Harmonizing Single-Question Instruments for Latent Constructs With Equating Using Political Interest as an Example

Singh, Ranjit Konrad

Survey Research Methods (2022)
Vol. 16, No. 3, pp. 353-368
doi:10.18148/srm/2022.v16i3.7916
European Survey Research Association

Harmonizing single-question instruments for latent constructs with equating using political interest as an example

Ranjit K. Singh
GESIS—Leibniz-Institute for the Social Sciences
Survey Design and Methodology
Mannheim, Germany

Many latent constructs in the social sciences, such as political interest, are measured with single-question instruments. Furthermore, survey programs often differ in the wording and response format of those instruments. This is problematic if we want to compare or combine data across different surveys or across changing instruments within a survey program. Consequently, we need robust methods to establish comparability in existing data (i.e., ex-post harmonization). In this paper I demonstrate the usefulness of an approach from psychometry: Observed score equating with a random groups design. Using two existing instruments for political interest, I show that harmonizing instruments with equating works well in transforming the numerical scores of instruments so that they can be compared across instruments. Since random groups equating needs data for both instruments from the same population, I also demonstrate the feasibility of equating using an online nonprobability sample as well as using two probability samples of the adult German population.

Keywords: survey measurement instruments; data harmonization; comparability; equating; integrative data analysis

1 Introduction

Quantitative research in the social sciences often relies on large scale survey programs. Such national and international programs provide a valuable service to the research community and accumulate a veritable treasure trove of data. At the same time, survey programs are very complex multi-stakeholder endeavors. One consequence of this complexity is that survey programs often have their own methodological idiosyncrasies. One area where this is most apparent are the measurement instruments for different concepts (in the following just *instruments*). Different survey programs often use instruments for the same concept that differ in their question wording, the number of response options, or the labeling of response options and many other design aspects as well (Tomescu-Dubrow & Slomczynski, 2016). And even within a survey program, instruments may change over time.

A drawback of this diversity is that these instrument differences reduce the comparability of data across time, populations, and different survey programs. Instrument diversity might also make it harder to satisfy the growing demand for FAIR (findable, accessible, interoperable, reusable) data (Link, Lumbard, Germonprez, Conboy, & Feller, 2017), since it makes survey data less interoperable and thus also less reusable. In practical terms, this means that researchers cannot easily compare findings based on one instrument with that based on another instrument for the same concept. Furthermore, researchers increasingly combine data from different sources to answer their research questions. For example to combine existing research data into an integrative dataset for meta-analytical purposes, to achieve adequate sample sizes for smaller subpopulations (e.g., LGBTQI, migrants, first generation academics etc.), or to combine data for a specific substantive topic (e.g., democratic values and protest behavior; Tomescu-Dubrow and Slomczynski, 2016). Different measurement instruments are challenging for such use cases as well.

Consequently, there is a need for efficient and methodologically sound approaches to *ex-post harmonize* data on the same concept measured with different measurement instruments. Ex-post harmonization encompasses all activities that aim to increase comparability of existing data after it was collected and in ways that the survey was not designed for. The specific focus of this paper lies on the ex-post harmonization of different single-question survey measurement instruments for the same latent construct. *Latent constructs* in this context are mental concepts that cannot be directly observed but only indirectly inferred (Bollen, 2002; Price, 2017), such as attitudes, values, emotions, or in our paper:

Ranjit K. Singh; GESIS—Leibniz-Institute for the Social Sciences, Survey Design and Methodology, B6, 4–5, 68159 Mannheim (Email: Ranjit.Singh@gesis.org).

political interest. Single-question instruments for latent constructs, meanwhile, are especially interesting (and challenging) in ex-post harmonization because many approaches to wrangle latent constructs in psychometry, such as confirmatory factor analysis, require multiple questions (i.e., items, or indicators) for the same construct (Price, 2017).

The common situation researchers (and often also data producers and archivists) find themselves in is that they are faced with two or more single-question instruments for the same latent constructs. This may mean different question wording, a different number of response options, different response labels, or many other differences in the instruments' design. In this paper, I use as a practical example two measures for political interest, with different question wording, response labels, and a different number of response options (i.e., four-point scale versus five-point scale). And now the question arises: How can we make responses to the two instruments comparable? This question is crucial, if researchers want to choose which instrument to use in their own research, if they want to compare results based on different instruments, or if they want to combine data gathered with different instruments.

