Comparability of survey measurements

Daniel L. Oberski
Joint Program in Survey Methodology
University of Maryland
Email: doberski@umd.edu

Version: 2011-12-02

Comparative surveys nowadays provide a wealth of survey data on a diverse range of topics covering most countries in the world. The online companion[1] to the ``SAGE handbook of public opinion research'', for example (Donsbach & Traugott, 2008), lists some 65 cross-national comparative social surveys that have been conducted around the world since 1948. Besides these general social surveys, many surveys on specific topics such as education, old age and retirement, health, working conditions, and literacy, to name just a few, are carried out continually.

Surveys may be conducted for different purposes. One purpose is estimation of population means, totals, and marginal distributions; another is the estimation of relationships between variables. Van de Vijver & Leung (1997) called studies with these goals respectively “level” and “structure” oriented. A comparative survey will then have as its goal to *compare* such level and/or structure parameters.

However, it is well-known that even estimates from surveys carried out with the utmost care and attention to quality will contain some amount of survey errors (Groves, 2004). A simple division can be drawn between *errors due to the selection of sample units*, and *errors due to the measurement instrument*. These error sources may have an effect on the estimates in the form of both bias and variance.

In comparative surveys, then, the estimates to be compared may each be influenced by survey errors, leading to the possibility that the comparison to be made is invalidated. That is, the estimates from different surveys might not be *comparable*.

The problem of comparability does not only occur when comparing the results of different surveys, but will also apply to the comparison of subpopulations in the same survey.

This chapter discusses the problem of comparability from the point of view of total survey error. The problem is illustrated with examples, and comparability issues related to errors due to both the selection of sample units and the measurement instrument are discussed. The discussion is, of necessity, limited and brief. For further details the reader is referred to the literature at the end of the chapter.

[1] http://www.gesis.org/en/services/data/portals-links/comparative-survey-projects/

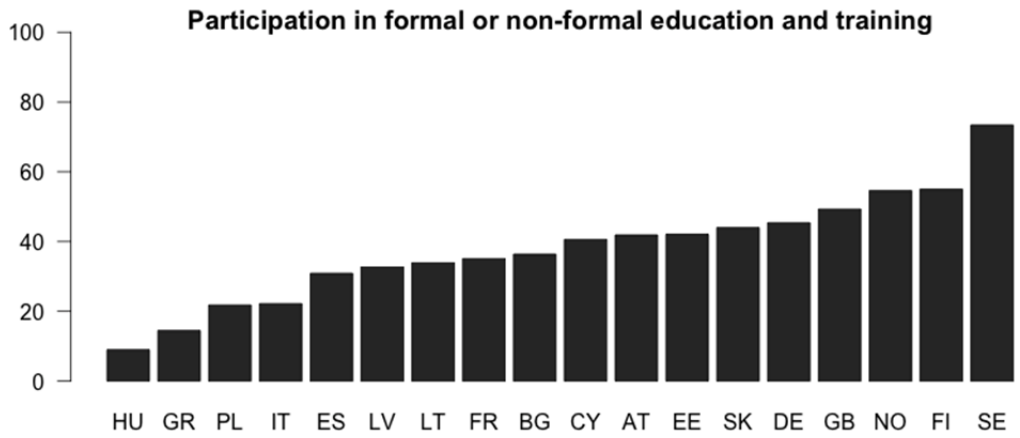

Figure 1. Comparative statistics by country on life-long learning in Europe. Shown are percentages of adult population participating in education, from Eurostat's 2007 Adult Education Survey. Source: Boateng (2009, table 1).

Figure 1 shows a typical application of comparative surveys: groups, in this case European countries, are compared on their means. Figure 1 compares the percentage of adults in each country that indicate participating in some type of educational or training activity, as estimated by Eurostat's 2007 Adult Education Survey.

Boateng (2009: 2) presented these data, remarking that "Total rates of participation vary between countries and the data shows the Nordic countries and the UK having high rates of participation. Low rates of participation are found in Hungary, Greece, Poland, and Italy." (p. 3). This certainly appears to be the case when looking at Figure 1. However, how might the estimated means shown there have been different if nonresponse were much higher in Greece and Hungary than it was in Finland and Sweden? Would the comparison still yield the same conclusion? Suppose nonresponse in Greece were predominantly due to noncontacts – for example because those in training programs are more difficult to contact – while nonresponse in Finland is more related to refusals of the lower educated?

It is clear, then, that although the differences shown in Figure 1 are statistically significant and appear to confirm a priori expectations one might have on differences between Nordic and other countries, there are possibly other *non-substantive* explanations for the findings shown. Nonresponse is only one of those factors that form possible alternative explanations for differences: among others are differences in frame errors, translation errors, differences in understanding of the concept of "job training" across countries, differences in interviewer training, differences in the printed answer scales, etc. In effect, *any difference in systematic survey error across countries potentially threatens the comparison*.

Whether these issues truly affect the comparison is not known, and it is by no means the intention here to suggest the Adult Education Survey provides incomparable numbers. The problem of comparability will arise on any occasion where groups are being compared on a survey statistic.

Besides means, measures of relationship between variables are also of central interest. Boateng (2009: 3), for example, mentions "varying gender differences [in educational participation]" across countries. Public health researchers have studied social class differences in self-rated health. Differences across countries in these health inequalities were studied by von dem Knesebeck, Verde, & Dragano (2006).

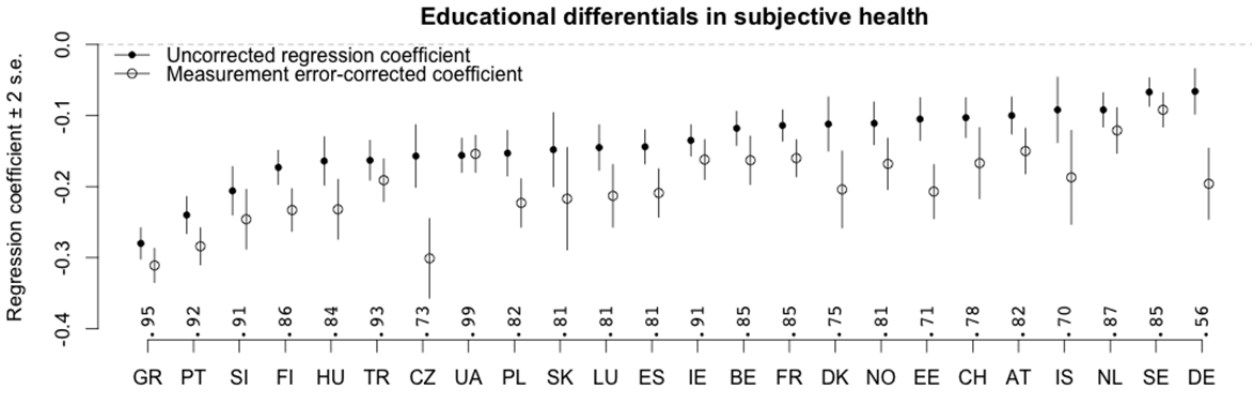

Figure 2. The estimated difference in self-rated health per education level by different countries (closed circles). Shown also is the same difference corrected for measurement error in the education variable (open circles). The estimated reliabilities used to correct the regression coefficients in each country are printed at the bottom.

Figure 2 shows country rankings on these health inequalities over educational levels. To estimate inequality and simplify comparison, the slope of a linear regression of self-rated health on education levels in each country is shown[2], based on data from the European Social Survey round 4. The size of the coefficients are shown as filled circles with two standard error intervals. Lower (more negative) values indicate larger inequalities in health across levels of education. It can be seen that Germany is by these estimates the most equitable, while Greece has the most inequality. Policy makers might be tempted to emulate Germany's health policies as a shining example of health equality in Europe.

Similarly to the comparative analysis of the proportion of life-long-learners discussed above, one may wonder whether these differences are due to a substantive reason such as countries' health policies, or whether there might be some other possible explanation for differences. Again, differential nonresponse might play a role. However, in the case of relationship parameters that would imply a high-order interaction among country, nonresponse bias, and the two variables of interest (Groves & Couper 1989). This explanation is still possible, then, but less plausible than it was for the comparison of means.

Measurement error, on the other hand, is well-known to have a direct and strong effect on simple regression coefficients (e.g. Fuller 1987). These are attenuated downward by unreliability in the independent variable, in this case education. Presumably, if there are strong differences between countries in the reliability of education level, the rank ordering of the countries may change accordingly.

Level of education is an objective variable, but it is not free of measurement error. Ganzeboom & Schröder (2009) found reliabilities of level of education between 0.7 and 0.9 (p. 9). The bottom of Figure 2 displays the estimated reliability of the level of education variable, showing that reliability varies considerably across countries (see also Oberski et al 2010). For example, the reliability is estimated as close to 1 in the Ukraine but is very low in Germany (only 0.56).

Unreliability can have a strong biasing effect on the regression coefficient. Having obtained these reliability estimates, one can then correct the regression coefficients used to rank the countries for

[2] For a complete description of the study design and the original questionnaires, please see http://ess.nsd.uib.no/

measurement error, obtaining corrected coefficients. These are shown in Figure 2 as open circles, again with two standard error intervals. Particularly the rank of countries where educational level is estimated with lower reliability is affected. Germany, for instance, moves from first place to the middle, while Greece is relieved from its role as the most unequal country in Figure 2 by the Czech Republic. Correction for measurement error dramatically affects the conclusions drawn from the comparison of these relationships.

