activities were performed, and the Psychological distress scale (Kessler et al., 2002) asked respondents to report how often they felt certain negative feelings. In Germany, the items included a short version of the Big Five personality scale measuring two dimensions: agreeableness and openness (Digman, 1990; Goldberg, 1993; Rammstedt et al., 2013). A second scale included items regarding interest in politics and political activity. Finally, two dimensions from the Need for Cognition scale were measured: cognitive persistence and cognitive complexity (Beissert et al., 2014; Cacioppo & Petty, 1982; Tanaka et al., 1988). The full wording of items and response categories can be found in Table A7 of the Online Supplement.

Method

To understand whether the measurement structure of multi-item scales differs between probability and nonprobability panels we apply confirmatory factor analysis (CFA; Bollen, 1989). This approach estimates a latent variable based on related observed variables. The statistical model is defined as

$$y_t = \tau_t + \lambda_t T + \varepsilon_t$$

where the observed variables, y_t, are explained by an unobserved latent variable T. The relationship depends on a slope/loading parameter λ_t, an intercept (or conditional mean) τ_t, and a residual term ε_t. The loading can be considered an indicator of the strength of the relationship between the observed variable and the latent variable—the larger the value the more closely the observed variable of interest measures the unobserved latent variable. The intercept can be interpreted as the conditional mean, or the expected value of y, when the latent variable is 0 (which typically refers to the average of T). The residual represents the unexplained variance conditional on the latent variable and is an indicator of measurement error.

The model can also be visualized as shown in Figure 1. Here the latent variable is represented by a circle and the observed variables are represented by squares. All the coefficients discussed before are represented with each observed score, y, being explained by the true score T with a slope of λ_t, an intercept of τ_t, and a residual of ε_t. In the figure the configural model is presented where the coefficients are allowed to be different across the two groups (probability and nonprobability data).

We use a simple CFA model (as seen in Figure 1) for all the scales analyzed using the Australian data. In Germany, due to the limited number of items that were part of a scale, some restrictions were made to the models. For the Big Five and Need for Cognition scales, even if they were comprised of four items each, they measure different sub-dimensions of the concept of interest (based on substantive and statistical reasoning). As a result, we estimate two latent variables with two indicators each for each scale. To estimate the models, loadings are fixed to 1 for all the indicators and the correlations of the two latent variables are fixed to 0 (they were very close to 0 when freely estimated). Similarly, for interest in politics, only two indicators were observed and, as such, the loadings had to be restricted to 1.

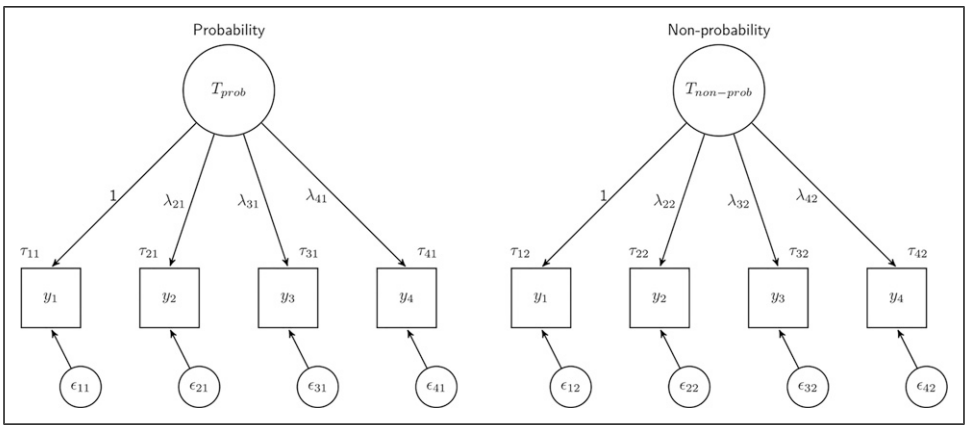

Figure 1. Visual representation of measurement model tested across groups.

The CFA estimation method has a number of advantages, but one of the most important ones is that it can be used to formally test for measurement equivalence between groups (Baumgartner & Steenkamp, 2006; Meredith, 1993). For example, multi-group CFA is often used for comparisons of attitudes across countries (Davidov et al., 2014) or between different survey modes (Cernat & Revilla, 2020). This is important for both substantive and methodological reasons. First, from a substantive point of view, the measurement of interest is typically assumed to be equivalent or invariant across groups. If this assumption does not hold, then one cannot meaningfully compare concepts across groups. This is also relevant in the present context, as attitudinal measurements and other social phenomena are commonly collected and compared between probability and non-probability surveys and within each type. For example, in political polling and election forecasting, estimates obtained from probability and nonprobability surveys are often compared (Sohlberg et al., 2017; Sturgis et al., 2018). The second reason is methodological. Comparing the factor models (also called measurement models) across groups can provide insights regarding the quality of the measures. For example, this might identify translation issues or problematic items in cross-country surveys. In the present study, this approach is used to understand whether probability and non-probability surveys differ in terms of their measurement structures for estimating latent variables.