In sum the present paper is intended as a proof of concept for using equating to ex-post harmonize single-question instruments for the same latent construct in the social sciences. I aim to demonstrate that a specific psychometric approach can help us increase the comparability of single-question instruments for latent constructs ex-post: *Observed score equating with a random groups design*. The approach is usually applied in the context of psychometric individual diagnostics, such as harmonizing scores of different versions of a psychometric test (Kolen & Brennan, 2014). However, while some equating techniques require multi-item scales, observed score equating formulas can be applied to increase the comparability of single-question instruments as well. As an example, I use political interest as a frequently used construct with comparatively straightforward question wordings.

The paper will first address the theoretical background of instrument harmonization in general and observed score equating in particular. Then the concrete research questions and study design are described. Afterwards, the methods and results section follow. Lastly, in the discussion, I will summarize core messages and point out some issues to consider when applying equating.

2 Theoretical Background

In the following, we will first consider the broader challenges that occur when we want to combine data on a latent construct measured with two different single-question instruments. We will then focus on a specific challenge: How can we transform scores measured with different instruments so that they have comparable units of measurement. We will consider why this is a pressing issue and why traditional approaches such as linear stretching fall short of what we need. Then, I explain how Observed Score Equating in a Random Groups Design works and why this is a suitable solution for the problem of incomparable units of measurement.

2.1 Harmonization of single-question instruments for latent constructs

Let us assume that we want to combine data on the same construct measured with two different instruments. The result we aim for is a variable combining values from both (or more) instruments but representing the same concept. However, we have to consider four issues before we can safely run analyses using this combined variable: (1) Do both instruments measure the same construct? (2) Are both instruments similarly reliable measures of the same construct? (3) Do both instruments have similar units of measurement. In the following, we will briefly address issues one and two because the method we propose—observed score equating—does not solve them. Instead, it solves the fourth, which will be the focus going forward.

First and foremost, we have to ensure that both instruments measure the same latent construct. In the language of psychometric test theory, we would like both instruments X and Y to be at least congeneric tests of the same latent construct (adapted from Raykov and Marcoulides, 2011, section 5.3.3).

$$X = d_X + b_X T + E_X \text{ and } Y = d_Y + b_Y T + E_Y \qquad (1)$$

This means we assume that the scores of both instruments (X and Y) represent respondents' true score of the same latent variable (T), albeit with different units of measurement (constants $d \wedge b$) and different errors (E) and different error variances. However, for our context—the harmonization of single-question instruments in the social sciences—we cannot formally test for these assumptions as we would with psychometric multi-item instruments (Price, 2017; Raykov & Marcoulides, 2011). Instead, we will often have to assess both instruments separately, by applying techniques of instrument validation, such as correlational measures of construct validity (Price, 2017). In this paper, I will briefly sketch out one such approach, by correlating both instruments of political interest with related constructs and comparing the correlation patterns of both instruments.

Second, there is the issue of reliability. Different instruments may be more or less susceptible to random measurement error. Lower reliability (and thus a higher level of random measurement error) leads to increased attenuation (i.e., an underestimation of correlations between constructs due to random measurement error). In harmonization, the issue is less attenuation itself, but rather different levels of attenuation depending on the source instruments. However, this again is a hard challenge since the reliability of single-question instruments is harder to assess then that of multi-item instruments. However, if the instruments' reliability can be assessed, the problem can be mitigated by correcting analyses for attenuation.

2.2 Comparable units of measurement

Third, we have to consider the units of measurement of the different instruments. The issue is already implied in the congeneric model, where even in the absence of measurement error, different instruments would still generate different scores in our data. Formally, even if we removed the error terms E_X and E_Y, we would still have to consider the constants d_X, b_X, d_Y, and b_Y which project the true scores onto the numerical scale of X and Y:

$$T = \frac{X - d_X}{b_X} = \frac{Y - d_Y}{b_Y} \tag{2}$$

In fact, even the assumption of equidistant response options is often unrealistic. Instead, scores to single-question instruments often behave more like ordinal projections of the underlying continuous true score continuum. Individual response options thus differ which range of true scores they represent. An idea which is formalized in item response theory (IRT), for example. We will get back to this idea when we discuss equipercentile equating.

