Survey Research Methods (2020)
Vol. 14, No. 4, pp. 399-415
doi:10.18148/srm/2020.v14i4.7416

ISSN 1864-3361
http://www.surveymethods.org

Investigating the relative impact of different sources of measurement non-equivalence in comparative surveys: An illustration with scale format, data collection mode and cross-national variations

Caroline Roberts
Institute of Social Sciences, Faculty of Social and Political Sciences
University of Lausanne

Oriane Sarrasin
Institute of Psychology, Faculty of Social and Political Sciences
University of Lausanne

Michèle Ernst Stähli
FORS - Swiss Centre of Expertise in Social Sciences
University of Lausanne

Different factors are known to affect the comparability of multinational, multicultural and multiregional ('3MC') survey data. These include factors relevant to the design of the questionnaire in different contexts (such as cultural differences in how a concept is understood, inaccurate or approximate translations of concepts, and variant adaptations to question formats). Others include factors relating to the survey design in general and how it is implemented across contexts (such as sample design, choice of mode(s), and contact strategies). While research to date has looked at the effects of these factors on measurement invariance individually, there have been few attempts to compare them directly and assess their relative impact. To illustrate how this can be done, the present paper tests for measurement invariance in a subjective wellbeing measure across scale formats, modes of data collection, and linguistic and cultural contexts. To do so, European Social Survey data from Switzerland, Germany and France were combined, enabling analyses of the effects of naturally present and experimentally induced variations (resulting from the use of variant question formulations and translations, and tests of mode and question wording effects) on data comparability. Overall, we find variant translations and other cross-national variations to be bigger sources of nonequivalence than scale format and mode. The findings are of interest both to survey designers making decisions about optimal resource allocation in the design of 3MC studies, as well as to analysts comparing countries with shared languages and interpreting cross-group differences.

Keywords: Measurement invariance; mixed mode; European Social Survey; wellbeing

1 Introduction

Surveys are designed with the aim of measuring variation between population members so that researchers can draw conclusions about the ways and extent to which they differ. In order to make such comparisons meaningful, measurement accuracy is paramount, and depends largely on the design of the questionnaire—the words selected to formulate the question, the format in which respondents must give their answer, and the labels used to define the available response alternatives (Fowler & Consenza, 2008). In comparative surveys, where the goal is to make comparisons across multiple nations, cultures and regions (sometimes referred to as 3MC studies; Survey Research Center, 2016) and the questionnaire must be administered in multiple languages, these questionnaire design challenges are compounded with the inherent complexity of how best to translate concepts and the words used to express them to ensure valid measurement in each context of interest. Validity relies on the 'equality or equivalence' of measurement across contexts (Jowell, 1998, p. 169). Designing the survey in a way that achieves an acceptable degree of equivalence is, therefore, of the utmost importance for comparative research, and establishing the extent to which equivalent measurements are obtained is thus increasingly recognised as an essential requirement for analysing differences between groups (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014).

While questionnaire design and effective translation are key determinants of data quality in comparative surveys, they

Contact information: Caroline Roberts, Institute of Social Sciences, Géopolis-5534, CH-1015 Lausanne, Switzerland (E-mail: caroline.roberts@unil.ch)

are not the only ones. Design features such as the target population, the sampling strategy, the mode of data collection and fieldwork protocols (timing, contact schedules, non-contact rates) are all important influences. Often, despite best efforts, competing resource constraints and institutional differences result in unintended or unavoidable differences in the design and implementation of 3MC surveys, making threats to measurement comparability inevitable. In this context, it is important—both for survey designers, as well as for comparative analysts—to establish which design features pose the greatest risk to comparisons. While there is a substantial body of research investigating the effects of specific sources of nonequivalence on the comparability of 3MC data (see ibid. for a recent review), there is little research comparing them directly or considering their combined effect (though there have been attempts to control for some potential sources of nonequivalence while testing for another in particular; e.g., E. Hu et al., 2019).

In this article, we present an analysis of the relative impact of a number of key sources of nonequivalence in comparative surveys. Drawing on methodological experiments implemented in the third round of the European Social Survey (2006) and the comparative design features implicit in this survey, we test for measurement equivalence in a measure of subjective wellbeing in a series of cross-group comparisons implemented sequentially, to draw conclusions about the extent to which 1) variations in scale format, 2) different modes of data collection, and 3) cross-national differences (in how the survey is implemented and how questions were adapted in different languages) affect the comparability of measurement. Before describing in detail our methods and presenting the findings, we consider the significance of establishing the relative impact of different sources of measurement nonequivalence, focusing on how this process can serve both a diagnostic purpose, as well as shed light on substantive cross-group differences of interest. By doing so, our goal is to illustrate how the impact of potential sources of nonequivalence can be compared, and which solutions can be applied if the measurement is found not to be invariant.

2 Sources of measurement nonequivalence in comparative surveys—the Total Survey Error framework

Survey design decisions in comparative surveys, just as in single-country studies, are guided by the Total Survey Error (TSE) framework (Biemer, 2010; Groves et al., 2009), taking into consideration all (i.e., sampling and non-sampling) error sources likely to affect the accuracy of the estimates produced and the comparisons to be made (Pennell, Cibelli Hibben, Lyberg, Mohler, & Worku, 2017; Smith, 2011), and deciding accordingly, how best to allocate resources to address the potentially most damaging. Because equivalence is so essential to effective comparative research, recommendations for designing 3MC surveys stress the importance of 'stringent and well-policed ground rules for comparable survey methods' (e.g. Jowell, 1998, p. 175), whereby researchers responsible for running the survey in different contexts are required to apply the same design, or follow the same specifications, to ensure as far as possible that differences observed between contexts cannot be attributed to differences in how the survey was implemented. In practice, the use of identical methods is not always possible, and functionally equivalent methods may be necessary (Smith, 2011; e.g., different sample designs, Häder and Lynn, 2007). In other cases, non-equivalent methods may end up being used either by accident or intentionally, either because of 'country-specific differences in methodological or procedural habits' (Jowell, Kaase, Fitzgerald, & Eva, 2007, p. 7), or especially, resource constraints, which themselves vary by context, and may be exacerbated by demands relating to harmonised specifications for a study.

Given the multiple survey design features with potentially conflicting influences on measurement comparability, the job of designing high quality comparative surveys is extremely complex and costly. This implies a need for survey methodological evidence that sheds light on which sources of nonequivalence are most detrimental to comparisons, to ensure limited resources are directed where the impact will be greatest. The TSE framework (Biemer & Lyberg, 2003; Groves et al., 2009) facilitates such an evaluation, and has been extended to take account of the multiple additional threats to comparative survey quality not present in single-population studies (e.g. Pennell et al., 2017; Smith, 2011). In 3MC studies, each error component contributes to comparison error, a conceptual error that influences the comparability of the data collected across different populations (Smith, 2011). As well as guiding survey design decisions, the TSE framework also provides a basis for evaluating and adjusting error structures post hoc to ensure comparison error is kept to a minimum during data analysis. The onus is on the analyst to assess the extent of the threat from comparison error, by evaluating the equivalence of measurements across the populations to be compared and even where equivalence is established empirically, ideally, eliminating competing explanations (other than true variance) for any differences observed between populations (Davidov et al., 2014).