The above examples demonstrate that comparability of survey measurements is an issue that must be carefully considered before substantive conclusions may be drawn from comparative surveys. The problems are not limited to the comparison of means but extend to the comparison of relationships.

Nor is the issue of comparability limited to "subjective" variables; comparisons across groups of an objective variable such as whether a person is in a training program may also be threatened by issues of comparability.

Another example that clarifies this is the comparability across time of the US Bureau of Labor Statistics' Current Population Survey, or the National Election Study which have switched from face-to-face to telephone to mixed mode data collection including web surveys over the years (Cohany, Polivka, & Rothgeb, 1994; Malhotra & Krosnick, 2007). Survey mode has been shown to affect estimates and therefore it is a question to what extent the time series can be compared. This also demonstrates that the issue of comparability is not limited to cross-country comparisons, but may also extend to comparison over time, different social groups, and so forth.

Two issues are of prime concern. First, what are possible sources of incomparability? This is discussed in the following section in the framework of Total Survey Error (TSE) (Groves, 2004). Second, in what way should survey methodologists and users consider comparability? This question is addressed in the remained of the chapter.

Incomparability as differential Total Survey Error

Incomparability of statistics across groups arises from systematic error or "bias". A framework for describing such errors in surveys is that of Total Survey Error (Groves 1989), shown in Figure 3. The figure shows the survey process and sources of errors that may arise during this process. The errors may be random or systematic. Each of these errors is described in detail elsewhere in the present volume.

Particular to comparative surveys is the existence of at least one other survey statistics, with which the statistic shown in Figure 3 is to be compared. This statistic is subject to different survey errors arising from the sources shown there. If the overall statistics are differently affected, this will cause "bias" in the comparison. That is, besides the substantive differences of interest, the comparison will also be partly affected by differences in systematic survey errors.

At this high level of abstraction, the problem is simple: differences in systematic total survey error. And so is the solution: systematic total survey error must either be reduced to zero or kept equal or close to equal across groups. This observation is the basis for the field of comparability of survey statistics.

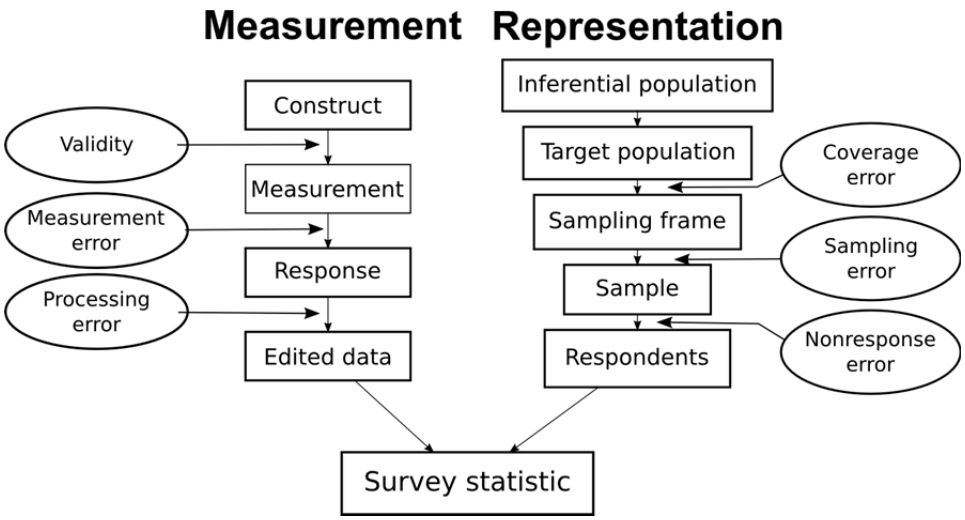

Figure 3. Total Survey Error framework, adapted from Groves (1989).

More specifically, it can be seen that Figure 3 has a "representation" and a "measurement" side. These correspond respectively to the goals of obtaining a sample statistic that is close to the population value, and measuring a quantity that corresponds to that intended by the survey researcher.

Systematic representation errors involve undercoverage, overcoverage, and nonresponse bias. These errors are so mathematically similar to each other that will be discussed jointly in the following section. Systematic errors in measurement arise from invalidity, processing, and measurement error. Measurement error is generally understood to include response error, interviewer-induced response effects, social desirability, method effect, response styles, and random error (unreliability).

It should be noted that errors that are random with respect to one statistic, usually the mean or total, are systematic with respect to another. For example, purely random measurement error will not affect means, but it will bias correlations and regression coefficients. Conversely, nonresponse bias may, in a particular instance, affect means but not correlations. Of importance for comparability are only those errors that may cause systematic deviations. From this point of view, sampling error, for example, is not relevant, because it is standard practice to account for sampling error in any statistical analysis.

Systematic errors due to survey mode have received quite a bit attention in the literature (Dillman, Smyth, & Christian, 2008). In the Total Survey Error framework adopted here, a survey mode is a combination of different aspects such as wording, design, interviewers, coverage, nonresponse propensities, etc. Differences in mode are likely to increase potential incomparability since many sources of error are changed at the same time. However, we do not consider it separately from its constituent errors here.

In the field of cross-cultural psychology, a different classification of biases was given by (Van de Vijver & Leung, 1997). These authors distinguished "construct", "item", and "method" bias. In the TSE framework, construct bias corresponds to differences in invalidity (top left of Figure 3), and item bias corresponds to differences in measurement error. "Method" bias, in their formulation, is essentially the rest of Figure 3.

Comparability

Making comparisons possible starts by "meeting the coming disease" of bias and preventing it. This involves such actions as putting in place adequate translation and adaptation procedures for cross-cultural studies (Harkness, 2003) ensuring uniformity of fieldwork arrangements (Jowell, 2007), and defining concepts that will be similarly understood in different groups (Hui & Triandis, 1985). Such actions are vital to the goal of bringing the ideal of comparability closer. However, they do not guarantee that this ideal is attained. Just as even the best survey has errors, even the most carefully designed, implemented, and monitored comparative survey will have some degree of bias in its comparisons; there will always be some possibility of incomparability.

There are four basic possible approaches to this problem.

First, the researcher may simply ignore the problem. This amounts to an implicit or explicit assumption that any bias in the comparison will be small. That is, that any biases will not be so large as to "explain away" the conclusions drawn. Such an assumption can be warranted if prior research suggests it. In other cases it will amount to a hope.

In psychometrics, comparability is formulated in terms of "item bias": the difference, *for equal values of the true score*, in expected value of the variable given the comparison group (Mellenbergh, 1989). This idea is the basis for the rich literature on differential item functioning or "DIF" (Holland & Wainer, 1993).

The second possibility is then that, by some means, the hypothesis is tested that some form of bias will occur. Item bias is usually associated with the comparison of parameters of item response theory (IRT) models over groups, to which we will return later in the chapter. This includes the approach taken in the invariance testing literature (Meredith, 1993; Steenkamp & Baumgartner, 1998; Millsap et al., 2007) discussed later in this chapter. If the test indicates that the measures are not comparable across group, the researcher will then avoid the comparison in question altogether. This approach corresponds to what will here be termed a "strong" interpretation of comparability: that any possibility of bias (in the population) is sufficient to invalidate comparisons.

Another interpretation of the idea of comparability is that a comparison is warranted if it is very unlikely that the bias due to differential errors will be so large that it can change conclusions. It might be called the "weak" interpretation of comparability. Weak comparability is implicitly assumed if the researcher ignores the issue of comparability. However, given certain research designs, it may actually be investigated. This, then, is the third possible approach to the problem of incomparability: to perform analyses that make plausible the assumption that any biases will not be so large as to "explain away" the conclusions drawn. Parts of the rest of this chapter will give particular attention to this approach.

A fourth and final approach is to attempt to directly estimate the survey errors, so they may be corrected for. Any differences across groups can then no longer threaten the conclusions drawn from comparisons. This approach was taken in the example of the educational differential in subjective health above: since it was known that differential measurement error in the education variable over countries could invalidate comparisons, each regression coefficient was corrected for the estimated reliability

before the final ranking was made. The correction approach based on models that estimate item bias is known as “equating” in the psychometrics literature (Holland, 1982).

In psychophysics, a notion similar to “item bias” is termed “differences in response function” and has been the subject of extensive research in that field since its inception (Stevens, 1975). Psychophysics is concerned with the relationship between a stimulus and a person’s perception of that stimulus. Classic examples are hues of light, highness of musical notes, and creaminess of butter. Many psychophysical experiments used response scales with arbitrary reference points, such as squeezing a ball, producing a louder or softer sound, or freely giving a number (with no restrictions) to allow the subject to indicate their perception of the stimulus.

The interpersonal differences in scale precluded direct comparison between people of the obtained scores, since, for instance, older people will squeeze less, and whether one decides that 100 or 1000 means “very creamy” is entirely up to the respondent. Classical psychophics therefore developed the notion of the response function, and methods to estimate it for individual respondents so as to correct for incomparability (Stevens, 1975). The method was applied to survey research by Saris (1988), and also forms the basis for the “anchoring vignettes” approach developed by King et al. (2004).