Therefore, the focus of the present study is on multi-group analysis (also known as equivalence or invariance testing), where the measures presented above are compared across different groups. More precisely, we focus on three groups of comparisons. The first comparison is between the probability surveys (two in each country) and nonprobability surveys (five in Australia and eight in Germany) to assess whether the measurement equivalence can be established between these two sampling streams (RQ1). The second comparison is between the different nonprobability surveys to each other, five surveys in Australia and eight in Germany (RQ2). All comparisons are conducted separately within country and topic, and with and without the adjustment weights (RQ3) to assess whether selection plays a role in affecting the measurement comparisons. Typically, multi-group analysis involves comparing a series of nested models. Here, we compare five models that become cumulatively restrictive (each model includes the restrictions of the previous models):

1. *Configural model:* the factor structure of the measurement model is the same across groups, but all coefficients are allowed to be different across groups.
2. *Loadings (metric) model*: the loadings are restricted to be equal across groups.[1]
3. *Intercepts (scalar) model*: the intercepts are restricted to be equal across groups.
4. *Means (of the latent variable)*: the mean of the latent variable is restricted to be equal across groups.
5. *Residuals:* the variances of the residuals are restricted to be equal across groups.

Each of these restrictions provides insights about the comparability of measurement across groups. The loadings and the residuals refer to variance and, as such, can be viewed as proxies of reliability. The intercepts and latent means refer to systematic differences that affect the averages. The models are also important as they indicate what can be appropriately compared across groups. If the best fitting model is the loadings model, then only covariances can be compared across groups but no other comparisons can be made. If the intercept model is the best fitting model, then the means of the latent variables can be compared across groups as well. If the means model is the best fitting model, this indicates that the means of the latent variables are the same across groups. If the residuals model is the best model, then the observed summative scores can be compared across groups. Because we compare data that should refer to the same population, using the same items, and the same mode (i.e., web), we expect that the full measurement structure will be the same across

groups (i.e., model 5 will be the best fitting one). If that is not the case, then there are differences in the measurement structure that are not accounted for.

There are multiple ways of assessing model fit and selecting the best fitting model. Given the large number of models analyzed here we concentrate on only one indicator that is often used in this context, the difference in the Comparative Fit Index (ΔCFI) between adjacent models. This metric has been found to perform well when used to investigate measurement equivalence and, unlike other metrics, it is not very sensitive to sample size. We adopt the commonly used threshold of a CFI difference of 0.01 or more to indicate a significant decrease in the fit of a given model (Chen, 2007). For the reader's information, we also report other commonly used model fit indicators (Chi^2, *p*-value, RMSEA, and BIC). All outcome variables are treated as continuous and listwise deletion is used for the small number of missing cases (below 5%). The models were run in R 3.6.2 (R Core Team, 2019).

Results

Before testing the equivalence of the measurements across groups we assess the fit of the scales used. Table 3 shows the main fit indicators for the six scales by country. The three scales in Australia show moderate-to-good fit with CFI values of 0.99, 0.91 and 0.97, and RMSEA values of 0.09, 0.18, and 0.12 for the New technology, Internet use, and Psychological distress scales, respectively. The models in Germany display moderate fit with the Big Five scale having a CFI value of 0.8 and an RMSEA value of 0.07 while the Need for Cognition has a CFI value of 0.98 and an RMSEA of 0.06. The political interest scale is just-identified (it has 0 degrees of freedom) and as such has no fit indicators.

Comparing probability and nonprobability surveys

To address the first research question (RQ1), we test for measurement equivalence between all probability data in one group and all nonprobability data in the second group for the six scales in the two countries. Looking at the three scales in Australia, we see that there are few significant differences in measurement, as the difference in the Comparative Fit Index (ΔCFI) between adjacent models does not exceed 0.01 (Table 4). For the new technology and psychological distress scales, the measurement seems to be equivalent across the probability and nonprobability surveys. However, for the Internet use scale, the intercepts, and the residuals differ between groups. This would indicate that there are systematic differences in the answering patterns of probability and nonprobability panels that have an impact on the observed averages for Internet use. This would also

Table 3. Fit indices for measurement models in the pooled data.

Country	Dependent	Chi^2	Df	*p*-value	CFI	RMSEA	BIC
Australia	New technology	146.5	5	0.00	0.99	0.09	31,153
	Internet use	220.1	2	0.00	0.91	0.18	58,226
	Psychological distress	477.4	9	0.00	0.97	0.12	47,948
Germany	Big Five	263.5	4	0.00	0.80	0.07	158,148
	Need for Cognition	175.0	4	0.00	0.98	0.06	197,329
	Politics[a]	—	—	—	—	—	—

[a] The indices for the politics scale are not shown as it is just-identified (no fit indicators).