2.1 Levels of equivalence

The TSE framework applied to 3MC studies builds on earlier work in the field of comparative sociology and cross-cultural psychology, which distinguishes three hierarchically-related levels of equivalence (e.g. Johnson, 1998; Van De Vijver, 1998). These include: construct equivalence, measurement unit equivalence and scalar equivalence (Van De Vijver, 1998), which translate into the three levels of measurement invariance typically tested for by compara-

tive analysts: configural, metric and scalar invariance (e.g. Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Analysts commonly use a latent variable approach to test the extent to which multi-item instruments designed to measure latent constructs attain these levels of invariance (for a summary of other approaches, see Van der Vijver, 2018), the goal being to demonstrate the equivalence of the measure (Johnson, 1998), which is considered a prerequisite for making cross-group comparisons of different kinds. First, configural invariance ensures that the general structure of the measure is equivalent enough across the groups under comparison. In other words, the instrument (e.g., scale, questionnaire) should comprise the same number of constructs, and the composition of each should be deemed similar enough. Then, metric invariance implies that all items play a similar role in forming the constructs. In more technical terms, item loadings are established to be highly similar. This ensures that comparisons of relationships between constructs are trustworthy. Finally, reaching scalar invariance means that intercepts are invariant across the groups: latent means can be compared. Note that at each stage it is possible to rely on partially invariant scores with some items differing across the groups, but as a general rule, a minimum of two items should be invariant (Byrne, Shavelson, & Muthén, 1989).

As well as pointing to the presence or absence of equivalence, the advantage of the latent variable procedure for invariance testing is that it serves a diagnostic purpose by highlighting which items in particular contribute to a problem of nonequivalence, allowing the analyst to investigate likely causes and make necessary adjustments to the measure (e.g. by excluding the problematic item(s) in order to proceed). Researchers exploiting such an approach have tested for invariance in a wide variety of different measures but very rarely put into perspective several sources of nonequivalence. As an illustration, the present study focuses on three sources in particular: 1) scale format, 2) mode of data collection, and 3) cross-national variations. We briefly review research relating to each in the following.

2.2 Scale format

The effect of response scale format and labelling—e.g., the number of scale points, the presence of a middle point, unipolar vs. bipolar labelling, full vs. end-point labelling, construct-specific vs. agree/disagree formats—has been widely studied (see Krosnick and Presser, 2010 for an overview). Each aspect of how response scales in nonfactual questions are constructed can play a role in influencing how respondents interpret the question and answer it, by constraining how they communicate their subjective evaluation of the object of interest (Fowler & Consenza, 2008). Numerous experimental studies have demonstrated how sensitive respondents' answers are to variations in response format, and how this in turn can affect data quality (see Roberts, 2016 for a recent review). Particularly where multiple items are presented in a battery with the same response format (such as in multi-item scales designed to measure latent constructs), the response strategies respondents adopt to simplify the task of completing the scale (e.g. over-preferring a particular scale point) may affect the observed structure of relationships between the variables (ibid.). In a 3MC setting, this implies that not using the same scale format could endanger equivalence.

Two challenges relating to response format are particularly relevant. The first relates to how scale labels are translated. It can be difficult to find equivalent vague quantifiers for scale labels, making it hard to maintain the intended scale structure. The format or intended meaning of modifiers may be altered without intention, or deliberately adapted in the pursuit of fluency (Villar, 2009). How far resulting format differences jeopardize comparability in multi-lingual surveys has so far not been addressed extensively (Harkness, Villar, & Edwards, 2010). The second challenge is the possibility of systematic cultural variations in how respondents make use of answer scales (e.g. Yao et al., 2007), and the manifestation of different response effects (e.g. social desirability bias, response styles) (Johnson & Van de Vijver, 2003; Van Herk, Poortinga, & Verhallen, 2004). Thus, even where equivalent scale formats/ labelling can be found, measurement equivalence may still be compromised. A large-scale programme of research investigating measurement quality using different response formats through the use of Multitrait-Multimethod (MTMM) experiments embedded in the European Social Survey (ESS) has demonstrated the need to correct for differential measurement error produced by different formats in order to improve comparability across variant formulations, as well as across countries (see Saris & Gallhofer, 2014).

2.3 Mode of data collection

A key aspect of the survey design with implications for the structure of errors in estimates, as well as for costs, is the mode of data collection. Available budget and infrastructure and the extent of population coverage offered by a given mode determine the feasibility of using different data collection methods in different contexts. This has meant that many comparative surveys inevitably involve the use of multiple modes between countries. Furthermore, survey designers are increasingly opting to combine modes of data collection within countries, for example, to reduce survey costs by switching to web-based methods and then reduce the noncoverage and nonresponse error associated with web (e.g. by following-up nonrespondents with an alternative mode). Thus, despite the recognised benefits of input harmonisation in 3MC studies, mixed mode designs will continue to prevail.

Mode influences who is able to participate in a survey (coverage), who chooses to participate (nonresponse),

and how respondents answer questions (measurement) (De Leeuw, 2018). For example, interviewer-administered modes tend to provoke more social desirability bias in respondents' answers than self-administered modes, while response strategies such as 'straightlining' tends to be more common in self-administered modes (Couper, 2011; Tourangeau, 2017). Other systematic measurement differences have been observed when comparing only interviewer modes or when comparing self-administered modes (De Leeuw, 2018). For example, in the US, the tendency to give socially desirable, socially normative or acquiescent answers tends to be greater in telephone interviews than in face-to-face interviews, as is the tendency to satisfice (Holbrook, Green, & Krosnick, 2003).

For these reasons, testing for measurement invariance is becoming common practice in the evaluation of mixed mode data comparability. To date, the results of this research have been mixed, finding varying degrees of equivalence, depending on factors such as the nature of the measure analysed, the number of scale items, the criteria used to establish invariance and whether or not, as is recommended, differential selection errors (due to noncoverage and nonresponse) between modes are adjusted for (Hox, De Leeuw, & Klausch, 2017). For example, De Leeuw, Mellenbergh, and Hox (1996) failed to establish measurement invariance between self- and interviewer-administered versions of a (Dutch) measure of subjective wellbeing fielded in the Netherlands, as did Klausch, Hox, and Schouten (2013) in measures of perceptions of traffic, policing and police obedience (also in the Netherlands. Comparing face-to-face and telephone modes, Gordoni, Schmidt, and Gordoni (2011) tested for measurement invariance in a sensitive measure relating to social integration attitudes among the Arab minority in Israel and found factor loadings to be invariant, but some differences in item intercepts, resulting in only weak partial scalar invariance overall between the modes for the measure analysed. A number of other studies find variation in the level of invariance between modes attained depending on the measure analysed (e.g. Heerwegh & Loosveldt, 2011; Martin & Lynn, 2011; Revilla, 2013), or the time of measurement, in the case of a longitudinal survey (Cernat, 2015). To the extent that data collection mode mediates how respondents react to scale format and the tendency for socially desirable responding, cross-cultural differences in response style may also influence the relative prevalence of mode effects observed in a comparative mixed-mode survey setting, though we are not aware of studies that have addressed this.