The last three approaches – strong comparability, weak comparability, and correction of statistics – all require some form of modeling. In what follows different models that can be used for these purposes will be discussed and demonstrated. The first concern, however, should always be to reduce the possibility of incomparability as much as possible.

Prevention of incomparability

When feasible, incomparability should be prevented as much as possible by reducing total survey error, and by making errors likely to be similar across groups. Sources of survey error and suggested methods to reduce them are discussed elsewhere in this volume. Here a short description of two methods useful for developing comparable questions are discussed: translation and quality control by coding of question characteristics. Another method is cognitive interviewing (Willis, 2005).

Translation or adaptation

One particular source of incomparability specific to comparative studies is that of translation or adaptation. Translation/adaptation usually refers to questions being asked in different languages, as is the case in cross-national research. But it may also refer to the comparison of other groups: questions designed for adults are not usually directly applicable to children, for instance.

Harkness (2003) gave an overview of different translation procedures intended to yield comparable questions, describing particularly the Translation, Review, Adjucation, Pretesting and Documentation (TRAPD) procedure developed for large cross-national surveys such as SHARE and ESS. For more information we refer to Harkness et al. (2010), where many other issues relating to cross-national comparative surveys are also discussed.

The difficulties inherent in translation or adaptation procedures can be demonstrated with the following item from the ESS, cited by Zucha in Hoffmeyer-Zlotnik & Harkness (2005):

> *"Please tell me how important each of these things should be in deciding whether someone born, brought up, and living outside [country] should be able to come and live here. Firstly, how important should it be for them to… be wealthy?"*

The stimulus "be wealthy" was translated in Italian as "…avere una buona salute", which means "…be healthy". Clearly this is simply a mistake in the process; it does go to show, however, that even with extremely careful procedures in place, mistakes are still made.

In French and Spanish "be wealthy" was translated as "être riche"/"ser rico", which means "be *rich*". These languages do not have a separate word for "wealthy", but only the word "rich", which has a related but different meaning. The word "wealthy" cannot be translated into these languages as closely as "rich" could be.

This example demonstrates that translation is not just an issue of finding the corresponding word in the other language. It entails translating the *concept intended* as closely as possible (Harkness, 2003). How close the translation is able to get to the intended meaning, in turn, depends on the question itself. In the example, if the original question writers had written "rich" instead of "wealthy", there would have been no conceptual difference for these languages[3].

Another example of this phenomenon is the word "fair", used in social trust scales: "people try to be fair". There appears to be no very close translation of this concept in many European languages. As solutions, "honest" (Dutch, Spanish, Swedish) and "behaving correctly" (French) are found in the ESS. Both come close to the intended concept but are not as close as they would have been if the original questions had used the word "honest".

Adaptation issues also arise when the mode of survey administration is different for the groups to be compared. For example, questions with many categories in self-administered modes when asked verbally are often broken into multiple steps. This may have an effect on the responses. There is no simple solution to this problem, and researchers may have to weigh the possibility of incomparability against the information loss incurred if the number of categories were reduced.

[3] There might of course still be difficulty in translating into other languages. Source questions formulated in the English language, which is often claimed to have more words than any other natural language, would appear to be particularly prone to this type of problem.

Table 1. Unwarranted differences in two survey question's design characteristics across three countries in the ESS.

	Item 1	Item 2
Country B	• Missing scale definition • Missing respondent instructions "please use this card" • Showcards have different layout	• Missing question introduction • Missing respondent instructions "please use this card" • Showcards have different layout
Country C	• Missing respondent instructions "please use this card" • Series of separate questions changed to battery of questions • Showcards had numbers before the categories, while source had no numbers	• Missing respondent instructions "please use this card" • Series of separate questions changed to battery of questions • Showcards had numbers before the categories, while source had no numbers
Country D	• Missing question introduction • Showcards had numbers before the categories, while source had no numbers • Showcards had boxes around the categories, while source had no boxes • Showcards have vertical instead of horizontal layout	• Showcards had numbers before the categories, while source had no numbers • Showcards had boxes around the categories, while source had no boxes • Showcards have vertical instead of horizontal layout

Question comparability quality control using a question coding system

An additional check on question comparability may be provided by comparing different characteristics of the questions in the original and adapted versions. One such coding system (SQP) will be discussed here. Another coding system, QAS, was developed by Lessler & Forsyth (1996).

Zavala (2011) gives the results of coding question characteristics with the SQP coding system (Oberski, Gruner, & Saris, 2011; see also Saris & Gallhofer, 2007) and comparing the results across translations in the European Social Survey round 5. Table 1 shows the differences found over three countries for two items. The countries and items have been relabeled, since this table is given merely as a demonstration of the type of problem that can be detected with this method.

Table 1 shows that problems that can be detected by this procedure are not semantic in nature; that type of equivalence must be established by another method, such as the translation procedure or cognitive interviewing. Rather, the differences are on aspects of the question design that are known from the survey methodology literature to affect question quality and response patterns (Alwin, 2007; Saris & Gallhofer, 2007). Common are differences in the visual layout of the response scales, and the omission of interviewer or respondent instructions. Since the layout of response options affects the response (Christian & Dillman, 2004) such differences form a threat to equivalence of the question and were corrected before the questionnaires were fielded.

Rigorous translation and pretest procedures such as the ones described here can help to reduce incomparability. However, even when such procedures have been followed it is still possible that the questions are not comparable. For this reason, models were developed that allow for the estimation and correction of the degree of incomparability across groups.

The representation side

The "representation" sources of total survey error shown in Figure 3 arise from the process that selects respondents into the sample. As will be shown below, if this selection process produces bias that differs over the groups being compared, a bias in the comparative statistic may occur. Since the literature on this topic focuses strongly on bias due to nonresponse, this section will follow that convention and discuss bias on the representation side in terms of nonresponse error. However, it may be kept in mind that the same arguments will apply to coverage error and other possibly differential selection processes such as self-selection into convenience samples, self-selected internet panels, and so forth.

Incomparability due to nonresponse

Nonresponse bias is well-known to affect survey statistics (Groves & Couper, 1998), and due to rising nonresponse rates, has become a prime concern among survey methodologists. In this section we develop the effect of nonresponse bias on the *comparison* of survey statistics. If this comparison is affected by nonresponse bias, comparability may be threatened.

Researchers may be interested in the comparison of means or totals, but also of odds ratios, correlations, regression coefficients, etc. across groups. Interest will typically focus on the difference between the groups, $T^{(1)} - T^{(2)}$ in some statistic. In each group, there may be nonresponse bias, so that the statistic observed by using only respondent data in each group, say $T_R^{(g)}$, equals $T_R^{(g)} = T^{(g)} + b^{(g)}$, where $b^{(g)}$ is the nonresponse bias in group *g*.

Differential nonresponse may make the groups incomparable, in the sense that differences between the groups might be due to differences in nonresponse bias between the groups. This can readily be seen by expressing the differences between groups as

$$T_R^{(1)} - T_R^{(2)} = T^{(1)} - T^{(2)} + \left(b^{(1)} - b^{(2)}\right)$$

= (*true difference*) + (*difference in bias*).

It can be seen that the threat to comparability is the difference in nonresponse bias between the groups (Groves & Couper, 1998, pp. 8–9). This difference in nonresponse bias, in turn, depends on the differences $T_R^{(g)} - T_N^{(g)}$ between respondents and nonrespondents with respect to the statistic, and the response rates in the groups:

$$b^{(1)} - b^{(2)} = \left(\bar{\rho}^{(1)} - 1\right)\left(T_N^{(1)} - T_R^{(1)}\right) + \left(\bar{\rho}^{(2)} - 1\right)\left(T_R^{(2)} - T_N^{(2)}\right),$$

where $\bar{\rho}^{(g)}$ is the response rate in group *g*. The bias in the difference will be zero if the response rates in both groups are 1. If the response rates are not 1 but equal, there may still be bias if the differences between respondents and nonrespondents are not equal across groups. Conversely, even if these differences between respondents and nonrespondents are equal, there may still be bias if the response rates are not equal. There will be zero bias only if the differences between respondents and

nonrespondents are equal *and* the response rates are also equal. Strong bias can be obtained if the biases are opposite to each other and the response rates are substantial. It can also be shown that higher response rates, similar response rates, lower nonresponse biases, and similar nonresponse biases will tend to lower the relative bias of the difference.

The problem may be more pronounced when comparing means than when comparing correlations, regression coefficients, and other relationship parameters. Goudy (1976) and Voogt (2004) argued the necessary assumptions to be more plausible when comparing (logistic) regression coefficients than when comparing means. It is also a common assumption in much of experimental psychology that nonresponse bias will tend not affect estimates of relationships. Whether this is actually true or not remains mostly a matter of assumption (Groves & Peytcheva, 2008, p. 182).

It is clear, then, that unless one is willing to make strong assumptions, nonresponse bias will generally cause groups to be incomparable in the strong sense. This will be especially true when comparing the results obtained from different surveys, such as is usually the case in cross-national research.