Table 4. Equivalence testing of probability versus nonprobability data in Australia, unweighted. Differences in ΔCFI larger than 0.01 are bolded.

Dependent	Model	Chi^2	df	*p*-value	CFI	RMSEA	BIC
New technology	Configural	175.1	10	0.00	0.984	0.097	31,075
	Loadings	185.6	14	0.00	0.984	0.083	31,053
	Intercepts	256.7	18	0.00	0.977	0.087	31,092
	Means	344.0	19	0.00	0.969	0.098	31,171
	Residuals	415.9	24	0.00	0.962	0.096	31,202
Internet use	Configural	233.9	4	0.00	0.908	0.180	58,233
	Loadings	236.9	7	0.00	0.908	0.136	58,211
	Intercepts	311.6	10	0.00	**0.880**	0.130	58,262
	Means	319.9	11	0.00	0.877	0.126	58,262
	Residuals	351.9	15	0.00	**0.866**	0.113	58,261
Psychological distress	Configural	488.8	18	0.00	0.969	0.121	47,929
	Loadings	510.3	23	0.00	0.968	0.109	47,910
	Intercepts	522.2	28	0.00	0.967	0.100	47,881
	Means	624.8	29	0.00	0.960	0.108	47,975
	Residuals	667.8	35	0.00	0.958	0.101	47,969

imply that comparing means across the two types of samples could lead to bias. The differences in the residuals imply the presence of differences in reliability across groups.

We next look at the coefficients within the Internet use scale that were indicated as significantly different by the ΔCFI. In the nonprobability sample, respondents seem to systematically underreport how often they look for information over the Internet (a4a) compared to the probability sample (the intercepts are 0.61 for nonprobability vs. 0.22 for probability). The opposite is true for using the Internet to post on blogs/forums/interest groups (the intercepts are 1.92 for the nonprobability sample and 2.37 for the probability sample). Mixed results are also found when looking at differences in the residuals with lower estimates (and thus more reliable data) for the nonprobability panel for using the Internet to post images to social media (a4b) and to conduct financial transactions (a4c), and higher estimates (and thus less reliable data) for looking for information over the Internet (a4a) and using the Internet to post on blogs and forums (a4d).

In Germany, one type of coefficient significantly differs between probability and nonprobability groups in each of the three scales (Table 5). For the Big Five, the intercepts are significantly different, while for Need for Cognition and Politics, the means are different. All indicators refer to systematic differences (i.e., differences in the means) that may occur due to differing response styles, such as acquiescence, or due to selection.

Looking at the model coefficients where measurement equivalence was not found, we find mixed results. For the Big Five scale there are no consistent differences in the intercepts across groups as two of them are higher in the probability data while two are lower compared to the nonprobability data. For the second scale, Need for Cognition, it appears that nonprobability surveys systematically underestimate the mean of the latent variable (2.84 vs. 3.68 for the cognitive persistence factor and 2.55 vs. 3.45 for the cognitive complexity factor) compared to the probability sample surveys. These differences could be the result of higher propensities to provide socially desirable answers in the probability panels (e.g., respondents tend to say they like to be intellectually challenged) or due to higher acquiescence (tendency to agree more). For the politics scale, however, we find that the average of the factor is higher in the nonprobability data (4.14 vs. 3.88), possibly indicating higher social desirability (e.g., more likely to say that they are active in or interested in politics).

Table 5. Equivalence testing of probability versus nonprobability data in Germany, unweighted. To estimate the models the loadings are fixed to 1 and cannot be compared across groups. Differences in ΔCFI larger than 0.01 are bolded.

Dependent	Model	Chi^2	df	*p*-value	CFI	RMSEA	BIC
Big Five	Configural	301.6	8	0.00	0.781	0.074	158,181
	Intercepts	370.6	10	0.00	**0.731**	0.073	158,231
	Means	382.4	12	0.00	0.724	0.067	158,224
	Residuals	400.4	16	0.00	0.714	0.060	158,204
Need for Cognition	Configural	198.5	8	0.00	0.974	0.059	197,131
	Intercepts	203.2	10	0.00	0.973	0.053	197,117
	Means	448.0	12	0.00	**0.940**	0.073	197,343
	Residuals	507.2	16	0.00	0.932	0.067	197,364
Politics	Configural	0.0	0		1.000	0.000	108,495
	Intercepts	3.7	1	0.06	0.999	0.020	108,489
	Means	82.8	2	0.00	**0.983**	0.077	108,559
	Residuals	85.1	4	0.00	0.983	0.055	108,542

In summary, of the six scales compared in this study, four did not achieve full measurement equivalence between probability and nonprobability panels. Only one of these scales (i.e., Internet use) is from the Australian data, where the loadings model is the best fitting one, meaning that covariances can be compared across groups, but no other comparisons can be made. For the German data, the lowest level of equivalence was found for the Big Five scale, where the configural model is the best fitting model, meaning that the factor structure holds across groups, but no other comparisons can be made. For the other scales (Need for Cognition and Politics), the intercepts models fit best, implying that the means of the latent variables can be compared in addition to covariances. Our results show that scales differ in terms of equivalence between probability and nonprobability data. Some scales reach full equivalence while others do not, limiting their use in social research.