In practical terms, this differently designed instruments project the underlying latent construct expressions onto different responses and thus result in a different numerical scale (Price, 2017). For example, people with the same level of political interest might choose different responses in different instruments. Meanwhile people who were assigned the same numerical score by different instruments may have very different true levels of political interest.

On the aggregate level in the data, this implies that different instruments represent the same population with different response distributions (Kolen & Brennan, 2014). If we applied two instruments to the same group of respondents, we might get response distributions with different means, standard deviations, skewness, or even completely different distribution shapes (e.g., uni- vs. bimodal). And if instruments were applied to different populations, then these instrument differences mingle with and thus bias the true population differences. In any case, comparability and the potential for integrated analyses suffers. To solve this problem, we must find some way to convert the "measurement units" of one instrument into the numerical format of the other.

2.3 Linear stretching

Data harmonization practitioners in the social sciences are, of course, aware of the issue of different measurement units. However, often the focus lies on the most obvious source of scale differences: The number of scale points. After all, if we measured the same construct with a four-point scale and a seven-point scale in the same population, we would not expect comparable response distributions. Instead, we would expect the mean and standard deviation of the seven-point scale to be higher than that of the four-point scale.

With that in mind, a popular approach to harmonizing measurement units is linear stretching (e.g Durand, Peña Ibarra, Rezgui, & Wutchiett, 2021; Tomescu-Dubrow & Slomczynski, 2016). The approach sets the minimum scores and the maximum scores of both instruments as equal and then stretches all other scores with equal distances in between (Cohen, Cohen, Aiken, & West, 1999; de Jonge, Veenhoven, & Kalmijn, 2017). Formally, linear stretching is a linear transformation solely based on the maximum possible scores of both instruments. For simplicities sake, let us score the minimum responses as zero with maximum scores of K_X and K_Y For instruments X and Y. Then the formula for linear stretching is:

$$\text{stretch}(x) = \frac{xK_Y}{K_X} \tag{3}$$

In our example, we would thus stretch the four-point instrument (X) towards the seven-point instrument (Y) by multiplying each score of X by six (K_Y) and then dividing by three (K_X). Thus, the scores 0, 1, 2, and 3 of the four-point instrument become 0, 2, 4, and 6. This approach certainly mitigates the problem of different measurement units in cases where the number of response options differ. However, the units of measurement also depend on other factors, such as the question wording, the response labels, and the visual layout of the question. This is again already implied in the congeneric model above. After all, the congeneric model is usually used in psychometry to describe multi-item scales which all share the same response options. And yet, the units of measurement differ. As a practical illustration of the shortcomings of linear stretching, imagine two instruments with the same number of response options. However, one instrument is worded more strongly than the other (e.g., "I am passionate about" vs. "I am interested in"). We would expect the stronger wording to result in lower average scores, because the stronger statement is harder to agree to. The units of measurement have shifted due to differences in item difficulty (Moosbrugger & Kelava, 2012). In the paper, I will use linear stretching as a baseline to contrast with the novel harmonization approach of observed score equating.

2.4 Observed Score Equating with a Random Groups Design

How then does observed score equating with a random groups design mitigate differences in measurement units? Equating in general emerged in psychometric diagnostics to solve a very similar problem: How can we transform scores of different (performance) tests, so that the results can be fairly compared? The general idea is to establish some way of transforming scores measured with one (source) instrument so that they become comparable to scores measured with another (target) instrument (Kolen & Brennan, 2014). Equating is hereby different from other ex-post harmonization approaches in that it does not only transform the current dataset. Its goal is instead to derive a transformation rule for two instruments that can be used in other data as well. In practice, we would derive a recoding table in one dataset and then use it in many other instances where the same instruments were used.