Perhaps, however, in a particular study, the biases due to nonresponse are not so strong so as to be able to "explain away" the found differences. One method of examining this possibility of weak comparability is to attempt to estimate nonresponse bias directly (Groves, 2006; Stoop, Billiet, & Koch, 2010). Another is to attempt to work out the *maximum possible bias* in the difference. Recent work by Schouten, Cobben, & Bethlehem (2009) has provided results that allow for estimation of a maximum bias in the comparison of means. This maximum bias can then be used to see whether such a bias could in fact form an alternative explanation for the observed difference.

Maximal absolute bias and the R-indicator

Survey response can be viewed as a stochastic process (Bethlehem, 1988; Groves & Couper, 1998: 11-2) with every potential respondent having a probability of participation, denoted as ρ, sometimes called the response propensity score. The decision to participate or not is determined by a person's ρ, with 1 being a certain respondent and 0 a certain nonrespondent. The response rate is then just the average, denoted $\bar{\rho}$, of the response propensities.

Schouten et al. (2009) developed an indicator of the representativeness of a survey based on the response propensities. They termed this quantity the "R-indicator". The R-indicator, $R(\rho)$, has a rather simple definition: it is defined as $R(\rho) = 1 - 2\sqrt{\mathrm{var}(\rho)}$. The R-indicator will lie in the (0, 1) interval, 0 meaning maximum unrepresentativeness, and 1 meaning perfect representativeness. Note that the R-indicator does not depend on the statistic or the target variable of interest, but only on the variance of the propensity scores among sampled units.

If an estimate of the propensity score can be obtained from auxiliary data, the R-indicator can then be estimated simply via calculating the variance of the estimated propensity scores (Shlomo, Skinner, Schouten, Bethlehem, & Zhang, 2008). Estimates of the R-indicator do assume, however, that response propensity is not related to the outcome statistic of interest after conditioning on the auxiliary variables.

That is, that the nonrespondents are Missing at Random (MAR) given the auxiliary variables. The disadvantage of this approach is therefore that its success depends wholly on the selection of adequate auxiliary variables.

By intuition, it seems reasonable that nonresponse bias, and therefore bias in the difference between groups, would depend on the R-indicator. And indeed this can be shown to be the case. In a single survey, the *maximum absolute bias* in a mean will be bounded by $|b| \leq \frac{[1-R(\rho)]\sigma_y}{2\,\bar{\rho}}$, where σ_y is the standard deviation of the target variable. Thus, the more unrepresentative the survey is, and the lower the response rate, the stronger the bias can potentially be.

That maximum bias depends on the R-indicator and response rate is an incredibly useful result, since the maximum absolute bias gives an upper bound on the amount of damage nonresponse can do to the estimate. Given an estimate of the R-indicator, it can be investigated, for instance, whether an obtained mean or total still differs significantly from some predetermined value, even after taking the maximum bias into account. If it does, then nonresponse bias becomes less plausible as a possible alternative explanation for observed differences.

The concept of maximum bias is here straightforwardly extended to the comparison of groups. When a difference in means is of interest, the maximum bias of the difference is bounded by

$$\left|b^{(1)} - b^{(2)}\right| \leq \frac{1}{2}\left(\frac{\sigma_y^{(1)}\left[1 - R\left(\rho^{(1)}\right)\right]}{\bar{\rho}^{(1)}} + \frac{\sigma_y^{(2)}\left[1 - R\left(\rho^{(2)}\right)\right]}{\bar{\rho}^{(2)}}\right).$$

This worst case will occur only when the nonresponse biases in the two groups are maximal *and in opposite directions*. When the biases can be assumed to be in the same direction, the maximum bias of the difference will be bounded by $\max\left(\frac{\sigma_y^{(1)}[1-R(\rho^{(1)})]}{2\bar{\rho}^{(1)}}, \frac{\sigma_y^{(2)}[1-R(\rho^{(2)})]}{2\bar{\rho}^{(2)}}\right)$.

The above is merely a reformulation of the theoretical results given earlier. Its advantage, however, is that, under the assumption of MAR given the covariates, it does not depend on unknown quantities: it can be estimated in a given study. In a particular study where two groups are compared on their means, weak comparability can be made more plausible if good estimates of the R-indicators can be obtained. An illustration of this approach is given below.

Table 2 gives the sample sizes, response rates, and estimated R-indicators for the Belgian and Norwegian samples of the European Social Survey (ESS), Round 3 (fielded in 2006). It can be seen that both countries had relatively high and similar values for both response rate and R-indicator. The values are neither perfect nor equal, however, as is to be expected in any applied study.

Table 2. Sample sizes, response rates, and estimated representativity (R) indicators for the European Social Survey Round 3 in two countries.

	Sample size	Response rate	R-indicator (estimate)
ESS 2006 (Belgium)	2,927	61.4%	0.807
ESS 2006 (Norway)	2,673	65.6%	0.762

Source: Shlomo et al. (2008, p. 33).

According to the strict definition of comparability, the Belgian and Norwegian samples are not comparable, because they do not have exactly the same response rate, R-indicator values, and (presumably) nonresponse bias. As shown by the equations above, this *might* lead to apparent differences that do not reflect population differences.

This shows that the strict definition of comparability is not very helpful in practical research, since some differences in response rates and representativity will be rule rather than exception. The weak definition of comparability then compels us to ask whether such differences can plausibly explain the observed differences between the samples or not.

To demonstrate the application of this principle we compare Norway and Belgium on average reported happiness, and the proportion of people who say they use the internet "every day". These comparisons are shown in Table 3, together with the maximal biases assuming bias is in the same or in opposite directions. The corresponding "worst-case scenario" estimates of differences between Belgians and Norwegians are displayed in the last four rows.

Table 3. Analyses of differences between Norway and Belgium in happiness and internet use, and the same analyses correcting for maximum bias based on R-indicators and response rates in the two surveys.

	"How happy are you?" (0-10)	**Use internet "every day"**
ESS 2006 (Belgium)	7.66 (1.6)	0.346 (0.48)
ESS 2006 (Norway)	7.93 (1.5)	0. 493 (0.50)
Difference	-0.27* (0.06)	-0.147* (0.01)
Max. abs. bias of difference (opposite directions)	0.52	0.17
Max. bias-corrected estimate (opposite directions)	+0.25*	+0.03
Max. abs. bias of difference (same direction)	0.27	0.09
Max. bias-corrected estimate (same direction)	0.00	-0.06*

**p-value < 0.05. Source: ESS data (2006). http://ess.nsd.uib.no/ess/round3/*

Table 3 shows that in the worst case scenario, the possibility cannot be ruled out that the differences in response rate and lack of representativity are the cause of the apparent differences between Belgium and Norway. This will only occur when the biases are maximal and in opposite directions; for example when the imbalance is on gender *and* males are *more* likely to use the internet in Norway, but *less* likely to do so in Belgium. Other scenarios where this might occur are of course also possible.

For the cross-country comparison of happiness, applying a correction for the maximum potential bias due to differential nonresponse drives home the point that incomparability due to nonresponse cannot be ruled out as an explanation for differences. Differential nonresponse even has the potential to make it look as though the average Norwegian is happier than the average Belgian, when in reality the opposite might be true.

Looking at daily internet use, the differences between Belgium and Norway could be due to nonresponse if the biases are maximal and in opposite directions. However, if we can assume that they are in the same direction, then the maximal bias due to differential nonresponse is 0.09. In that case, even the worst-case bias adjustment will still yield the conclusion that Norwegians have a statistically significantly higher daily internet use than Belgians. So a possible alternative explanation for these differences in terms of incomparability due to differential nonresponse bias is less plausible in comparison of internet use.

The estimated R-indicator, when it can be applied, is a useful device for examining whether weak comparability with respect to nonresponse is plausible or not. It must be kept in mind that estimates of the R-indicator are only as good as the auxiliary variables that are used to estimate them. Furthermore, although some have contended that nonresponse is less likely to affect comparisons of relationships than it is to affect comparisons of means, work on comparing relationships across groups is, at the time of writing, still a field open for investigation.

In this section the theoretical effects of nonresponse on comparability were given. Through the R-indicator, it was possible to perform an analysis of weak comparability, under certain assumptions. The methods and results presented here were originally developed to deal with the issue of nonresponse, but can be equally well applied to assess the possible effects of over- or undercoverage or self-selection in nonrandom samples on comparability.

The measurement side

The previous section discussed the goal of obtaining a comparative sample statistic that is close to the "population" value. The second goal identified in Figure 3 is that of measuring a quantity that corresponds to that intended by the survey researcher, the "measurement" side of total survey error.

Tourangeau, Rips, & Rasinski (2000) gave an account of the process of survey response, based on a body of experimental findings in cognitive psychology. In their process model, the respondent goes through the stages of comprehension, retrieval, judgment, and response. Response behavior is then tied to these different processes. For example, consider the question (p. 38):

> *During the past 12 months, since January 1, 1987, how many times have you seen or talked to a doctor or assistant about your health? Do not count any time you might have seen a doctor while you were a patient in a hospital, but count all other times you actually saw or talked to a medical doctor of any kind.*

The structure of this question is very complex, containing many subordinate clauses and conditions. The complexity is in the concept being described, leading Tourangeau et al. (2000, p. 39) to remark that "processing this question into its underlying logical form is likely to impose quite a burden on working memory, one that may exceed the capacity of the respondent". They therefore argue that the step of comprehension may not correspond to that intended by the researcher.