Comparing nonprobability surveys

Having identified some measurement differences between probability and nonprobability surveys, we turn to assessing whether measurement equivalence can be established between the nonprobability surveys (RQ2). Comparing the different providers of nonprobability surveys can show whether the measurement structure of latent variables differ between nonprobability panel vendors and hint at the potential impact of methodological choices made by different vendors.

In Australia, as with the previous comparison of probability and nonprobability panels, only the Internet use scale is significantly different across the five nonprobability panels (Table 6). As before, both the intercepts and the residuals are different across groups. The remaining scales, New technology and Psychological distress, achieve full measurement equivalence. When examining the intercepts and residuals for Internet use, we find that the coefficients range widely in the different nonprobability panels although no systematic differences stand out. The largest differences in the intercepts across the panels refers to looking for information over the Internet (a4a) and using the Internet to post on blogs and forums (a4d) with the values ranging between 0.31 and 0.85 for the former and 1.70 and 2.20 for the latter. Similarly, the residuals vary across the panels but there seem to be no systematic patterns where one panel outperforms the others.

In comparing the eight nonprobability panels in Germany we found significant differences in each of the three scales (Table 7). For the Big Five scale both the intercepts and the means are significantly

Table 6. Equivalence testing of nonprobability data in Australia, unweighted. Differences in ΔCFI larger than 0.01 are bolded.

Dependent	Model	Chi^2	df	*p*-value	CFI	RMSEA	BIC
New technology	Configural	171.3	25	0.00	0.984	0.098	26,638
	Loadings	192.0	41	0.00	0.984	0.078	26,530
	Intercepts	218.3	57	0.00	0.983	0.068	26,428
	Means	239.2	61	0.00	0.981	0.069	26,417
	Residuals	271.7	81	0.00	0.980	0.062	26,289
Internet use	Configural	194.5	10	0.00	0.918	0.174	50,261
	Loadings	227.2	22	0.00	0.909	0.124	50,197
	Intercepts	340.0	34	0.00	**0.864**	0.122	50,214
	Means	348.8	38	0.00	0.862	0.116	50,190
	Residuals	403.5	54	0.00	**0.845**	0.103	50,117
Psychological distress	Configural	538.1	45	0.00	0.964	0.134	42,097
	Loadings	563.6	65	0.00	0.963	0.112	41,962
	Intercepts	661.6	85	0.00	0.958	0.105	41,900
	Means	667.7	89	0.00	0.958	0.103	41,874
	Residuals	737.8	113	0.00	0.954	0.095	41,751

different, while for the Need for Cognition and the Politics scales, the means are different across groups. Looking at the coefficients that are different across groups, we do not see any systematic patterns. For the Big Five scale, for example, we see that differences in the intercepts have a range of approximately 0.08 while for the means of the factors, it is 0.1. For the Need for Cognition scale, the Cognitive persistence factor shows more variation in the mean (range of around 0.6) than the Cognitive complexity factor (range around 0.2). For the mean of the Politics scale the range between different nonprobability panels is 0.2.

In a separate analysis, we also compared measurement models across the two probability panels in Germany and Australia. These results can be found in Tables A1 and A2 of the Online Supplement. The German probability panels differed in Politics (intercepts, means and residuals), Big Five (intercepts), and Need for Cognition (residuals), but other differences were not larger than the ΔCFI threshold of 0.01. In Australia, differences in ΔCFI larger than 0.01 were only found for Internet use (intercepts, means, and residuals) when comparing probability panels. The New technology and Psychological distress scales were fully equivalent across the probability surveys. Thus, in both countries it is apparent that measurement differences (especially of the systematic type) can arise not only between nonprobability panels but also between probability panels.

Do weights correct for differences in measurement?

The final research question (RQ3) is addressed by assessing whether the measurement differences are influenced by differential selection. Differential selection has been highlighted in prior research when comparing probability and nonprobability surveys and nonprobability panel vendors (Dutwin & Buskirk, 2017; Macinnis et al., 2018; Sturgis et al., 2018). To investigate this, we re-ran all previous models with weights that control for selection based on demographic characteristics found in official statistics (see the Method section for a description of the weighting approach for each survey). Next, we compared the model results from the unweighted analyses (previous tables) with those from the weighted models (not shown).

To facilitate the interpretation of all models and comparisons, Table 8 summarizes the differences between the groups with and without weights. While the weights do have an impact for half of the

Table 7. Equivalence testing of nonprobability data in Germany, unweighted. To estimate the models the loadings are fixed to 1 and cannot be compared across groups. Differences in ΔCFI larger than 0.01 are bolded.