The core idea of all equating approaches is the so-called *equity property* of equating. After an ideal equating process, respondents with the same true score for a construct should, on average, get the same score in the transformed source instrument than they would get in the target instrument (Kolen & Brennan, 2014). Formally, we want participants with a particular true score to have the same expected transformed score in X than the expected score they would have in Y. With $E(\cdot)$ being the expectation operator (i.e., the mean) and eq_Y being an equating function that transforms scores of X into their instrument Y equivalents (Kolen & Brennan, 2014, p. 10):

$$E\Big(\text{eq}_Y(X|\tau)\Big) = E(Y|\tau) \quad \text{for all } \tau \tag{4}$$

However, for single-question instruments, we cannot extract true score estimates and thus many equating approaches cannot be applied. However, observed score equating does not require true score estimates. Instead, the observed score equity property focuses on the cumulative distribution of the observed scores in our datasets. If we applied instrument X and instrument Y in the same population and then equated instrument towards instrument Y, we would expect the transformed scores of X to have the same cumulative response distribution (G^*) than the cumulative response distribution of instrument Y (G) (adapted from Kolen & Brennan, 2014, p. 11)):

$$G^*\Big(\text{eq}_Y(x)\Big) = G(y) \tag{5}$$

That may sound abstract, but it is a very desirable property in ex-post harmonization. If we measured the same population, we would get the same (harmonized) response distribution shape regardless of the instrument we used. This means, for example, that different instruments no longer bias the mean, the standard deviation, and in more advanced equating methods, we also mitigate bias in skewness, kurtosis, or even multi-modal distributions.

Additionally, observed score equating is *symmetrical*, in the sense that we can transform instruments in both directions with no loss of information (Kolen & Brennan, 2014). The result of the equating process is, in essence, recoding information with which one instrument can be transformed into the format of another. This means that equating harmonizes instruments and not just the present data. It also means that we can perform equating with data well suited to equating and then use the resulting recoding information in other data we want to harmonize and where the instruments were used.

However, before delving into the concrete process of observed score equating with random groups design, I want stress that harmonizing with equating does not harmonize differences in content, meaning differences in the constructs that the two instruments measured are not corrected for (Kolen & Brennan, 2014). It also does not mitigate differences in instrument reliability (Kolen & Brennan, 2014).

Random groups design. After all that preamble: How does observed score equating in random groups design work, exactly? The key here is the random groups design; a research design for collecting data suitable for observed score equating. In a random groups design, we collect samples for both instruments drawn randomly from the same population. In psychometry, this usually means performing a split-ballot experiment where one half of participants randomly answers instrument X and the other random half instrument Y. However, in this paper, we will explore other equating data sources as well.

Through the random groups design, we can expect approximately the same cumulative true score distribution in both samples. In other words: The experimental design has set the true score distribution equal. The cumulative observed score distributions, meanwhile, will differ between the samples of instrument X and Y. This is because the two instruments transform the true scores differently into observed scores. However, since those differences represent differences in measurement units and not in true scores, removing those differences aligns measurement units. Observed score equating, in other words, transforms scores of instrument X so that the cumulative response distribution now matches that of instrument Y in the same population. Please note, however, that this approach does not account for the error term in measurement. Random measurement error, for example, is simply passed on, as mentioned earlier.

Aligning response distributions with linear or equipercentile transformations. Now only a practical issue remains: How to align response distributions? In this paper, I demonstrate two widely used distribution transformation approaches: Linear and equipercentile. In the following, I describe the logic and the mathematical formulas behind both approaches. However, in practice, equating can be effortlessly performed using specialized software. For example, the equate package for R (Albano, 2016) which I also used for this paper.

Linear equating. The *linear equating* logic is very straightforward. Response distributions of instruments A and B are assumed to be approximately normally distributed and response options to be equidistant. Linear equating thus sets the z-scores of the two instruments equal, meaning that converted scores of X will have the same mean and standard deviation then scores of Y in the same population (Kolen &

Brennan, 2014, p. 31).

$$\frac{x-\mu(X)}{\sigma(X)} = \frac{y-\mu(Y)}{\sigma(Y)} \tag{6}$$

This means we only have to align the mean and the standard deviation with a linear equating function $l_Y(x)$ (Kolen & Brennan, 2014, p. 31):

$$l_Y(x) = \sigma(Y)\left(\frac{x-\mu(X)}{\sigma(X)}\right) + \mu(Y). \tag{7}$$

As a result, the responses to B now have the mean and the standard deviation of A. This is, of course, very similar to a z-standardization. The difference is that linear equating then makes scores of instrument X interpretable in the numerical format of instrument Y (or vice versa) whereas z-standardization makes instruments interpretable in terms of a specific population at a specific point in time. This means that the result of linear equating can be applied to other populations and still be interpretable in terms of the target instrument.