2.4 Cross-national variations

3MC studies generally imply relying on data collected in different national, cultural and/or linguistic settings. Culture and language represent two closely intertwined potential sources of nonequivalence. Instruments cannot simply be exported from one culture to another (Chen, 2008); indeed, "an appropriate translation requires a balanced treatment of psychological, linguistic, and cultural considerations" (Van De Vijver & Tanzer, 2004, p. 122). First, the manifestations of psychological, social and political phenomena are likely to differ across cultures. For instance, depression is expressed through somatization to a greater or lesser extent in some cultures, which should be taken into account when comparing levels of depression across ethnic or national groups (Dere et al., 2015). Second, the use of different languages may also harm the comparability of data collected in different places. Besides translation errors (not only in question wording, but in the response alternatives too), which happen despite the great care devoted to survey preparation in large-scale comparative studies, inaccurate translations may stem from overliteral interpretations of instruments that should have been adapted (Harkness et al., 2010). Finally, cross-cultural differences are also observed in the way individuals give their responses. Response styles, such as using extreme vs. middle points or showing acquiescence have been shown to vary systematically across national contexts (Batchelor & Miao, 2016; Van Vaerenbergh & Thomas, 2012). For example, survey respondents in collectivistic cultures—in which memberships of social and familial groups are highly important—have a greater tendency to agree with the content of items (acquiescence bias), compared to respondents in more individualistic cultures (Johnson, Kulesa, Cho, & Shavitt, 2005).

With a few exceptions, studies that have assessed the cross-cultural equivalence of survey instruments have considered differences between countries. However, the invariance of measurement is not guaranteed within countries with more than one official language, since both linguistic and compositional differences across regions may harm the comparability of data (Zavala-Rojas & Saris, 2018). With this assumption in mind, Davidov and De Beuckelaer (2010) used data from the 2004 and 2006 European Social Survey (ESS) to assess the equivalence of the Schwartz Human Values Scale in multilingual countries (Switzerland and Belgium), in neighbouring countries sharing a language (France, German, Austria, and the Netherlands), and in countries with a different language. Measurement invariance was higher across countries sharing a language than it was either across countries with no shared languages or within multilingual countries. Meanwhile, using the 2002 ESS data from Switzerland, Sarrasin, Green, Berchtold, and Davidov (2012) found that while support for stricter immigration criteria was higher among members of the German-speaking majority than among members of the French-speaking minority, a reverse pattern emerged in the case of a specific criterion. Further examination of the items used revealed that this was most probably entirely due to non-equivalent translations, with an additional verb in French distorting the meaning of the question.

3 The present study

The present study illustrates a way to compare different sources of measurement nonequivalence through an analysis of data from a module of questions on personal and subjective wellbeing from Round 3 (2006) of the European Social Survey (later repeated in ESS Round 6). The ESS wellbeing data have been widely analysed (e.g. Clark & Senik, 2011; Group., 2017; Kööts-Ausmees, Realo, & Allik, 2013; Soons & Kalmijn, 2009) as they provide a basis for constructing key social indicators of quality of life across European countries. Comparisons of national wellbeing are frequently reported in the media, and are often based on simple country-level comparisons across single-item measures or multi-item indexes, in some cases based on survey data involving multi-mode data collection. However, to carry out unbiased comparisons, measurement invariance needs to be ensured first.

We used ESS 2006 data from Switzerland and two of its neighbours with a common language, France and Germany (European Social Survey, 2012). Switzerland is a multilingual country, the most widely spoken national languages of which are German (roughly 64%) and French (roughly 23%). The selection of these three countries, therefore, allows us to investigate measurement invariance across different languages and cultures, both within and between countries. In addition, we take advantage of the presence of a number of variations in adaptations to the English-language source questionnaire identified in the different questionnaires used in each country (described in further detail below). These variations—which affected both the format (response scales and scale labels) and wording of questions—may partly be the result of translation errors and a failure to harmonise translations across countries with shared languages, or of deliberate choices made by the translation teams about how best to formulate questions in a particular language, in a particular cultural context. Finally, the selection of these three countries allows us to take advantage of two methodological experiments that were conducted alongside the main fieldwork in Switzerland and Germany in this survey round (see sections 3.4 and 3.5).

Through this combination of natural and experimentally designed differences, the present study compares the extent of measurement invariance between groups to draw conclusions about the relative impact of the three different sources of comparison error reviewed above: 1) scale format; 2) mode of data collection and 3) cross-national variations. Regarding the latter, while the impact of culture and language on data comparability are conceptually closely intertwined, for analytical purposes we had to consider country and language separately in the empirical analyses. Overall, we use a step-by-step analytic approach, assessing measurement invariance between different groups at each step, which enables us to draw conclusions about the relative impact of each error source through a process of elimination.

3.1 Data

As mentioned, we use data from the Swiss, German and French editions of the 2006 European Social Survey (European Social Survey, 2006) and data from methodological studies that were run alongside the survey in the same year (described below). The population for the ESS is all (legal) residents in a country, aged 15 and over, and the survey is conducted by face-to-face (CAPI) interviews (lasting around one hour). Details of the sample designs and fieldwork documentation are available at www.europeansocialsurvey.org. The combination of natural (languages, countries, translations) and experimental (question format and mode) features resulted in 16 different groups (see Table 1 for description and sample sizes). A document in the online appendix also shows all translations used in our analyses.

3.2 Questions analysed

The ESS 2006 module on personal and subjective wellbeing includes 55 items intended to be combined in different ways to capture different dimensions of wellbeing (see Huppert, Clark, Frey, Marks, & Siegrist, 2005; Huppert et al., 2009). This makes a latent variable approach to the evaluation of measurement invariance particularly suitable and this is the approach we adopt. To ensure a sufficient 'respondents by parameter' ratio (Kline, 2011), we decided to focus on two dimensions of subjective wellbeing. The first dimension (three items) is what is referred to by the authors of the ESS module (ibid.) as "evaluative wellbeing" and includes two measures of life satisfaction (satisfaction with your life "so far" and "overall") and a measure of happiness (question wording from the source questionnaire is shown in Table 2) to constitute a global evaluation of an individual's wellbeing. The second dimension is referred to as "emotional wellbeing" and is here comprised of four items measuring negative affect (frequency of feeling depressed, lonely, sad, and anxious during the past 7 days), which should be negatively associated with the evaluative dimension.

3.3 Differences in question adaptation between languages

Different adaptations to the English source questions resulted in different formulations of the items in the French and German language questionnaires used in Switzerland, France and Germany. The differences between the questionnaires are documented in column 3 of Table 2. In the measure of evaluative wellbeing, the adaptations concerned the scale format. The modifier "extremely" dissatisfied/ satisfied in the overall life satisfaction measure was translated as "very" in French-speaking Switzerland, which might be expected to attract more responses at the end-point than the source modifier. In France, the "extremely dissatisfied" endpoint label was changed to "not at all satisfied", which has the

Table 1
Groups compared at each analytic step and sample sizes

Nr.	*N*	Sources of measurement invariance	Sample
Step 1: Scale Format			
1	120	Bipolar scale as in source (Extremely dissatisfied)	Switzerland (CH-FR)
2	146	Unipolar (Not at all satisfied)	Switzerland (CH-FR)
3	133	Bipolar with less extreme modifiers (Very dissatisfied)	Switzerland (CH-FR)
4	425	Bipolar scale as in source (Extremely dissatisfied)	Switzerland (CH-DE)
5	439	Unipolar (Not at all satisfied)	Switzerland (CH-DE)
6	443	Bipolar with less extreme modifiers (Very dissatisfied)	Switzerland (CH-DE)
Step 2: Mode			
7	400	CAPI[a,b]	Switzerland (CH-FR)
8	237	CATI	Switzerland (CH-FR)
9	2681	CAPI[b]	Germany (DE)
10	199	CATI	Germany (DE)
Step 3: Language (within country)			
11	400	French	Switzerland (CH-FR)
12	1308	German	Switzerland (CH-DE)
Step 4: Country (within language)			
13	1760	France	France (FR)
14	400	French-speaking Switzerland	Switzerland (CH-FR)
15	1308	German-speaking Switzerland	Switzerland (CH-DE)
16	2681	Germany	Germany (DE)

[a] Respondents in French-speaking and bilingual French and German cantons only.
[b] Respondents with fixed-line telephones only.

effect of modifying the bipolar response scale into a unipolar scale. This could affect how respondents map their answer to the scale, particularly at the endpoint affected, but also at the midpoint, because unipolar scales do not offer a neutral "neither/nor"-type alternative. In the happiness item, the Swiss-French again uses "very" instead of "extremely" as the modifier, while in the life satisfaction so far item, both France and French-speaking Switzerland used "very" instead of "extremely" as the modifier.