Dependent	Model	Chi^2	df	*p*-value	CFI	RMSEA	BIC
Big Five	Configural	272.4	32	0.00	0.737	0.087	93,671
	Intercepts	298.1	46	0.00	**0.724**	0.074	93,571
	Means	330.0	60	0.00	**0.704**	0.067	93,477
	Residuals	364.4	88	0.00	0.697	0.056	93,260
Need for Cognition	Configural	155.0	32	0.00	0.974	0.062	116,279
	Intercepts	172.9	46	0.00	0.973	0.052	116,171
	Means	362.4	60	0.00	**0.936**	0.071	116,234
	Residuals	405.6	88	0.00	0.933	0.060	116,026
Politics	Configural	0.00	0		1.000	0.000	64,442
	Intercepts	17.5	7	0.01	0.996	0.039	64,396
	Means	66.3	14	0.00	**0.981**	0.061	64,382
	Residuals	98.4	28	0.00	0.975	0.050	64,289

Table 8. Types of coefficients different by group based on a ΔCFI of 0.01. Bold denotes coefficients that are made similar across groups by weighting while italics denotes those that are made different by weighting.

Comparison	Country	Scale	No weights	With weights
Probability versus nonprobability	Australia	New technology		*Residuals*
		Internet use	Intercepts, **residuals**	Intercepts
		Psychological distress		
	Germany	Big Five	Intercepts, **residuals**	Intercepts, *means*
		Cognition	Means	Means
		Politics	Means	Means
Within nonprobability	Australia	New technology		
		Internet use	Intercepts, residuals	*Loadings*, intercepts, residuals
		Psychological distress		
	Germany	Big Five	Intercepts, means	Intercepts, means, *residuals*
		Cognition	Means	Means
		Politics	**Means**	*Residuals*
Within probability	Australia	New technology		*Loadings*
		Internet use	Intercepts, **means**, **residuals**	*Loadings*, intercepts
		Psychological distress		*Residuals*
	Germany	Big Five	Intercepts	Intercepts, *mean*
		Cognition	Residuals	Residuals
		Politics	Intercepts, **means**	Intercepts

scales (including six coefficients made more similar, nine made more different, and twenty-six unaffected), they have different effects on different scales. For example, when comparing probability and nonprobability surveys the weights make the residuals of the Internet use and Big Five scales similar across groups but make the residuals for the New technology scale and the means for

the Big Five different. The weights have no effect on the Psychological distress, Cognition, and Politics scales.

When comparing the different nonprobability panels, the weights shift the means of the Politics scale to be the same across groups while causing the loadings for Internet use and the residuals for the Big Five and Politics to be different. The weights have no effect on measurement equivalence for the New technology and Psychological distress scales. We also find mixed results for the probability online panels where three types of coefficients become similar across panels with the weights, but four other types of coefficients become more divergent. Only the cognition scale is unaffected by the weights.

In short, weighting to account for differential selection does affect several coefficients, but in mixed and unpredictable ways. Some coefficients are made more similar, others more divergent, while others are unaffected by weighting. No discernable pattern is evident, and effects are found with each type of comparison (between probability and nonprobability and within each type). Weighting by demographics can improve measurement equivalence in certain cases, but the effects are heterogeneous and insufficient as a general remedy for ensuring measurement equivalence.

Discussion

This article investigated measurement equivalence between probability and nonprobability panel surveys for several multi-item scales measured in Germany and Australia. Confirmatory factor analysis was used to formally test measurement equivalence between (and within) these two sampling types. Failing to achieve measurement equivalence can be problematic in this context as it means that multi-item scale data cannot be meaningfully combined or compared across the different sample types. Such differences can be indicative of differences in the response behavior or response styles of the respondents.

When comparing probability and nonprobability panels, we found differences in five (out of 27) coefficients tested across the six scales. In Australia, differences were found for two coefficients related to Internet use, while one type of coefficient differed in each of the three scales used in the German surveys. Interestingly, most of the coefficients that were different were related to the average (mean of the factor and intercept). This may be an indication of systematic (related to the mean) differences between the groups. These results suggest that differences in measurement can (but do not always) arise when comparing data from probability and nonprobability samples. Therefore, researchers comparing or combining data from the different types of samples should be wary of drawing inferences before measurement equivalence has been established.

By examining the coefficients that were different we found mixed patterns. There were some signs of systematic differences that could be due to social desirability or acquiescence in some of the means of the latent factors but, again, the pattern was mixed. Although we found differences between probability and nonprobability panels there were no consistent differences to indicate that respondents in nonprobability samples produce more measurement error compared to those in the probability samples and vice versa.