As a result, the transformed responses to X now have the mean and the standard deviation of Y. This is, of course, very similar to a z-standardization. The difference is that linear equating then makes instrument X interpretable in the numerical format of instrument Y (or vice versa) whereas z-standardization makes instruments interpretable in terms of a specific population at a specific point in time. This means that the result of linear equating can be applied to other populations and still be interpretable in terms of the target instrument.

Equipercentile equating. *Equipercentile equating*, meanwhile, does not assume normal response distribution shapes. Instead, it aligns the cumulative response distributions by matching the percentile ranks of responses. The basic idea is that we transform each score of instrument X into its corresponding percentile rank in the random group's population. Then we transform each of that percentile ranks into a corresponding score of instrument Y with the same percentile rank. This seems intuitive: If the median response is a "3" in instrument X and a "2" in instrument Y, we would match those two scores.

However, the actual mathematical process is a bit more complicated, for two reasons. First, each response option represents a whole segment of construct intensities and thus a whole segment of percentiles. If 14% of respondents choose the first response option, then respondents from the 0th to the 14th percentile most likely choose this option. Second, the response options of different instruments never match perfectly in the percentiles they represent. Equipercentile equating solves both these issues with linear interpolation. Figure 1 illustrates this interpolation process which transforms the relative frequency distribution into a function that assigns each response scores a percentile rank. Importantly, the function is continuous, which allows us to find the percentile ranks of non-integer response scores. This means we can find a continuous, interpolated response score value for any arbitrary percentile rank. Please note, that all figures and formulas for equipercentile equating formulas assume scores to start at zero for the first response option. This simplifies all formulas and is if no practical consequence since equating software, such as the equate R-package (Albano, 2016), perform this internal transformation automatically. For now, just note that a 0 in figure 1 would most likely be a "1" in the dataset.

Mathematically, equipercentile equating thus needs two functions: (1) A percentile function, which transforms scores of an instrument into linearly interpolated percentiles and (2) an inverted percentile function, which transforms percentiles into linearly interpolated, "continuized" response scores (Kolen & Brennan, 2014). Please note that for simplicities sake the formulas below assume that there is at least one response per possible response score. In single-question instruments, this is a reasonable assumption. However, if zero frequency scores occur, the formulas must be slightly modified (Kolen & Brennan, 2014, 42ff). The equate package (Albano, 2016) does this automatically.

Let x be a decimal score of instrument X from zero to the maximum possible response score K_X. Let x^* be the integer score closest to a decimal score x, so that $x^* - 0.5 \leq x < x^* + 0.5$. Finally, let $f(x)$ be the relative frequency function for scores of x and $F(x)$ be the cumulative relative frequency function for scores of x.

Then the percentile function (adapted from Kolen & Brennan, 2014, p. 42) can be described as:

$$P(x) = 100 \cdot \Big(F(x^* - 1) + (x - (x^* - 0.5)) \cdot f(x^*)\Big) \tag{8}$$

In essence, the formula interpolates the percentile rank as the cumulative relative frequency of the integer response option one lower than the x^* plus a portion of the relative frequency of x^*.

What the formula implies is this: A response option, for example "2", represents an interpolated score range from 1.5 to 2.5. It thus encompasses percentiles from $P(1.5) = F(1) \cdot 100$ to $P(2.5) = F(2) \cdot 100$. An integer score, meanwhile, is halfway between those cumulative frequencies: $P(2) = \Big(F(1) + \frac{1}{2}f(2)\Big) \cdot 100$. If 20% of respondents choose responses lower than "2", and 10% chose "2", then the percentile rank of "2" would be 15.

To transform a percentile rank P^* back into the corresponding interpolated score, we need the inverted percentile rank function $P^{-1}(P^*)$ so that:

$$P^{-1}(P^*) = P^{-1}\Big(P(x)\Big) = x \tag{9}$$

For this we need to first find x^*_U, which is the smallest integer score where $P^* < F(x^*_U)$. This might seem confusing, but x^*_U is nothing else than the integer score closest to the interpolated x that we will get as a result of $P^{-1}(P^*)$.