Suppose that one wished to compare the number of doctor's visits made by people with high cognitive capacity to people with lower cognitive capacity – for example, intellectually disabled adolescents compared with their peers. Considering that the comprehension hurdle is likely to be cleared differently by these two groups, the response process is likely different as well. Thus, different answers to this question may be obtained for two people, one with a very low intelligence and the other with a high intelligence; *even if they visited the doctor the same number of times*. Such variation of responses for the same true score will be called "measurement error" here.

Differences in the response process across the groups to be compared may threaten comparability. Differences in comprehension may arise due to translation or cultural differences, a topic that has received some attention in the literature (Harkness, Vijver, & Johnson, 2003). Differences in the retrieval of information may occur due to differences in the salience and frequency of an event asked about (Tourangeau et al., 2000 chapters 3-4), and in opinion research due to differences in the attitude "strength" or "crystallization" (Petty & Krosnick, 1995), among other factors. An example of a cause of differences in judgment – the way in which retrieved information is combined – is differences in motivation; for example, if one group answers the question after a long list of filter questions and the other does not. Response mapping differences may occur if the meanings of the labels are not equal across countries (Harkness et al., 2003), not equally spaced (Oberski, Saris, & Hagenaars, 2008) or if there are differences in response style (Baumgartner & Steenkamp, 2001; Harzing, 2006). Finally, differences in the overall strategy of response may be related to differences in task difficulty, respondent ability, and motivation to perform the task (Krosnick, 1991).

Factor analysis and IRT model the process by which the respondent arrives at an answer, given the "true" variable of interest. This relationship between the underlying score and the respondent's final answer at is called the "response function". Invariance (equivalence) testing and DIF are the comparison of these response functions across the groups to be compared.

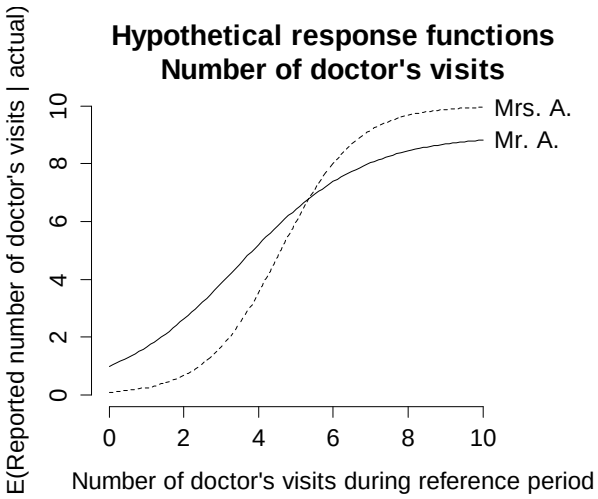

Figure 4. Hypothetical response functions to a question querying number of doctor's visits for two respondents.

For illustration, two hypothetical response functions are given in Figure 4. In a survey, both Mrs. and Mr. A. are asked how many times they visited the doctor in a certain period. The plot shows the expectation of the number reported given the actual number of visits for both respondents. Of course, this expectation cannot be observed in practice; the curves represent a series of counterfactuals, the answers respondents would give in expectation if they had visited the doctor a certain number of times.

Figure 4 shows that Mr. A, if he cannot recall any visit in the reference period will still report one; this might happen, for instance, if he incorrectly places a prior visit in the reference period (forward telescoping). If he has had close to 10 visits, he might be ashamed to admit it and reports fewer. The effect is that Mr. A overestimates the low and underestimates the high numbers. The reverse is the case for Mrs. A.: she underreports low numbers, claiming "it was nothing", while as the numbers get higher she tends toward overestimation. She might include events that happened after the reference period, for instance (backward telescoping).

The final result of the differences in response functions shown in Figure 4 is that, for example, if Mr. and Mrs. A. had both visited the doctor 10 times in the reference period, Mrs. A. will report 10 visits, while on average Mr. A. reports 8.8 visits. A researcher might be tempted to conclude that Mrs. A. is more ailed than her spouse. However, this conclusion would be incorrect. In general, *any conclusion resulting from the comparison of the reported number of visits of Mr. and Mrs. A can be potentially explained by the difference in their response functions.*

If Mr. and Mrs. A. are representative of their genders, for instance, a statistically significant difference in subclass means by gender of the number of doctor's visits, has a difference in response function as an possible alternative explanation. Invariance testing, differential item functioning, and anchoring vignettes are all methods of attempting to rule out this alternative explanation.

The logic behind these methods when applied to the comparison of means and relationship parameters is demonstrated by considering the comparison of a three-item "social trust" scale across countries in the ESS (e.g. Reeskens & Hooghe, 2008).

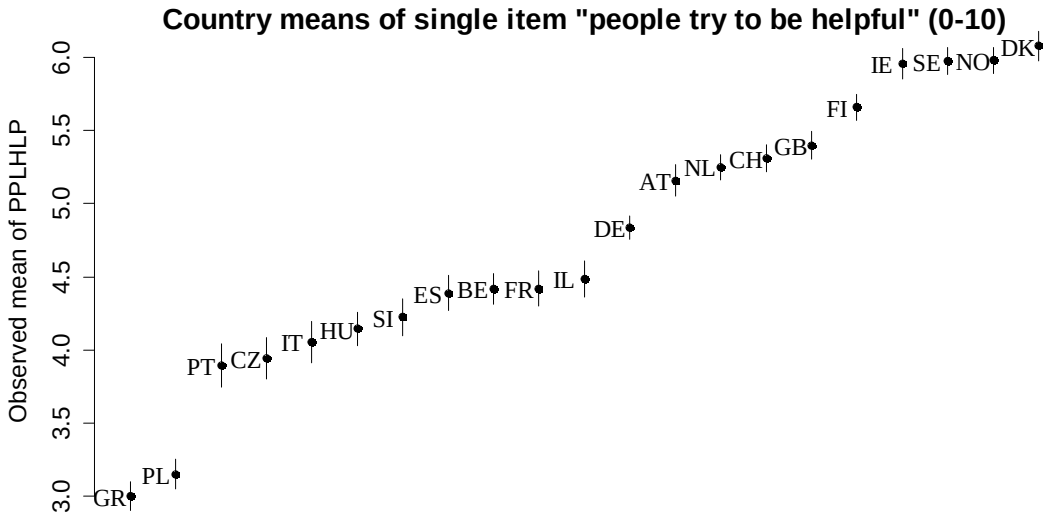

Figure 5. Mean of the item PPLHLP in different countries of round 1 of the ESS, with 2 s.e. confidence intervals.

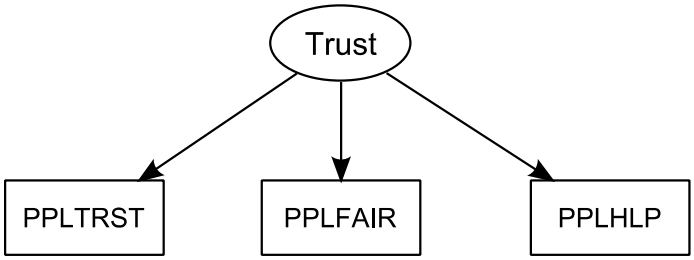

Figure 6. Latent variable model of the three observed indicators.

The three items assumed to measure "social trust" in this questionnaire are:

PPLTRST *Would you say that... most people can be trusted, or that you can't be too careful in dealing with people?* (0-10)
PPLFAIR *...most people would try to take advantage of you if they got the chance, or would they try to be fair?* (0-10)
PPLHLP *...most of the time people try to be helpful or that they are mostly looking out for themselves?* (0-10)

Figure 5 shows the country means on one of the three items, PPLHLP. Confidence intervals using the survey weights are also shown. It can be seen that Greece has the lowest mean on this item and Denmark the highest. Below we will illustrate comparisons by considering Greece (GR), Poland (PL), and the United Kingdom (GB). For example, the difference of 0.18 between the mean of the PPLHLP item in Greece and Poland is significant (p = 0.037). The differences with the United Kingdom are also clearly statistically significant. However, it is not clear whether such differences can be explained by variation in response functions over the countries.

There are many models that have as a goal the estimation of the response function. The most commonly used are the multiple group factor model (or MGSEM), and Item Response Theory (IRT) models. These models can only be estimated by assuming an underlying latent variable "Trust", for which the three items PPTRST, PPLFAIR, and PPLHLP form a scale. This model is shown as a graph in Figure 6. The graph shown there may represent different types of model: factor, IRT, and latent class models can all be applied. The key assumption is that of conditional independence of the items given the latent Trust variable (uncorrelatedness for factor models).