In the measure of emotional wellbeing, one of the four negative affect measures was affected by variant adaptations. Specifically, "felt anxious" in the source was translated as "felt worried" in France and Switzerland (both in French and German), while in Germany the translation "waren Sie ängstlich" was used. While closer to the origin of the word "anxious", to be "ängstlich" has stronger negative connotations than to feel "worried", capturing both an idea of fearfulness and a more persistent trait-like characteristic than does "worry".

3.4 Scale format experiment

In Switzerland, a split-ballot experiment was used to compare the variant scale formats and labels that were used in France and in French-speaking Switzerland (described below). The data allow us, therefore, to assess the relative impact of some of the above-mentioned adaptation errors on comparability. The scale format experiment was embedded in the Swiss supplementary questionnaire (administered by the face-to-face interviewer at the end of the main interview) as part of some additional country-specific tests not included in the main survey (Ernst Stähli et al., 2019; Joye, Schöbi, Pollien, & Kaenel, 2010). Respondents to the main interview were randomly assigned to one of three versions of the overall life satisfaction question, allowing an experimental comparison between the three variant modifiers in use (1: "Extremely satisfied" for an 11-point symmetric bipolar scale with a literal translation of the modifiers, 2: "Not at all satisfied", for an 11-point asymmetric unipolar scale, and 3: "Very dissatisfied", for an 11-point symmetric bipolar scale as in the source version, but with more "natural" sounding, but less extreme modifiers), as well as an analysis of within-subject response reliability between question formats (not presented here).

3.5 ESS Round 3 CATI Experiment

The second experiment was a mode comparison study, carried out by the Core Scientific Team of the ESS in collaboration with the Swiss and German national coordina-

Table 2
Measures of wellbeing analysed, source formulation and variants

	ESS		Source Question		Variant Question	
Variable	Name	ID	Formulation	Used in	Formulation	Used in
Evaluative Wellbeing						
Overall life Satisfaction	stflife	B24	All things considered, how satisfied are you with your life as a whole nowadays?			
			0 “Extremely dissatisfied” 10 “Extremely satisfied”	DE, CH-DE, CH-FR (Test)	0 “Not at all satisfied” 10 “Extremely satisfied”	FR, CH-D, CH-FR (Test)
					0 “Very dissatisfied” 10 “Very satisfied”	CH-FR, CH-D, CH-FR (Test)
Happiness	happy	C1	Taking all things together, how happy would you say you are?			
			0 “Extremely unhappy” 10 “Extremely happy”	FR, DE and CH-DE	0 “Very unhappy” 10 “Very happy”	CH-FR
Satisfication with life so far	stflfsf	E31	How satisfied are you with how your life has turned out so far?			
			0 “Extremely dissatisfied” 10 “Extremely satisfied”	DE, CH-DE	0 “Very dissatisfied” 10 “Very satisfied”	FR, CH-FR
Emotional Wellbeing (Negative Affect)			Please tell me how much of the time during the past week <*phrase*>. 1 “None or almost none of the time”, 2 “Some of the time” 3 “Most of the time” 4 “All or almost all of the time”			
Depressed	fltdpr	E8	. . . you felt depressed	FR, DE, CH-DE, CH-FR		
Lonely	fltlnl	E12	. . . you felt lonely?	FR, DE, CH-DE, CH-FR		
Sad	fltsd	E14	. . . you felt sad?	FR, DE, CH-DE, CH-FR		
Anxious	fltanx	E17	. . . you felt anxious?		. . . waren Sie ängstlich?	DE
					. . . Sie sich Sorgen gemacht haben?	CH-DE
					. . . vous vous êtes senti/e inquiet/ète?	FR, CH-FR

(FR) France, (DE) Germany, (CH-FR) French-speaking Switzerland, (CH-DE) German-speaking Switzerland. For full formulations and translations, please refer to the online appendix.

tors of the ESS in 2006–2007 (Roberts, Eva, Lynn, & Johnson, 2010)[1], designed to test the feasibility of conducting ESS interviews by computer-assisted telephone interviewing (CATI). Conducted in French-speaking Switzerland and Germany, the CATI experiment involved the random assignment of random probability samples (of ESS population members with (at least one) fixed-line telephone number) to one of three treatments involving questionnaires of different lengths (Version A: 1 hour, Version B: 45 minutes and Version C: 30 minutes). For the analyses presented here, we used data from respondents assigned to versions A and B.[2]

4 Results

4.1 Analytic approach

To compare the quality of measurements obtained in the different modes, it is necessary to first control for potential mode effects on selection error due to noncoverage and nonresponse (Hox et al., 2017). To decide how to render as comparable as possible the samples surveyed in each mode, we performed preliminary analyses of the data (Section 4.2). Then (Section 4.3), we used a multi-step procedure to enable us to draw conclusions about the relative impact of scale format, mode of data collection, and cross-national variations (country and language) through a process of elimination (starting with scale format and mode—countries and regions could be compared if the preliminary steps were proved to be invariant)[3]. Across the 16 groups in the study design, we conducted a total of seven separate tests of measurement invariance in four steps (described below). We started with scale format, since differences too large in scale format would preclude any further analysis.

4.2 Preliminary analyses

We used two approaches to control for potential mode effects on selection error. Firstly, as the greatest source of selection error in the CATI sample was noncoverage (the samples were restricted to residents in households with a fixed line telephone number), we decided to select from the main face-to-face samples only those respondents who reported (in the survey) having a fixed-line telephone in their household. Secondly, we created a propensity score weight (based on a logistic regression model) to control for the remaining (observed) socio-demographic variables on which the samples in each mode were found to vary, following procedures recommended for mode comparison studies testing for measurement invariance (e.g. Hox et al., 2017) and general recommendations for addressing questions of causal inference in social and psychological research (Harder, Stuart, & Anthony, 2010; Rosenbaum & Rubin, 1983). As no auxiliary data were available for the total samples, we examined relative differences across categories of socio-demographic characteristics of the achieved (selection-probability weighted) samples in each mode, on the assumption that these variables were least likely to be affected by mode effects on measurement. The results[4] indicated differential selection effects between modes to be adjusted for when assessing measurement invariance between modes. Our tests of measurement invariance between the modes were conducted first on the selection-probability weighted data to assess the overall comparability of the measure in both modes, then using the combined (selection probability plus propensity-score) weights. The weight is assumed to adjust for the main sources of selection error in the two modes, such that remaining differences observed can be attributed to mode effects on measurement.