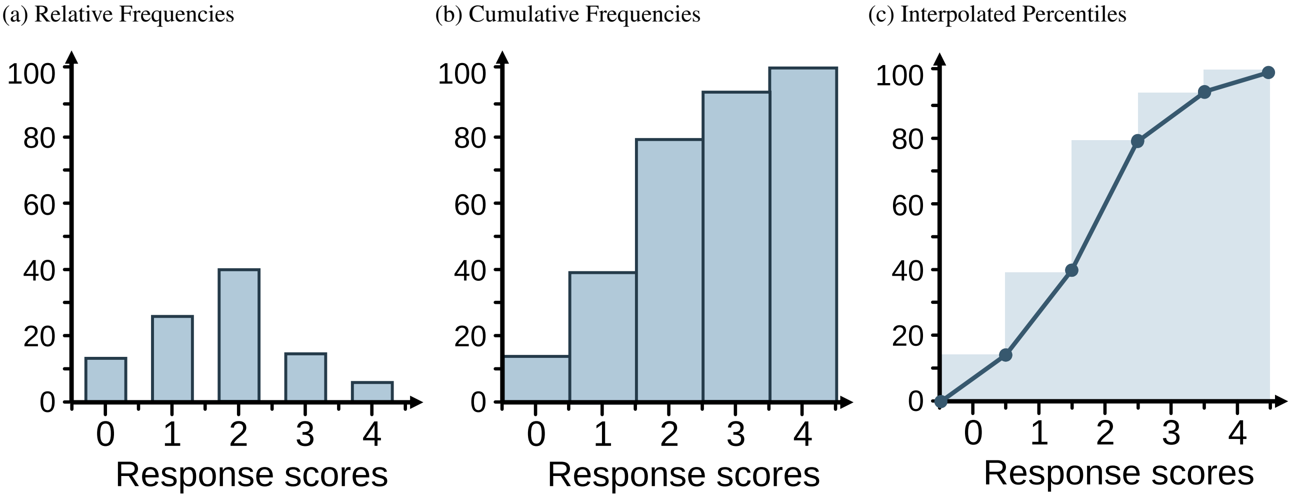

Figure 1. Linear interpolation of continuous percentile ranks in equipercentile equating

The inverted percentile function (Kolen & Brennan, 2014, p. 43) then is:

$$P^{-1}(P^*) = \frac{\frac{P^*}{100} - F(x_U^* - 1)}{f(x_U^*)} + (x_U^* - 0.5) \qquad (10)$$

The formula works like this: With x_U^* we have determined the closest integer score to the result. Thus, we know the score segment (if $x_U^* = 2$, we expect the result to be between 1.5 and 2.5). The fraction then simply determined how far we are from the lower bound (e.g., 1.5) to the upper bound (e.g., 2.5). If we supply exactly the percentile rank of an integer score, for example, the fraction resolves to 0.5 and thus the result x is identical with x_U^*.

To finally perform equipercentile equating, we create the percentile function and inverted percentile function both for instrument X (P and P^{-1}) and instrument Y (Q and Q^{-1}). The equipercentile equating function $e_Y(x)$ (Kolen & Brennan, 2014, p. 44) thus becomes:

$$e_Y(x) = Q^{-1}\Big(P(x)\Big) \qquad (11)$$

Every score of instrument X is transformed into a percentile rank and then transformed into an interpolated score in the format of instrument Y. This setup also demonstrates the symmetry property of equating. After all, we can just as easily transform scores of Y into the format of X with $P^{-1}\Big(Q(y)\Big)$. Figure 2 represents this process visually.

The result are usually not integer scores, but approximations with decimal places. A "2" in X might be best represented by a "2.8" in Y. This might seem unusual at first, but those approximations have desirable properties. Once transformed, the response distribution parameters (e.g., mean, sd, skewness) align well, allowing for unbiased comparisons and joint analyses (Kolen & Brennan, 2014).