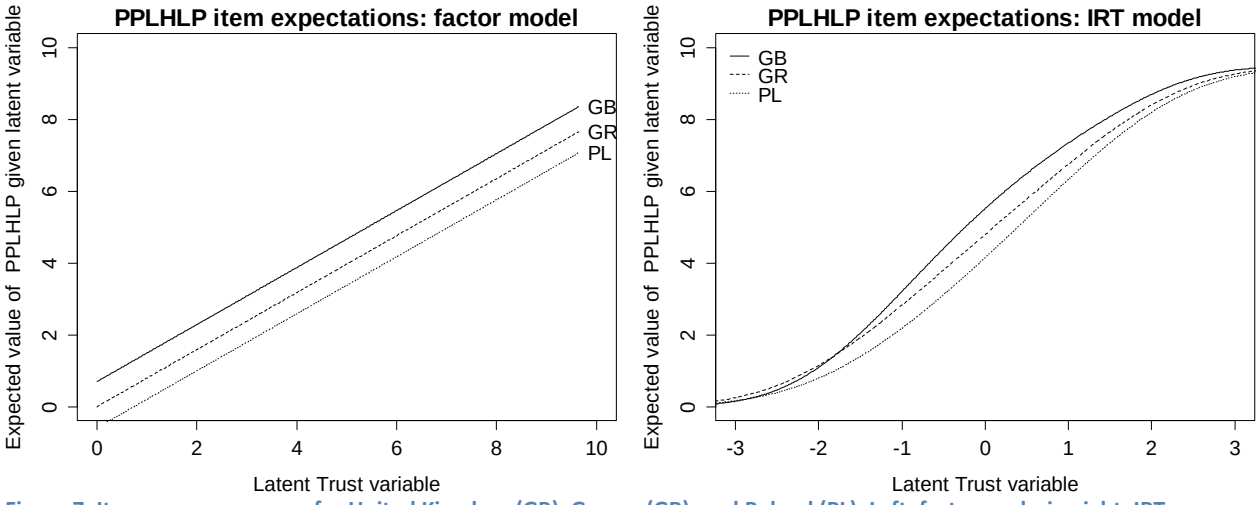

Figure 7. Item response curves for United Kingdom (GB), Greece (GR), and Poland (PL). Left: factor analysis; right: IRT.

A literature has developed around the aim of establishing whether the measures can be compared across groups: the procedure of so-called "invariance" testing (Meredith, 1993; Steenkamp & Baumgartner, 1998; Millsap et al., 2007; Davidov, Schmidt, & Billiet, 2010). Invariance testing aims to assess whether 1) the model shown in Figure 6 holds in all countries ("configural invariance"); 2) the slopes (loadings) are equal across countries ("metric invariance") and 3) the intercepts are equal in addition to the slopes ("scalar invariance"). Under the model assumptions, this amounts to evaluating whether response functions vary over countries. When scalar invariance holds, the groups are said to exhibit "full score equivalence"; when only metric invariance holds, "scale equivalence".

Reeskens & Hooghe (2008) performed such invariance tests on the social trust data presented here using the model shown in Figure 6. Employing multiple group structural equation models (linear factor analysis), they concluded that the three items are not scalar invariant across countries, but that there is metric invariance[4]. What this means is illustrated in the left-hand side of Figure 7.

Figure 7 plots the expected value of the PPLHLP item for different values of the theoretical latent factor "Trust". In the case of the linear factor models employed by Reeskens & Hooghe (2008) this is simply a matter of plotting *(intercept) + (loading) Trust.* This has been done for the United Kingdom (GB), Greece (GR), and Poland (PL). That metric invariance holds can be seen because the lines are parallel. The lack of scalar invariance is reflected in the distance between the response functions for the different countries.

Overall, Figure 7 shows that there is variation in estimated response functions across these three countries. In the factor model, a British person is expected to choose a value 0.7 higher than a Greek person, and a Greek person chooses values 0.6 above those chosen by Poles, even when their unobserved Trust scores are all equal. For instance, Figure 5 showed a statistically significant difference between Poland and Greece of 0.18. But this is much less than the item bias of 0.6 shown in Figure 7. Therefore the raw means plotted in Figure 5 should not be taken at face value: the lack of scalar invariance causes the countries to be incomparable on their raw means.

[4] A test of whether the factor model holds in each country is not possible in this case because with only three indicators the model without equality constraints has zero degrees of freedom.

The factor model commonly used in the invariance testing literature is not the only possible model to estimate response functions based on the graph in Figure 6. Other possibilities are latent class analysis (Kankaraš, Moors, & Vermunt, 2010), and IRT modeling with its theory of differential item functioning (Mellenbergh, 1989; Millsap & Yun-Tein, 2004). As an illustration of the differences and similarities of these approaches, the right hand side of Figure 7 provides estimated item response functions for the PPLHLP item based on an IRT model[5].

Figure 7 shows that the spaces between the IRT model response curves at the mean (zero) are approximately equal to the spaces between the parallel response curves produced by the factor model. Away from the mean, however, the IRT model predicts less score inequivalence than the factor model. In this model whether two scores from different countries on the PPLHLP item can be compared would depend on the score on the underlying scale.

Another point of interest in this comparison is that even though only the intercept parameters (thresholds) were allowed to vary over countries, mimicking the metric invariance found in the factor analysis, the lines in Figure 7 are not exactly parallel. This signals that lack of scalar invariance has more far-reaching consequences under the IRT model than under the linear factor analysis model.

It should be remarked that Figure 7 shows the lack of equivalence only for the single item PPLHLP. One might also compare means on the composite score of the three trust items together, as an estimate of the social trust construct. In this case the lack of invariance for one item does not automatically imply lack of invariance for the mean of the composite score (Byrne, Shavelson, & Muthén, 1989).

Scalar and metric invariance guarantee the comparability of means/totals and covariances, respectively. Regression coefficients, correlations, and other measures of bivariate association might still be incomparable, however. These quantities are affected by the *reliability* of the measures (Fuller, 1987). Under the true score model implied by factor analysis, the relationship between the true and observed regression coefficient can be expressed as

$$\beta_{YX} = \beta_{yx} \cdot \kappa_X,$$

where Y and X are the observed variables, y and x are their error-free counterparts, and κ_X is the reliability "ratio" (Fuller, 1987) of the observed independent variable.

If the reliability differs across groups, regression coefficients will be affected, since the observed coefficient equals the true coefficient multiplied by the reliability of the independent variable. A correlation or standardized regression coefficient is affected by the reliability of both variables. If reliabilities differ, observed regression coefficients and correlations will differ also even if the true difference across groups is zero.

[5] A two-parameter normal ogive model was estimated. Expected values were calculated by multiplying country-specific item characteristic curves with the scores 0-10 and summing over categories.

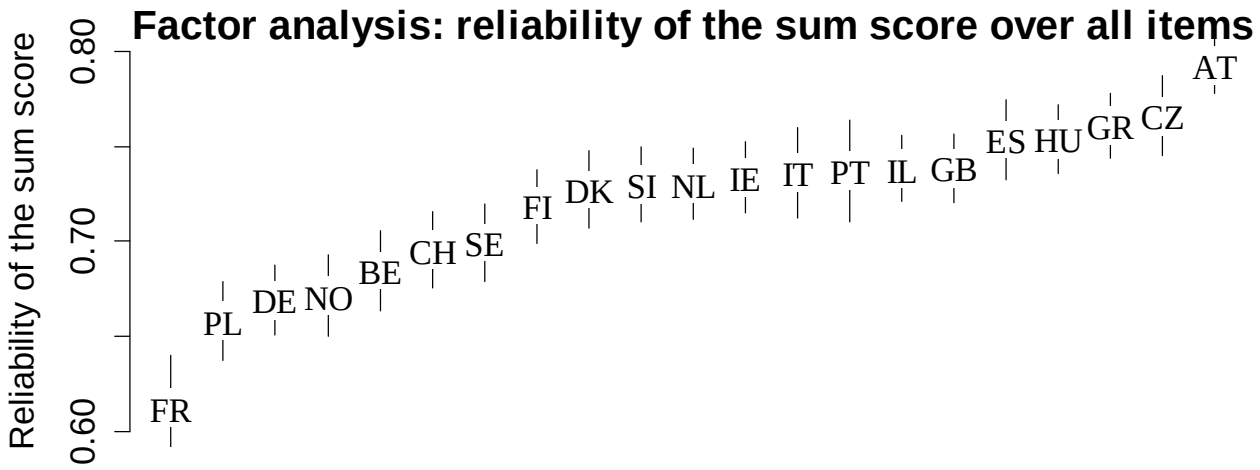

Figure 8. Sum score reliabilities from the factor model with approximate 2 s.e. confidence intervals.

Oberski, Saris, & Hagenaars (2007) showed that there can be considerable differences in the reliability across countries. The same can be seen at the bottom of Figure 2 for the reliability of the single variable level of education. Figure 8 shows the composite score reliability for the trust scale in different countries, with 2 s.e. intervals. It can be seen that reliabilities range between 0.6 and 0.8 over countries.

To illustrate the consequences of these differences, suppose a researcher regressed voting for right-wing political parties on the average of the three social trust items in Austria and France. Using data from the ESS round 2, in Austria the coefficient for the linear regression of voting for the FPÖ on the average of the three trust items is -0.0203 (0.004). In France the same coefficient for voting for the Front National is -0.0204 (0.005). These two coefficients are remarkably close to each other, suggesting that in both countries the relationship between social trust and voting for right-wing parties is equal.