4.3 Assessing the relative impact of sources of measurement nonequivalence

Multigroup Confirmatory Factor Analysis (MGCFA) performed with Mplus 7.4 was used to test for measurement invariance. To do so, we proceeded by testing the three usual steps—each one stricter than the previous—of measurement invariance: configural, metric and scalar. Models based on latent variables, including MGCFAs, are considered to fit the data adequately when the Comparative Fit Index (CFI) is equal or superior to 0.95 and when the Root Mean Square of Approximation (RMSEA) is equal to or lower than 0.05 (L.-T. Hu & Bentler, 1999), although values between 0.05 and 0.08 are also considered acceptable (Schermelleh-Engel, Moosbrugger, & Müller, 2003). A stricter level of invari-

[1]ESS ERIC reserves the right to make the relevant data sets available upon request

[2]The analysis reported here is not concerned with these treatments. However, the design affected the location of the personal and social wellbeing module in one of the questionnaire versions (C). To avoid the potentially confounding effects of questionnaire length on response quality, we use CATI data from respondents who completed versions A and B only. In these groups, the design of the questionnaire was almost identical, though the omission of the rotating module preceding the module on wellbeing in version B may affect the context of the questions and potentially affect the comparability of the answers given. However, to ensure sufficiently large sample sizes for the analysis of measurement invariance, we decided to overlook this design limitation.

[3]Replication materials for data preparation and analyses are provided as supplementary material to this article.

[4]Significant relationships between the covariates and the likelihood of being interviewed by telephone were relatively weak and the Nagelkerke pseudo R-square values were low at 0.13 for Switzerland and 0.09 for Germany. However, controlling for the effects of the other covariates, place of residence, main activity, household size and years of education were significantly related to the probability of being interviewed by telephone compared to face-to-face in both French-speaking Switzerland and Germany (and for the latter, being hampered in daily life by disability or illness was also a statistically significant covariate).

ance is considered as reached if chi-square values do not differ significantly (note that due to the estimation method chi-square values were rescaled prior to comparison; Satorra & Bentler, 2001).[5] If the differences between two models are deemed too large, modification indices (MI) indicate which parameters (due to their contribution to the chi-square statistic) should be modified. Results are present in Table 3.

Step 1: Impact of scale format. We first compared responses to the three scale formats separately within French-speaking Switzerland (1.1: groups 1, 2 and 3 in Table 1) and German-speaking Switzerland (1.2: groups 4, 5 and 6). If the variant question formats used in the different language questionnaires were found to affect data comparability, further comparisons between the countries and language regions would not be possible. In French-speaking Switzerland, results show that configural, full metric and full scalar invariance were achieved. Similar results were obtained in German-speaking Switzerland. These two sets of results indicate that in the present case, question format does not affect data comparability. This means that at further steps of the analysis, for instance when the effect of mode is investigated (Step 2), we can use data from the main questionnaire, despite the variant question formats used across the different languages.

Step 2: Impact of mode. At Step 2, we investigate the effect of mode, comparing the CAPI and CATI samples both in French-speaking Switzerland (2.1: groups 7 and 8) and in Germany (2:2: groups 9 and 10), controlling for selection errors between modes as described above (note that both in Switzerland and Germany similar conclusions were reached with the unweighted data). This provides two tests of mode, controlling for language and national cultural differences, allowing us to draw conclusions about the potential risks of combining modes within and between countries.

In French-speaking Switzerland, while configural invariance was achieved, full metric invariance was not. Modification indices show that the loading for the life satisfaction variable (Stflife) differed across the two groups. If the constraint equality placed on this loading is relaxed, the model is no longer significantly different from that testing for configural invariance. Partial metric invariance can thus be considered as having been reached. In a last step, we test for partial scalar invariance (items whose loadings are free to vary cannot have their intercept constrained to equality). The model received adequate fit indices. In Germany, all three steps of invariance could be considered as having been reached.

Step 3: Impact of language within countries. At Step 3, we investigated the effect of language within Switzerland by comparing the French and German-speaking regions (groups 11 and 12), controlling for shared national culture and socio-economic conditions that could impact on people's wellbeing (although note language differences between regions are confounded here with both translation and socio-cultural differences). Here, while configural and full metric invariance were reached, attaining full scalar invariance appeared to be an issue. Modification indices indicated first, that the happiness variable (Happy) was the most problematic item (i.e., it contributed most to the chi-square value). The resulting model—testing for partial scalar invariance—was however still significantly different from that testing for full metric invariance. It was necessary to relax the equality constraint for the loneliness item (Fltlnl) in order to get a model that could be considered as equivalent. Thus, despite the difficulties, since at least two items were found to be invariant per factor, it would still be possible either to merge the data from the two linguistic regions or to compare their means in a safe way.

Step 4: Impact of country within languages. Finally, at Step 4, we investigated the effects of cross-national variations on measurement invariance, by comparing France with French-speaking Switzerland (4.1: groups 13 and 14), and Germany with German-speaking Switzerland (4.2: groups 15 and 16). Results indicated that the latent means for the subjective wellbeing measure from French-speaking Switzerland and France should not be compared. Indeed, while configural invariance was reached, achieving full metric invariance appeared not to be possible. The equality constraint for the loading for the life satisfaction variable (Stflife) had to be relaxed, which resulted in a model that did not differ significantly from that testing for configural invariance. The next model—testing for partial scalar invariance—was, nevertheless, still significantly different. Modification indices suggest that the intercept of the two other satisfaction items should be freed. It was not possible to go ahead, since the satisfaction factor consisted of three items only, and at least two items need to be invariant to perform reliable latent means comparisons. The results also indicate that German data should not be compared to German-speaking Swiss data either. Here again, while configural invariance was achieved, full metric invariance could not be reached. Modification indices indicate that the equality constraint first for "anxious" then for "happy" should be relaxed. The resulting model was however still significantly different from that testing for configural invariance. The modification indices obtained in this last model only suggested cross-loadings, which means that we were not able to go further with the invariance measurement testing.

5 Discussion

When designing surveys and making comparisons between groups using survey data, ensuring that all respondents receive(d) the same or functionally equivalent question stim-

[5] Chen (2007) has suggested using differences in CFI and RSMEA values. However, due to the relatively small sample sizes, chi-square difference tests were preferred here.