Equating terminology. Lastly, a note on the equating terminology which will be helpful in navigating the primary psychometric literature. Equating is part of a larger literature on linking. Linking encompasses many different approaches which all seek to establish some form of comparability. Equating, meanwhile, is a specific family of linking approaches (Kolen & Brennan, 2014). However, in psychometry the term "equated" also implies that a set of comparability criteria have been met. The ideal of equating is that two psychometric tests become fully exchangeable. This means for example that they are identically reliable, and that test-takers have no preference over one test over the other (Kolen & Brennan, 2014; Price, 2017). These quality requirements are important for lawful and fair diagnostics of individual test-takers, such as tests used for professional aptitude diagnostics.

However, this distinction is less relevant for the ex-post harmonization of survey instruments. It is just important to note that while we can apply equating formulas to survey instruments, psychometrists would call the result linked, aligned, or calibrated, but not equated (Kolen & Brennan, 2014; Price, 2017). However, these terms each encompass a wide and sometimes contradictory range of procedures (Kolen & Brennan, 2014). In this paper we thus use equating to mean applying the mathematical procedures described above, which makes it easy to find pertinent literature.

2.5 Study design and research questions

The paper will use the example of two instruments for political interest, to explore the usefulness of observed score equating. Here, a remaining issue is how to get random groups data with which to perform the equating. As you will recall, equating with a random groups design requires data for both instruments randomly drawn from the same popu-

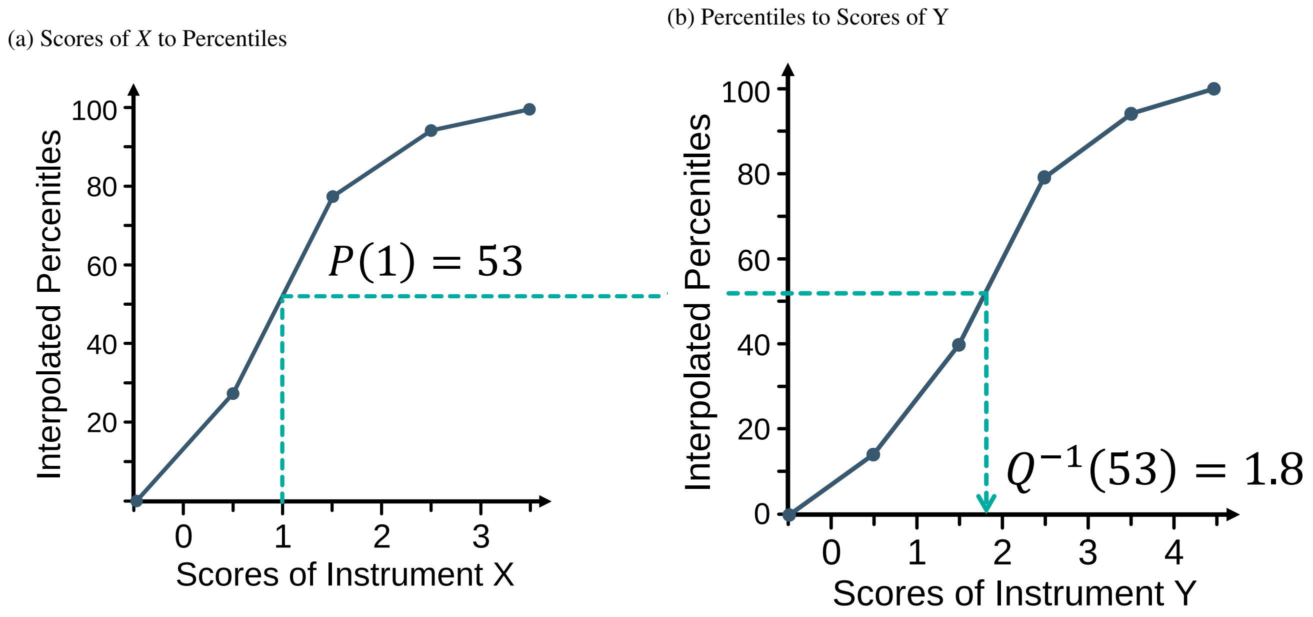

Figure 2. Transforming response scores from instrument X into the format of instrument Y. *Note:* Scores are set to start with zero as the first response option to simplify the formulas. Thus, the score 1 here represents a score of "2" in the dataset, i.e., the second response option.

lation. This can be achieved with experimental studies, but such data are usually not readily available for different instruments. Hence the paper also explores two approaches to acquire the necessary data. Specifically, the paper has four parts.