In addition to exploring differences between probability and nonprobability panels, we examined whether measurement equivalence could be established between different nonprobability panels in the same country. The results were similar, as six (out of 27) coefficients were different, and these were the same coefficients that differed in the prior analysis (with the addition of one more difference in Germany). Therefore, comparisons of nonprobability panels face the same types of problems as comparisons between probability and nonprobability panels. Although not a research question, we also tested for measurement equivalence between the probability panels. Those results

(presented in the Online Supplement) showed that measurement equivalence is not established for four of the six scales analyzed: Internet use in Australia (intercepts, means, and residuals), and for all scales in Germany (Big Five (intercepts), Need for Cognition (residuals), and Politics (intercepts, means, and residuals)). This is also not ideal, as probability surveys are typically assumed to be the gold standard for measurement and are commonly compared, but they too can sometimes produce non-equivalent measurement results.

Applying weights to adjust for differential selection by demographics did not improve comparability across groups, as some scales became more similar in their coefficients, while others became more divergent. Still, there were some interesting trends in the effects of the weights, as the coefficients that diverged tended to be related to variance (loadings and residuals) while the coefficients that became more similar tended to be related to the average of the latent variable (means and intercepts). This may suggest that compositional differences affect the means, but weighting adds noise to the measurements by inflating the variance. When comparing the probability panels to the nonprobability ones, the weights improved the differences in residuals. These findings show that correcting for selection based on official statistics variables is insufficient to ensure measurement equivalence between different data sources; see also Hox et al. (2015). These results highlight the fact that more work remains in understanding under which circumstances valid comparisons can be made between measurements collected from different sample types.

Also evident from the comparison is that not all scales are affected by the sampling design. For example, all coefficients for the Psychological distress and New technology scales were equivalent, while issues arose with the Internet use scale in every comparison made (probability vs nonprobability, and comparisons within each type). This indicates that measurement equivalence might be more robust for some topics than others.

As with all studies this one has limitations. Our analyses cannot establish whether one sampling approach (probability or nonprobability) produces more accurate measurements than the other, only that they do not measure all scales in the same way. Furthermore, the study was restricted to two countries with a limited number of surveys (Australia: N = 7, Germany: N = 10), which did not use the same scales, and was not designed to assess country-specific effects on measurement equivalence. Future research should focus on testing the effects of the same scales in probability and nonprobability panels in multiple countries and across more survey questions. The samples were also recruited using different survey designs, which may affect response behavior in different ways. This includes reaching the offline population in Germany which may respond in different ways due to their lack of experience interacting with questionnaire forms compared to those who were web users before being recruited to the probability surveys. Finally, we accounted for selection by using weighting adjustments based on demographics to control for compositional differences between the samples; however, there is always the risk that weights do not completely explain the selection mechanism. Future research could examine the effects of weighting on variables other than demographic characteristics, or on different weighting adjustment approaches than the logistic regression procedure that was used here.

This study has found differences in measurement for probability and nonprobability panel surveys, and within each sampling type. The differences between the two sample types were not larger than those found within each type. Full measurement equivalence can therefore not be guaranteed in any of the comparisons, although many coefficients were equivalent. The practical implications of these findings are that before combining or comparing data from different types of samples, in addition to accounting for differences in selection, measurement equivalence should be tested and established for multi-item scales. This could be done using multi-group CFA, as we have done. If differences in the measurement structure of latent variables are identified using CFA, then

users can attempt to establish partial equivalence by relaxing the parameter restrictions for the problematic questions (Byrne et al., 1989), given that all parameter restrictions hold for at least two of the other questions measuring the latent construct. This approach can also be used to correct for differences in measurement across samples. By using the latent variables, as opposed to the observed variables or sum scores, researchers will then be able to validly combine or compare the concepts between probability and nonprobability samples.

In conclusion, we find that measurement equivalence is not ensured when combining or comparing data from different probability and nonprobability panels. Differences in measurement equivalence were found between probability and nonprobability panels and within each type and demographic weighting adjustments did not consistently resolve these discrepancies. While this study has shown the importance of testing for measurement equivalence when using survey data from different sources, it is only a first step. Understanding the specific mechanisms that lead to nonequivalence is still an open question and an important topic for future research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by German Research Foundation (139943784 (SFB 884)) and German Federal Ministry of Education and Research.

ORCID iDs

Hafsteinn Einarsson https://orcid.org/0000-0001-9623-487X
Alexandru Cernat https://orcid.org/0000-0003-2176-1215
Carina Cornesse https://orcid.org/0000-0002-5437-2999
Annelies G Blom https://orcid.org/0000-0003-0377-301X

Supplemental Material

Supplemental material for this article is available online.

Note

1. In Germany, this type of equivalence was not investigated as all the loadings were restricted to 1 for estimation reasons.