Table 3
Results of measurement invariance testing

Model	Δdf	$\Delta_{\text{scaled}\chi^2}$	*p*	CFI	RMSEA
Step 1: Scale format					
1.1. Within French-speaking Switzerland (Groups 1,2, and 3)					
Configural				0.990	0.039
Full metric	10	10.61	0.39	0.990	0.036
Full scalar	10	6.82	0.74	0.993	0.027
1.2. Within German-speaking Switzerland (Groups 4,5, and 6)					
Configural				0.979	0.039
Full metric	10	6.72	0.75	0.983	0.031
Full scalar	10	7.12	0.71	0.984	0.027
Step 2: Mode					
2.1. Switzerland (Groups 7 and 8)					
Configural				0.980	0.043
Full metric	5	11.34	0.05	0.972	0.047
Partial metric	4	1.14	0.89	0.984	0.036
Partial scalar	4	2.44	0.66	0.986	0.032
2.2. Germany (Groups 9 and 10)					
Configural				0.990	0.025
Full metric		4.33	0.50	0.991	0.022
Full scalar		2.37	0.80	0.991	0.020
Step 3: Impact of language within countries					
3.1. Within Switzerland (Groups 11 and 12)					
Configural				0.991	0.027
Full metric	5	2.47	0.78	0.993	0.023
Full scalar	5	36.63	0.00	0.979	0.036
Partial scalar I	4	13.78	0.01	0.988	0.027
Partial scalar II	3	5.22	0.16	0.991	0.024
Step 4: Impact of country within languages					
4.1. Across French-speaking regions and countries (Groups 13 and 14)					
Configural				0.983	0.044
Full metric	5	19.88	0.00	0.978	0.045
Partial metric	4	5.42	0.25	0.982	0.041
Partial scalar I	3	16.25	0.00	0.978	0.043
4.2. Across German-speaking regions and countries (Groups 15 and 16)					
Configural				0.988	0.032
Full metric	5	56.26	0.00	0.975	0.042
Partial metric I	4	17.37	0.00	0.984	0.034
Partial metric II	3	8.96	0.03	0.986	0.032

uli is key to drawing valid conclusions about group differences. Where the focus is on comparing populations living in different national, sociocultural, and/or linguistic contexts, this requirement is imperative, but entails numerous challenges for those responsible for designing and implementing the survey, as well as for analysts, on whom the onus is to demonstrate empirically that measurement invariance has been achieved. From a total survey error perspective (Biemer & Lyberg, 2003), identifying potential sources of nonequivalence and evaluating which is likely to be the most damaging to the accuracy of estimates and comparisons is crucial for guiding survey design decisions about where to invest limited resources. For the analyst, uncovering the reasons for an absence of measurement invariance across groups is not only important for deciding how to proceed with comparisons, but can also shed light on important substantive differences of

interest (Davidov et al., 2014; Meuleman & Schlüter, 2018).

In this study, we illustrated a way to evaluate the respective impact of three principal sources of nonequivalence in comparative research (that is, adaptations to scale format, different modes of data collection, and cross-national variations) using a combination of survey data from the French, Swiss and German editions of the European Social Survey (2006) and data from methodological experiments conducted alongside the main ESS in Switzerland and Germany. We conducted tests of measurement invariance for a measure of subjective wellbeing across 16 groups, sequentially, using a process of elimination to evaluate the relative influence of the three aforementioned sources of nonequivalence. Overall, variant translations and other cross-national variations were found to be greater sources of nonequivalence than scale format and mode of data collection.

5.1 Scale format and mode

The results suggest that in the present case small modifications to scale point labels (affecting the extremeness of modifiers) and scale format (unipolar vs. bipolar) had no significant effect on the measurement of wellbeing and measurement invariance was unaffected. Having established that these variant item formulations produced equivalent measurements within linguistic regions in Switzerland, we worked on the assumption that we could proceed to making comparisons between linguistic regions and countries that had used the variant formulations in the main questionnaire. On this basis, we turned to the question of whether and if so, how, different modes of data collection affect measurement invariance. We compared modes of data collection within a linguistic region (French-speaking Switzerland), and within a country (Germany), to control other possible sources of measurement nonequivalence, on the assumption that if data gathered in different modes are found to be comparable, modes might be mixed across cultural and linguistic groups. Our analyses revealed only one problematic item (in French-speaking Switzerland), and we were able to achieve partial scalar invariance (in Switzerland) and full scalar invariance in Germany. This means that in the present case data collected with different modes could be safely merged for further analyses.

5.2 Language and culture

Having established that (with the data used here) face-to-face and telephone interviews produced adequately comparable measurements to permit combining these modes in a cross-national context, we turned to the question of whether different language questionnaires produced equivalent measurements within a given national culture, by testing for measurement invariance between the French- and German-speaking regions of Switzerland, controlling for the mode of data collection. Our tests confirmed that despite the use of slightly different modifiers for scale points ("extremely" versus "very", which were previously established to not affect the comparability of the measures), it was possible to attain partial scalar invariance. However, two items appeared to behave differently between the two regions: happy and lonely. Residents of the German-speaking regions were less likely than those in the French-speaking region to report feeling lonely and were marginally more likely to report feeling happy, though the mean differences were small (descriptive statistics are shown in the Appendix). One possible explanation for such findings (besides true variance) could be that the translations for these terms have different connotations in the different regions, or that social norms governing the willingness to self-report happiness and loneliness in surveys vary regionally for cultural reasons.

Finally, we turned to the question of how comparable the wellbeing measure was across different national contexts, controlling for language, by comparing respondents in France with those in French-speaking Switzerland, and respondents in Germany with those in the German-speaking part of Switzerland. This step of analysis proved to be the most problematic. Only partial metric invariance was reached in French, and nothing at all beyond configural invariance was found in German. In both cases, comparisons of latent means across the groups under investigation should not be attempted. Once again, the "happy" item was problematic, perhaps for the same reasons as for cross-lingual comparisons within Switzerland. However, invariance was also curtailed here by the presence of the different translation of the word "anxious" in Germany (ängstlich) compared to the Swiss-German version (Sorgen machen), the former having stronger negative connotations (see 3.3). This is reflected in the lower mean score observed for this item for the German respondents compared with the Swiss German respondents (see Appendix). These findings highlight the need for caution even when comparing countries with shared languages, despite the more reassuring findings of earlier research (Davidov & De Beuckelaer, 2010).

5.3 Comparing sources of nonequivalence

Thus, our research suggests that variant translations of concepts and the inherent national and cultural differences in how constructs are understood and communicated about (what Van der Vijver (1998) terms "construct bias") play a more important role in determining the comparability of survey measures than the other sources of nonequivalence that we analysed. In the case of subjective wellbeing, it seems that differences in the social norms governing how people express how they feel and their willingness to report experience of negative emotions may vary culturally, which, in turn, affect how these experiences can be measured. In this respect, the process of testing for measurement invariance, and the findings obtained, shed light not only on the relative impact

of different artefactual influences on measurement equivalence, but also on the substantive phenomenon of interest: cross-cultural variation in the experience and expression of subjective wellbeing. Such conclusions might suggest possible avenues for more in-depth research for comparative analysts interested in comparing the groups we analysed here.

Our findings correspond with those of other studies that have demonstrated the difficulties of finding full scalar equivalence of measures in 3MC studies, particularly where variant translations are discovered between questionnaires (e.g. Davidov et al., 2014; Sarrasin et al., 2012). They also extend the findings of that research, however, by not only uncovering the effects of mistranslations, for example, but by investigating their impact relative to other sources of nonequivalence through the use of data from experiments.

5.4 Limitations

A number of caveats should be discussed, as our findings derive from analyses of measurement invariance of a single, two-dimensional latent measure of wellbeing, across a limited number of groups and modes of data collection (at a particular point in time). Furthermore, the analyses were only possible due to the availability of a rare combination of comparative survey data exhibiting divergent questionnaire adaptations across countries with shared languages and within a multi-lingual country, and data from methodological experiments (testing different modes and scale formats) conducted alongside the main survey. The specificity of the design limits the generalizability of the results, and certain features should be borne in mind when interpreting the findings. For one, the test-retest design of the Swiss question format experiment might have affected the observed comparability of answers obtained within the two linguistic regions. The test was included in the supplementary questionnaire at the end of the survey, after respondents had already answered the life satisfaction measure in the main questionnaire. If respondents recalled their prior answer to the question, they may have adapted their later answer to make it consistent with their initial report, thus producing more comparable measurements than might be obtained in a standard split-ballot design (though if this were a major concern, it would also invalidate the results of the ESS MTMM experiments, which rely on the same design).