Figure 8 shows that the reliability of the trust score is 0.8 in Austria but only around 0.6 in France. This suggests that the *true* relationship is stronger in France than it is in Austria, possibly threatening the comparability of the regression coefficients. Correction for unreliability changes the regression coefficients to -0.025 and -0.033 for France and Austria, respectively. Even after correction for differential measurement error, then, the two coefficients are similar, and not statistically significantly different from each other. The same result is obtained using probit regression with correction for differential measurement error[6]: the *p*-values for the test of different regression coefficients across countries before and after correction are 0.987 and 0.364, respectively. Thus, in this case the differential measurement error shown in Figure 8 did not change the result of the cross-country comparison. A case where the cross-country comparison *is* radically different after correction was already shown in Figure 2.

[6] The analysis and correction using probit models were done using Mplus 5.2.

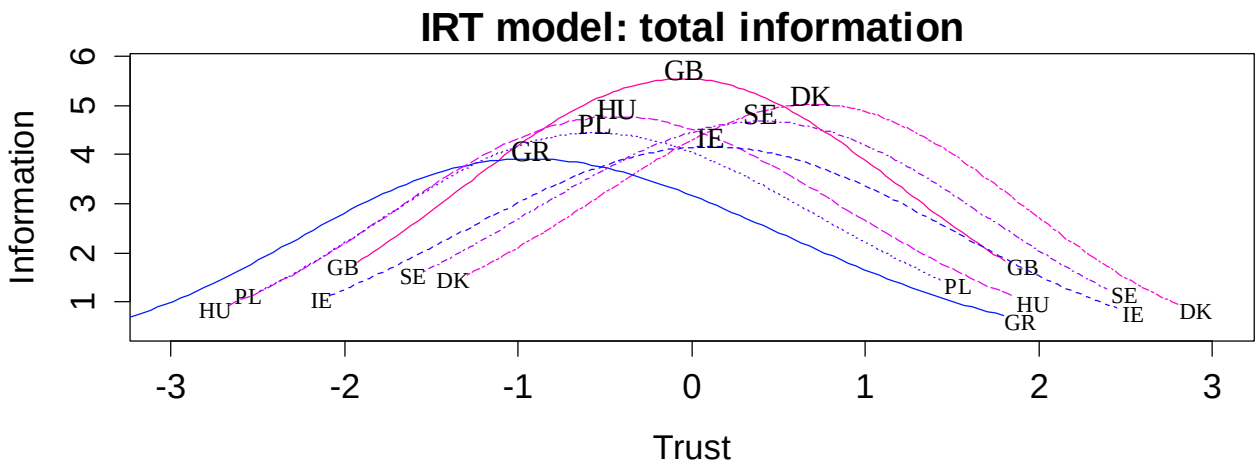

Figure 9. Total information curves in different countries for the trust IRT score.

One may also obtain the measurement error estimates from IRT models rather than factor analysis. In that case the total information curve gives the reciprocal of the conditional variance of the IRT trust score. This information curve is shown for six different countries in Figure 9, plotted against the rescaled trust composite score.

The higher the curve in Figure 9 at a point, the higher the reliability will also be (Mellenbergh, 1994). For example, at the mean of the reference country (Austria), which is set to zero, the measurement quality is highest in the UK, and lowest in Greece. But starting at about 1.3 standard deviations of Trust below the reference country's mean, the measurement quality is highest in Greece.

Overall comparability of measures of association in IRT will depend not only on the differences between the curves, but also on the marginal distribution of the Trust variable in each country (e.g. Oberski, 2011). For instance, if Greeks are less trusting than Britons, the average reliability may still be higher in Greece than in the UK. Therefore, the determination of comparability is more complicated than in linear factor models, but the same principles of weak and strong comparability apply.

Models with a stochastic systematic component

So far only models with bias (intercepts) and random measurement error (slopes and/or variance parameters) were considered. Another possibility is that each respondent has their own method of answering questions: there may be a systematic "response style" or "method effect".

A response style occurs when a respondent tends toward certain choices over all questions regardless of the content (Billiet & McClendon, 2000). Acquiescence, extreme response style, and middle response style are the most often considered. A method effect occurs when a respondent tends toward certain choices over those questions asked by a certain method regardless of the content (Saris & Andrews, 1991). For example, if acquiescence were to apply only to agree-disagree scales, it should be considered a method effect rather than a response style.

The common denominator of response style and method effect is that they represent a systematic response error on the part of the respondent that is *stochastic*, i.e. differing over respondents. I use the term "stochastic systematic error" to distinguish this type of error from relative bias, which, confusingly, is sometimes also called "systematic error" (Groves, 2004).

The responses to two different variables answered by the same person will be dependent if that person has a stochastic systematic error: response style and method effects cause correlated measurement error. Besides causing possible bias in comparison of means, then, comparisons of *relationships* will also be affected by differing stochastic systematic errors. If the response style or method effect obeys a linear factor model, observed regression coefficients will equal

$$\beta_{YX} = \kappa_X \beta_{yx} + (1 - \kappa_X) \beta_{e_Y e_X} ,$$

where $\beta_{e_Y e_X}$ is the regression coefficient that would be obtained by regressing the measurement error of Y on the measurement error of X. The coefficient $\beta_{e_Y e_X}$ will increase as the variance of style or method factors over survey respondents increase. It can therefore be seen that variations in response style or method factors bias the regression coefficient upwards in each group. Therefore, if there is more variation in response styles in one group than in the other, the regression coefficients will differ, threatening comparability.

Question*: Overall in the last 30 days, how much of a problem did [name/you] have with moving around?*

Response categories:

1. *None*
2. *Mild*
3. *Moderate*
4. *Severe*
5. *Extreme/Cannot Do*

Vignettes:

V1. *[Mary] has no problems with walking, running or using her hands, arms and legs. She jogs 4 kilometres twice a week.*

V2. *[Rob] is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometre or climbing up more than one flight of stairs. He has no problems with day-to-day physical activities, such as carrying food from the market.*

V3. *[Anton] does not exercise. He cannot climb stairs or do other physical activities because he is obese. He is able to carry the groceries and do some light household work.*

V4. *[Vincent] has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feel heavy.*

V5. *[David] is paralyzed from the neck down. He is unable to move his arms and legs or to shift body position. He is confined to bed.*

Figure 10. Anchoring Vignettes for Mobility, used in World Health Survey instrument in 2002. Source: http://gking.harvard.edu/vign/eg/?source_extra=mobility.shtml.

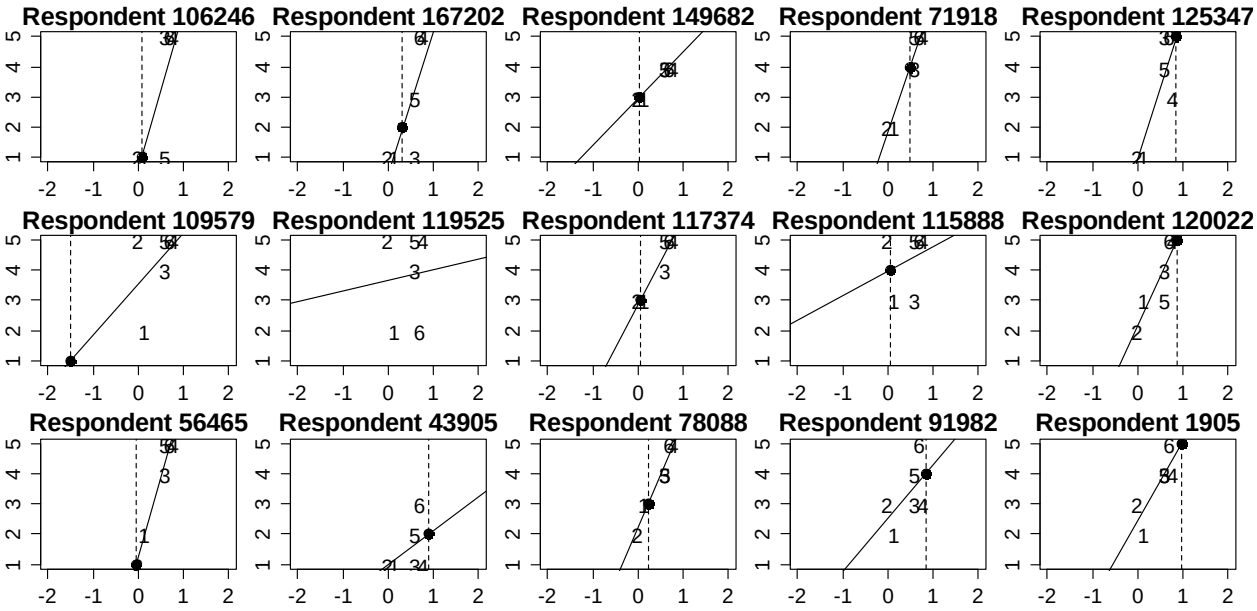

Figure 11. Anchoring vignettes: estimated response functions for respondents from Eastasia (top), Eurasia (center), and Oceania (bottom). The dotted line shows the estimated "self" score when the response is mapped back onto the response function. Each column shows randomly chosen respondents giving answers in categories 1 through 5, respectively.