References

AAPOR. (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*(3), 285–303. https://doi.org/10.1093/pan/mpt025

Baumgartner, H., & Steenkamp, J. B. E. M., (2006). "An extended paradigm for measurement analysis of marketing constructs applicable to panel data". *Journal of Marketing Research*, *43*(3), 431–442. https://doi.org/10.1509/jmkr.43.3.431

Beissert, H., Köhler, M., Rempel, M., & Beierlein, C. (2014). *Eine deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition: Die Need for Cognition Kurzskala (NFC-K), GESIS-Working Papers*. Mannheim: GESIS-Working Papers.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., & Weimer, D. L. (2003). "The advent of internet surveys for political research: A comparison of telephone and internet samples". *Political Analysis*, *11*(1), 1–22. https://doi.org/10.1093/pan/11.1.1

Blom, A. G., Ackermann-Piek, D., Helmschrott, S. C., Cornesse, C., & Sakshaug, J. W. (2017a). "The representativeness of online panels: Coverage, sampling and weighting". In General online research conference, Berlin, Germany, 11–13 March 2020.

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A. S., Das, M., Douhou, S., & Krieger, U. (2016). "A comparison of four probability-based online and mixed-mode panels in Europe". *Social Science Computer Review*, *34*(1), 8–25. https://doi.org/10.1177/0894439315574825

Blom, A. G., Gathmann, C. & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, *27*(4), 391-408. https://doi.org/10.1177/1525822X15574494.

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2017b). "Does the recruitment of offline households increase the sample representativeness of probability-based online panels? evidence from the german internet panel". *Social Science Computer Review*, *35*(4), 498–520. https://doi.org/10.1177/0894439316651584

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). "Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS panel". *Social Science Computer Review*, *36*(1), 103–115. https://doi.org/10.1177/0894439317697949

Brüggen, E., Van den Brakel, J., & Krosnick, J. A. (2016). *Establishing the accuracy of online panels for survey.* Statistics Netherlands.

Buck, S. F., Fairclough, E. H., Jephcott, J. S. G. & Ringer, D. W. C. (1977). "Conditioning and bias in consumer panels - some new results". *Journal of the Market Research Society*, *39*(1), 1-11. https://doi.org/10.1177/147078539703900102.

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin*, *105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456.

Cacioppo, J. T., & Petty, R. E. (1982). "The need for cognition". *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Callegaro, M., Baker, R., Göritz, A. S., Krosnick, J. A. & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look at the future. In M. Callegaro, R. Baker, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 1-22). John Wiley & Sons.

Callegaro, M., Villar, A., Yeager, D. & Krosnick, J. A. (2014). A critical review of studies investigating the quality of data obtained with online panels. In M. Callegaro, R. Baker, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 23-53). John Wiley & Sons. https://doi.org/10.1002/9781118763520.ch2.

Cernat, A., & Revilla, M. (2020). "Moving from face-to-face to a web panel: Impacts on measurement quality". *Journal of Survey Statistics and Methodology*, *9*(4), 1–19. https://doi.org/10.1093/jssam/smaa007

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*(4), 641–678. https://doi.org/10.1093/poq/nfp075

Chen, F. F. (2007). "Sensitivity of goodness of fit indexes to lack of measurement invariance". *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 465–504. https://doi.org/10.1080/10705510701301834

Cornesse, C., & Blom, A. G. (2020). "Response quality in nonprobability and probability-based online panels". *Sociological Methods and Research*. https://doi.org/10.1177/0049124120914940

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., Leeuw De, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). "A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research". *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Cornesse, C., & Schaurer, I. (2021). "The long-term impact of different offline population inclusion strategies in probability-based online panels: Evidence from the German Internet Panel and the GESIS Panel". *Social Science Computer Review*, *39*(4), 30–32. https://doi.org/10.1177/0894439320984131

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P. & Billiet, J. (2014). "Measurement equivalence in cross-national research,". *Annual Review of Sociology*, *40*(1), 55-75. https://doi.org/10.1146/annurev-soc-071913-043137.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417-440. https://doi.org/10.1146/annurev.ps.41.020190.002221.

Dutwin, D., & Buskirk, T. D. (2017). "Apples to Oranges or Gala versus golden delicious?" *Public Opinion Quarterly*, *81*(S1), 213–249. https://doi.org/10.1093/poq/nfw061

Goldberg, L. R. (1993). "The structure of phenotypic personality traits". *American Psychologist*. https://doi.org/10.1037/0003-066X.48.1.26

Greszki, R., Meyer, M., & Schoen, H. (2014). The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels. In M. Callegaro, R. Baker, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 238–262). John Wiley & Sons.