For another, while our results pertaining to the equivalence of measures across modes suggested mode had less influence on measurement equivalence than the combined effects of language, translation and culture, the fact that we only compared interviewer-administered modes means that we likely underestimate the importance of mode mixing as a source of nonequivalence. Our findings are consistent with those of other studies (e.g. Gordoni et al., 2011—though ours are more optimistic) in that they provide additional evidence that interviewer-administered modes can produce comparable composite measurements (and thus be mixed within research designs without compromising equivalence), despite small differential effects on the distribution of responses to individual items. However, it is well established that mode effects on the comparability of measurement are most problematic where interviewer modes are combined with self-administered modes (De Leeuw, 2018; Tourangeau, 2017). Another key point is that the total survey error of estimates (i.e. the combined effect of selection and measurement effects that are known to be affected by modes) and hence, the accuracy of comparisons in comparative research, is measurement-specific. Thus, there is no basis on which to assume that other measures—even taken from the same study—would be free from differential mode effects to the same extent.

6 Conclusion

Assessments of the relative importance of different sources of nonequivalence are valuable both from the perspective of survey designers and comparative analysts. For the comparative researcher, being able to rule out methodological confounds affecting comparability makes it possible to shed light on true sources of variation in substantive phenomena of interest, and the possibility to measure them. Despite its limitations, our study presents a helpful illustration of how to combine data from methodological experiments with survey data for this purpose, to make it possible to eliminate conflicting explanations for possible causes of nonequivalence. To facilitate this, one recommendation would be for survey designers to incorporate methodological experiments in the main fieldwork for comparative surveys in such a way as to enable analysts to draw conclusions about the causes of measurement invariance, if it is observed for key survey measures. This could include within-country/within-language tests of variant scale formats and translations, and where multiple modes may be used within and across groups, randomised tests of mode effects on measurement. Careful design could also get analysts closer to disentangling the effects of language, translation and culture when considering cross-national variations as a source of nonequivalence, which were partly confounded in the present study.

For survey designers, however, the results presented here also lend support to the conclusion that data comparability in 3MC settings can be enhanced through greater investment in questionnaire development and translation procedures (see e.g. De Jong, Dorer, Lee, Yan, & Villar, 2019; Miller, 2019). This recommendation is consistent with current practice on the ESS, which, since 2006 (the round of data analysed here), has introduced a number of measures to improve translation procedures, with the aim of limiting cultural comparability issues prior to data collection. These include advance translations to assess translatability, qualitative pretests in

the form of cognitive interviewing (in addition to quantitative pretests), as well as post-hoc controls of the outputs of the collaborative translations procedures. As ever in survey design, the financial costs of such procedures should be weighed up against the data quality improvements they offer, as well as against reductions in total survey error (including comparison errors), which could potentially be gained from investing in other parts of the survey design (such as embedding methodological experiments).

References

Batchelor, J., & Miao, C. (2016). Extreme response style: A meta-analysis. *Journal of Organizational Psychology*, *16*(2), 51–62.

Biemer, P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(817–848).

Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality.* Hoboken: Wiley.

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*(3), 456–466.

Cernat, A. (2015). *Using equivalence testing to disentangle selection and measurement in mixed mode surveys.* Understanding Society Working Paper Series, No. 2015-01.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018.

Clark, A., & Senik, C. (2011). Is happiness different from flourishing? Cross-country evidence from the ESS. *Revue D'Economie Politique*, *121*(1), 17–34.

Couper, M. (2011). The future of modes of data collection. *Public Opinion Quarterly*, *75*(5), 889–908.

Davidov, E., & De Beuckelaer, A. (2010). How harmful are survey translations? A test with Schwartz's human values instrument. *International Journal of Public Opinion Research*, *22*(4), 485–510.

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

De Jong, J., Dorer, B., Lee, S., Yan, T., & Villar, A. (2019). Overview of questionnaire design and testing. In T. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC) (wiley series in survey methodology)* (pp. 115–138). Hoboken, NJ.: John Wiley & Sons.

De Leeuw, E. (2018). Mixed-mode: Past, present, and future. *Survey Research Methods*, *12*(2), 75–89.

De Leeuw, E., Mellenbergh, G., & Hox, 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 and Research*, *24*(4), 443–472.

Dere, J., Watters, C. A., Yu, S. C.-M., Bagby, R. M., Ryder, A. G., & Harkness, K. L. (2015). Cross-cultural examination of measurement invariance of the Beck Depression Inventory—II. *Psychological Assessment*, *27*(1), 68–81.

Ernst Stähli, M., Joye, D., Sapin, M., Pollien, A., Ochsner, M., & Van den Hende, A. (2019). Translation and question format experiments: ESS 2006, ESS 2008, ESS 2010, ESS 2012, ESS 2014, EVS 2008, ISSP Pilot 2015. Dataset. Distributed by FORS, Lausanne, 2019. doi:10.23662/FORS-DS-709-1

European Social Survey. (2006). *Round 3 data.* Norwegian Social Science Data Services, Norway. Data Archive and distributor of ESS data.

European Social Survey. (2012). *Round 6 data. data file edition 2.1.* Norwegian Social Science Data Services, Norway. Data Archive and distributor of ESS data.

Fowler, F., & Consenza, C. (2008). Writing effective questions. In E. D. Leeuw, J. J. Hox, & D. A. Dillmans (Eds.), *International handbook of survey methodology* (pp. 136–160). New York: Lawrence Erlbaum Associates.

Gordoni, G., Schmidt, P., & Gordoni, Y. (2011). Measurement invariance across face-to-face and telephone modes: The case of minority-status collectivistic-oriented groups. *International Journal of Public Opinion Research*, *24*(2), 185–287.

Group., T. (2017). Comparative impact study of the European Social Survey (ESS) ERIC. final report. Retrieved from https://www.europeansocialsurvey.org/docs/findings/ESS-Impact-study-Final-report.pdf

Groves, R., Fowler Jr., F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken: Wiley.

Häder, S., & Lynn, P. (2007). How representative can a multi-nation survey be? In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally. lessons from the european social survey* (pp. 33–52). London: Sage Publications.

Harder, V., Stuart, E., & Anthony, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234–249.

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. P. Mohler, … T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 115–140). Hoboken: Wiley.

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. (2003). Telephone versus face-to-face interviewing of national probability samples with long 1 questionnaires. *Public Opinion Quarterly*, *67*(1), 79–125.

Hox, J., De Leeuw, E., & Klausch, L. (2017). Mixed mode research: Issues in design and analysis. In P. Biemer, E. D. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, … B. West (Eds.), *Total survey error in practice: Improving quality in the era of big data* (pp. 511–530). Hoboken: Wiley.

Hu, E., Stavropoulos, V., Anderson, A., Clarke, M., Beard, C., Papapetrou, S., & Gomez, R. (2019). Assessing online flow across cultures: A two-fold measurement invariance study. *Frontiers in Psychology*, *10*, 407.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fix indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Huppert, F., Clark, A., Frey, B., Marks, N., & Siegrist, J. (2005). ESS module proposal: Personal and social well-being, creating indicators for a flourishing Europe. Retrieved from https://www.europeansocialsurvey.org/docs/round3/questionnaire/ESS3_huppert_proposal.pdf.