In order to either establish comparability or correct for differing method or style influences, special models are necessary. For estimating the method factor influence, so-called "multitrait-multimethod" (MTMM) models can be used (Saris & Andrews, 1991). Such models, as well as models with style factors are discussed in more detail by Billiet in the present volume.

Another method to estimate individual response functions is that of "anchoring vignettes" (King et al., 2004). An example of this method is given in Figure 10. A question on degree of problems with mobility is asked each time of the same respondent using different stimuli. Afterward, the same question is asked of the respondents themselves. Respondents' answers to the "self" questions can then be corrected for the individual response functions, as estimated from the responses to the vignettes.

Under certain model assumptions, each individual response to the "self" questions can be corrected for that individual's response function. Therefore cross-group comparison on the transformed scores cannot be threatened by measurement incomparability if the assumptions of the model hold[7].

Figure 11 shows a very rudimentary example analysis. Response functions for 15 respondents are plotted, as estimated from their answers to the vignettes. Each of the five columns in Figure 11 shows a respondent who gave the response 1, 2, 3, 4, or 5 to the "self" question, respectively. These responses have been "mapped back" onto the respondents' response functions, shown by the vertical dotted lines. The three rows represent respondents from three different continents. Thus, differences between the vertical lines over rows represent evidence of incomparability of the responses. For example, respondent 106246 in the top left corner and respondent 109579 just below that gave the same response (1), but this implies a much lower true opinion for the second respondent than for the first.

[7] The principle of anchoring vignettes is identical to that of response function analysis in classical psychophysics.

Figure 11 serves only as an illustration of the principle behind anchoring vignettes. In practice response functions are estimated with a more advanced model than the rather simplistic linear regression employed here (King et al., 2004; Wand, King, & Lau, 2007).

Factor analysis, IRT, MTMM models, response style models, and anchoring vignettes are all measurement models intended to estimate the response function. These models are relevant for cross-group comparability because they can estimate variation in response functions over the groups to be compared under the appropriate model assumptions. The models are used for different purposes:

1. Establishing equivalence through invariance testing or DIF analyses (strong comparability);
2. Evaluating whether inequivalence is large enough to invalidate cross-group comparability (weak comparability);
3. Correction for inequivalence, if possible, either through
    a. estimation directly in the model, or
    b. post-estimation adjustments of comparison parameters (two-step approach).

The first two applications, establishing weak or strong comparability, were already discussed. Correction of the comparison for inequivalence can be done either by direct estimation in the model conditional on comparability (Byrne et al., 1989), or by adjustment of the comparison after the estimation, as was done in Figure 2, for instance (see also Saris & Gallhofer, 2007). The three applications are not mutually exclusive, but complement each other.

Conclusion

Survey measurements are subject to survey errors: thus, whenever different groups are compared on some survey statistic, survey errors in each of the groups may invalidate that comparison. The existence of this possibility was called strong incomparability. The possibility that the effect of the total survey errors is strong enough to invalidate the particular comparison made was termed weak incomparability.

Incomparability may stem from the “representation” side of total survey error, or from the “measurement” side. Traditionally the literature has focused on measurement issues, i.e. the fields of IRT, DIF, invariance testing, MTMM modeling, response style modeling, and anchoring vignettes.

These approaches do not usually consider the possibility that nonresponse, frame, or other selection biases may also invalidate comparisons. It was shown here that recent developments in the “representation” side, in particular the R-indicator, may in some cases provide useful evidence as to the comparability of survey statistics with respect to such representation errors. Future work in this field may extend the notion of comparability to the representation side.

The methods and models discussed in this chapter may be employed to prevent, detect, and correct for incomparability. The purpose of this chapter was to give the reader an impression and demonstration of the issues and possible solutions, but it was not possible to provide an exhaustive overview here. For further information the reader is referred to the specialized literature cited at the end of this chapter.

References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement* (Vol. 547). New York: Wiley.

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 143–156.

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Billiet, J. B., & McClendon, M. K. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*(4), 608–628.

Boateng, S. K. (2009). Significant country differences in adult learning. *Population and social conditions*, Statistics in focus. Eurostat.

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.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly*, *68*(1), 57.

Cohany, S. R., Polivka, A. E., & Rothgeb, J. M. (1994). Revisions in the current population survey effective January 1994. *Emp. & Earnings*, *41*, 13.

Davidov, E., Schmidt, P., & Billiet, J. (2010). *Cross-cultural Analysis: Methods and Applications*. New York: Taylor and Francis.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2008). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (3rd ed.). Wiley.

Donsbach, W., & Traugott, M. W. (2008). *The SAGE handbook of public opinion research*. Sage Publications Ltd.

Fuller, W. A. (1987). *Measurement error models*. Wiley Online Library.

Ganzeboom, H. B. G., & Schröder. (2009). Measuring Level of Education in the European Social Survey. *Keynot speech.* Presented at the European Survey Research Association (ESRA), Warsaw.

Goudy, W. J. (1976). Nonresponse effects on relationships between variables. *Public Opinion Quarterly*, *40*(3), 360.

Groves, R. M. (2004). *Survey errors and survey costs* (Vol. 536). New York: Wiley.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, *72*(2), 167.

Harkness, J. A. (2003). Questionnaire translation. *Cross-cultural survey methods*, *325*, 35.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., Pennell, B. E., et al. (2010). *Survey Methods in Multicultural, Multinational, and Multiregional Contexts* (Vol. 552). Wiley.

Harkness, J. A., Vijver, F. J. R., & Johnson, T. P. (2003). Questionnaire design in comparative research. *Cross-cultural survey methods*, 19–34.

Harzing, A. W. (2006). Response styles in cross-national survey research. *International Journal of Cross Cultural Management*, *6*(2), 243.

Hoffmeyer-Zlotnik, J. H. P., & Harkness, J. A. (2005). *Methodological aspects in cross-national research*. Mannheim: Zentrum für Umfragen, Methoden und Analysen (ZUMA).

Holland, P. W. (1982). *Test equating*. Academic Press.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. NJ: Lawrence Erlbaum Associates Hillsdale.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology. *Journal of cross-cultural psychology*, *16*(2), 131–152.

Jowell, R. (2007). *Measuring attitudes cross-nationally: Lessons from the European Social Survey*. Sage Publications Ltd.

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for Measurement Invariance With latent Class Analysis. *Davidov, E., Schmidt, P. and Billiet, J. (eds.). Cross-Cultural Analysis: Methods and Applications*. New York: Taylor and Francis.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(01), 191–207.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*(3), 213–236.

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires.

Malhotra, N., & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. *Political Analysis*, *15*(3), 286–323.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300.

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515.

Millsap, R. E., Meredith, W., Cudeck, R., & MacCallum, R. (2007). Factorial invariance: Historical perspectives and new problems. *Factor analysis at 100: historical developments and future directions*, 131.

Oberski, D. (2011). *Measurement error in comparative surveys*. Tilburg: Tilburg University.

Oberski, D., Gruner, T., & Saris, W. E. (2011). *SQP 2*. Retrieved from http://www.sqp.nl/

Oberski, D., Saris, W. E., & Hagenaars, J. (2007). Why are there differences in measurement quality across countries. *Measuring Meaningful Data in Social Research. Acco, Leuven*.

Oberski, D., Saris, W. E., & Hagenaars, J. (2008). Categorization errors and differences in the quality of questions across countries. *Loosveldt, G., Swyngedouw, D., Cambré, B. (eds.). Measuring meaningful data in social research*. Leuven: Acco.

Petty, R. E., & Krosnick, J. A. (1995). *Attitude strength: Antecedents and consequences.* Lawrence Erlbaum Associates, Inc.

Reeskens, T., & Hooghe, M. (2008). Cross-cultural measurement equivalence of generalized trust. evidence from the european social survey (2002 and 2004). *Social Indicators Research*, *85*(3), 515–532.

Saris, W. E. (1988). *Variation in response functions: A source of measurement error in attitude research* (Vol. 3). Sociometric Research Foundation.

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. *Measurement errors in surveys*, 575–597.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research* (Vol. 548). New York: Wiley.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101–113.

Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J., & Zhang, L. (2008). Statistical Properties of R-indicators. *RISQ Work Package*, *3*. Retrieved from http://www.risq-project.eu/papers/RISQ-Deliverable-2-1-V2.pdf

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

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Transaction Publishers.

Stoop, I., Billiet, J., & Koch, A. (2010). *Improving survey response: Lessons learned from the European Social Survey*. John Wiley & Sons Inc.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge Univ Pr.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (Vol. 1). Sage Publications, Inc.

von dem Knesebeck, O., Verde, P. E., & Dragano, N. (2006). Education and health in 22 European countries. *Social Science & Medicine*, *63*(5), 1344-1351. doi:10.1016/j.socscimed.2006.03.043

Voogt, R. J. J. (2004). *I'm not interested: nonresponse bias, response bias and stimulus effects in election research*. University of Amsterdam.

Wand, J., King, G., & Lau, O. (2007). Anchors: Software for anchoring vignette data. *Journal of Statistical Software*, *42*, 1–25.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications, inc.

Zavala, D. (2011, February 7). Deviations found through SQP coding in the ESS Round 5 questionnaires. Report given to the European Social Survey's national coordinator's meeting.