Hillygus, D. S., Jackson, N., & Young, M. (2014). "Professional respondents in nonprobability online panels". In *Online panel research: A data quality perspective* (pp. 219–237). Wiley. https://doi.org/10.1002/9781118763520.ch10

Hitchman, S. C., Brose, L. S., Brown, J., Robson, D., & McNeill, A. (2015). Associations between E-Cigarette type, frequency of use, and quitting smoking: Findings from a longitudinal online panel survey in Great Britain. *Nicotine and Tobacco Research*, *17*(10), 1187–1194. https://doi.org/10.1093/ntr/ntv078

Hox, J. J., de Leeuw, E. D., & Zijlmans, E. A. O. (2015). "Measurement equivalence in mixed mode surveys". *Frontiers in Psychology*, *6*(770), 1–11. https://doi.org/10.3389/fpsyg.2015.00087

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., Mcgeeney, K., & Gimenez, A. (2016). *Evaluating online nonprobability surveys*. Pew Research Center.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L. T., Walters, E. E., & Zaslavsky, A. M. (2002). "Short screening scales to monitor population prevalences and trends in non-specific psychological distress". *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Keusch, F., Batinic, B., & Meyerhofer, W (2014). Motives for joining nonprobability online panels and their association with survey participation behavior. In M. Callegaro, R. Baker, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 171–191). John Wiley & Sons.

Kish, L. (1965). *Survey sampling*. Wiley.

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2020). "Social media, web, and panel surveys: Using non-probability samples in social and policy research". *Policy and Internet*, *13*(1), 1–22. https://doi.org/10.1002/poi3.238

Lohr, S. L. (2019). *Sampling: Design and analysis*. Chapman and Hall/CRC.

Loosveldt, G., & Sonck, N. (2008). "An evaluation of the weighting procedures for an online access panel survey". *Survey Research Methods*, *2*(2), 93–105. https://doi.org/10.18148/srm/2008.v2i2.82

Macinnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. J. (2018). "The accuracy of measurements with probability and nonprobability survey samples: Replication and extension". *Public Opinion Quarterly*, *82*(4), 707–744. https://doi.org/10.1093/poq/nfy038

Meredith, W. (1993). "Measurement invariance, factor analysis and factorial invariance". *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Pasek, J. (2016). "When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence". *International Journal of Public Opinion Research*, *28*(2), 269–291. https://doi.org/10.1093/ijpor/edv016

Pennay, D., Borg, K., Neiger, D., Misson, S., Honey, N., & Lavrakas, P. (2016). *Online panels benchmarking study (Technical Report)*. The Social Research Centre.

Powell, J., Inglis, N., Ronnie, J. & Large, S. (2011). The characteristics and motivations of online health information seekers: Cross-sectional survey and qualitative interview study. *Journal of Medical Internet Research*, *13*(1), 1-11. https://doi.org/10.2196/jmir.1600.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rammstedt, B., Kemper, C., Klein, M., Beierlein, C. & Kovaleva, A. (2013). Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: 10 Item Big Five Inventory (BFI-10). *Methoden, Daten, Analysen (mda)*, *7*(2), 233-249. https://doi.org/10.12758/mda.2013.013.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. In *Biometrika* (Vol. 70, pp. 41–55). Cambridge University Press. https://doi.org/10.1093/biomet/70.1.41

Scherpenzeel, A. C., & Bethlehem, J. G (2011). How Representative Are Online Panels? In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 105–132). Routledge.

Skitka, L. J., & Sargis, E. G. (2006). "The Internet as psychological laboratory". *Annual Review of Psychology*, *57*(1), 529–555. https://doi.org/10.1146/annurev.psych.57.102904.190048

Sohlberg, J., Gilljam, M., & Martinsson, J. (2017). "Determinants of polling accuracy : The effect of opt-in Internet surveys". *Journal of Elections, Public Opinion and Parties*, *27*(4), 433–447. https://doi.org/doi.org/10.1080/17457289.2017.1300588

Sparrow, N. (2006). "Developing reliable online polls". *International Journal of Market Research*, *48*(6), 659–680. https://doi.org/10.1177/147078530604800604

Struminskaya, B., & Bosnjak, M (2021). "Panel conditioning: Types, causes, and empirical evidence of what we know so far". In P. Lynn (Ed.), *Advances in Longitudinal survey methodology* (pp. 272–301). Hoboken. https://doi.org/10.1002/9781119376965.ch12

Sturgis, P., Kuha, J., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Lauderdale, B. E., & Smith, P. (2018). "An assessment of the causes of the errors in the 2015 UK general election opinion polls". *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *181*(3), 757–781. https://doi.org/10.1111/rssa.12329

Taichman, D. B. (2020). "Knowledge and perceptions of COVID-19 among the general public in the United States and the United Kingdom: A cross-sectional online survey". *Annals of Internal Medicine*, *172*(1), ED1. https://doi.org/10.7326/AWED202001070.

Tanaka, J. S., Panter, A. T., & Winborne, W. C. (1988). Dimensions of the need for cognition: subscales and gender differences. *Multivariate Behavioral Research*, *23*(1), 35–50. https://doi.org/10.1207/s15327906mbr2301_2

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples". *Public Opinion Quarterly*, *75*(4), 709–747. https://doi.org/10.1093/poq/nfr020

Zhang, C., Antoun, C., Yan, H. Y., & Conrad, F. G. (2019). "Professional respondents in opt-in online panels: What do we really know?" *Social Science Computer Review*, *38*(6), 1–17. https://doi.org/10.1177/0894439319845102.