Huppert, F., Marks, N., Clark, A., Siegrist, J., Stutzer, A., Vittersø, J., & Warendorf, M. (2009). Measuring well-being across Europe: Description of the ESS well-being module and preliminary findings. *Social Indicators Research*, *91*(3), 301–315.

Johnson, T. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In J. A. Harkness (Ed.), *Zuma-nachrichten spezial volume 3: Cross-cultural survey equivalence* (pp. 1–40). Mannheim: ZUMA.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*(2), 264–277.

Johnson, T., & Van de Vijver, F. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–206). New York: Wiley.

Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist*, *42*(2), 168–177.

Jowell, R., Kaase, M., Fitzgerald, R., & Eva, G. (2007). The European Social Survey as a measurement model. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally. Lessons from the European Social Survey.* London: Sage Publications.

Joye, D., Schöbi, N., Pollien, A., & Kaenel, C. (2010). European Social Survey, Switzerland—2006. Dataset. Service suisse d'information et d'archivage de données pour les sciences sociales - SIDOS, Neuchâtel. Distributed by FORS,Lausanne, 2010. doi:10.23662/FORS-DS-568-4

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

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.

Kööts-Ausmees, L., Realo, A., & Allik, J. (2013). The relationship between life satisfaction and emotional experience in 21 European countries. *Journal of Cross-Cultural Psychology*, *44*(2), 223–244.

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). West Yorkshire: Emerald Group.

Martin, P., & Lynn, P. (2011). *The effects of mixed mode on simple and complex analyses.* Centre for Comparative Social Surveys Working Paper Series, Paper no. 04.

Meuleman, B., & Schlüter, E. (2018). Explaining cross-national measurement inequivalence. In E. Davido, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: Methods and applications* (2nd ed., pp. 363–390). New York: Routledge.

Miller, K. (2019). Conducting cognitive interviewing studies to examine survey question comparability. In T. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 203–226). Hoboken, NJ.: John Wiley & Sons.

Pennell, B.-E., Cibelli Hibben, K. L., Lyberg, L., Mohler, P. P., & Worku, G. (2017). A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts. In P. Biemer, E. D. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, … B. West (Eds.), *Total survey error in practice.* New York: Wiley.

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

Roberts, C. (2016). Response styles in surveys: Understanding their causes and mitigating their impact on data quality. In C. Wolf, D. Joye, T. Smith, & Y.-C. Fu (Eds.), *The Sage handbook of survey methodology* (pp. 579–596). London: Sage Publications.

Roberts, C., Eva, G., Lynn, P., & Johnson, J. (2010). *Measuring the effect of interview length on response propensity and response quality in a telephone survey—final report of the ESS CATI experiment.* ESSi JRA1 Deliverable 5. London.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Saris, W., & Gallhofer, I. (2014). *Design, evaluation and analysis of questionnaires for survey research.* (2nd ed.). New York: Wiley.

Sarrasin, O., Green, E. T., Berchtold, A., & Davidov, E. (2012). *Measurement equivalence across subnational groups: An analysis of the conception of nationhood in Switzerland.*

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(8), 23–74.

Smith, T. (2011). Refining the total survey error perspective. *International Journal of Public Opinion Research*, *23*(4), 464–484.

Soons, J. P., & Kalmijn, M. (2009). Is marriage more than cohabitation? Well-being differences in 30 European countries. *Journal of Marriage and Family*, *71*(5), 1141–1157.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90.

Survey Research Center. (2016). *Guidelines for best practice in cross-cultural surveys.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Tourangeau, R. (2017). Mixing modes: Tradeoffs among coverage, nonresponse, and measurement error. In P. Biemer, E. D. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, … B. West (Eds.), *Total survey error in practice: Improving quality in the era of big data* (pp. 115–132). Hoboken: Wiley.

Van De Vijver, F. (1998). Towards a theory of bias and equivalence. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence. ZUMA-Nachrichten Spezial, 3.* (pp. 41–65). Mannheim: Zentrum für Umfragen, Methoden und Analysen.

Van De Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135.

Van der Vijver, F. (2018). Capturing bias in structural equation modeling. In E. Davido, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), (2nd ed., pp. 3–43). New York: Routledge.

Van Herk, H., Poortinga, Y., & Verhallen, T. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360.

Van Vaerenbergh, Y., & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Villar, A. (2009). *Agreement answer scale design for multilingual surveys: Effects of translation-related changes in verbal labels on response styles and response distributions.* Survey Research and Methodology Program (SRAM)-Dissertations & Theses, 3.

Yao, E., Lim, J., Tamaki, K., Ishii, R., Kim, K. O., & O'Mahony, M. (2007). Structured and unstructured 9-point hedonic scales: A cross cultural study with American, Japanese and Korean consumers. *Journal of Sensory Studies*, *18*(2), 115–139.

Zavala-Rojas, D., & Saris, W. E. (2018). Measurement invariance in multilingual survey research: The role of the language of the questionnaire. *Social indicators research*, *140*(2), 485–510.

Appendix A

Table

Table A1 follows on next page.

Table A1
Descriptive statistics for scale items

Measure		France		French-speaking Switzerland				German-speaking Switzerland				Germany	
		CAPI		CAPI		CATI		CAPI		CAPI		CATI	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Evaluative Wellbeing													
Stflife (B24):	Overall life satisfaction: How satisfied with your life as a whole (0 Extremely dissatisfied – 10 Extremely satisfied)	6.51	2.35	7.72	1.90	7.79	2.0	8.25	1.50	6.94	2.11	7.16	2.17
Stflfsf (E31)	Life satisfaction so far: How satisfied with how your life has turned out (0 Extremely dissatisfied – 10 Extremely satisfied)	7.08	1.72	7.48	1.71	7.91	1.46	7.90	1.44	7.05	1.87	7.41	1.74
Happy (C1)	Happiness: How happy would you say you are (0 Extremely unhappy - 10 Extremely happy)	7.31	1.65	8.05	1.50	8.29	1.55	8.16	1.37	7.16	1.86	7.56	1.81
Emotional Wellbeing (Negative Affect)													
	How much of the time during the past week…(1 None or almost none of the time – 4 All or almost all of the time)												
Fltdpr (E8)	Depressed: …you felt depressed?	1.50	0.70	1.43	0.58	1.45	0.57	1.41	0.60	1.50	0.67	1.47	0.70
Fltlnl (E12)	Lonely: …you felt lonely?	1.44	0.74	1.34	0.59	1.34	0.62	1.21	0.48	1.33	0.60	1.28	0.60
Fltsd (E14)	Sad: …you felt sad?	1.55	0.68	1.53	0.62	1.53	0.61	1.42	0.57	1.43	0.60	1.45	0.64
Fltanx (E17)	Anxious: …you felt anxious?	1.76	0.76	1.76	0.64	1.70	0.80	1.72	0.65	1.21	0.46	1.20	0.49

Selection probability weighted.

Appendix B

*

Questions

1. Table 1 re-designed. Please confirm acceptance or advise
2. Table 1 re-designed. Please confirm acceptance or advise. If acceptable please provide the missing information (see questions marks)
3. please provide bibliographic information about Batchelor & Maio 2016