First, I provide some empirical evidence that both instruments for political seem to measure the same construct. This is not the focus of the paper, but it is an important issue because equating cannot mitigate differences in content. Fortunately, we have several questions about related concepts in the ALLBUS-ISSP 2014 dataset that we can use.

Second, we tackle the problem of different measurement units, by applying both the traditional linear stretching approach as well as the two observed score equating approaches: linear and equipercentile equating. Then the distributions of the transformed instrument are compared to the reference instrument. This allows us to fairly compare the harmonization quality of linear stretching, linear equating and equipercentile equating.

Third, I explore a pragmatic approach to gain data suitable for equating: Collecting affordable data specifically to equate the two instruments in a nonprobability online access panel. Specifically, the study was set up as a split-half experiment in which respondents randomly answered one instrument or the other. This is the random groups design in its purest form. However, the data quality and the undefined sample in nonprobability online access panels may cast doubt on this approach. To test the resulting equating relationship, it is then applied to the ALLBUS–ISSP data used in the first two parts, to validate the result.

Fourth, I explore another possible approach to gain data suitable for equating, which makes the most of the highly developed survey landscape in the social sciences: Using data from two probability surveys of the same population ideally in the same year. For example, if two probability surveys sample the adult German population, an argument can be made that those two samples are random samples of the same population (i.e., a random groups design). To test this approach, I use data from the ALLBUS and from the GLES which uses the same instrument for political interest as the ALLBUS. Performing an equating between two samples of the same instrument seems counterintuitive. However, this is an ideal test of how much differences in sampling, survey mode, or other survey characteristics bias equating. After all, if we equate an instrument with itself, we know the perfect result: an identity relationship. This means that the resulting recoding table would recode a "1" to a "1", a "2" to a "2" and so on. A substantial empirical deviation from this ideal thus represents a bias in equating brought on by survey or sample differences.

3 Methods

3.1 Samples

Probability Samples. This study uses data from three survey programs: First, the ALLBUS, the German general social survey. Second, the German part of the ISSP, the International Social Survey Programme. And third, one of the surveys of the GLES German Longitudinal Election Study; specifically, the pre- and postelection cross-sectional survey (in the following just GLES).

The most central data set hereby is the combined dataset of the ALLBUS and the German part of the ISSP in 2014 (GESIS-Leibniz-Institut Für Sozialwissenschaften, 2018). The ISSP shares the same sample as the ALLBUS in Germany. After completing the ALLBUS, conducted every two years, participants are randomly given one of the yearly ISSP waves to answer. Hence, the sample we use are respondents who answered both the ALLBUS 2014 and the ISSP 2014 (N = 1704). This sample allows us to focus on comparability in a setting where we can easily validate answers. After all, respondents answered one instrument in the ALLBUS and then the other in the ISSP. This allows for a comparison without sampling differences. At the same time, the two instruments are spaced far apart and embedded in many other questions. Learning effects are unlikely. In the following, this specific dataset is called ALLBUS-ISSP 2014 data.

For the fourth part of this paper ("Equating across different probability surveys") we use data from the GLES survey from 2017 (GLES, 2019), which uses the same instrument for political interest as the ALLBUS (N = 4290). We then equate the GLES 2017 data with ALLBUS data from 2016 (GESIS-Leibniz-Institut Für Sozialwissenschaften, 2017; N = 3490) and 2018 (GESIS-Leibniz-Institut Für Sozialwissenschaften, 2019; N = 3475). I use both ALLBUS waves to interpolate a response distribution for an ALLBUS survey in 2017 (which does not exist due to the biyearly ALLBUS schedule). Equating the same instrument with itself allows us to clearly isolate possible biases introduced by survey differences.

Web experiment sample and procedure. Lastly, to assess the potential of affordable nonprobability samples for equating, a web-experiment was conducted. The nonprobability sample was recruited via the commercial online access panel of the respondi AG (Respondi, 2021). The sample size was N = 2171. The sample had a median age of 41. The youngest participants were 18 years old. Half of participants reported their sex as female (50%); two respondents chose the offered option "divers". In Germany, "divers" is now the official category for people whose biological sex does not fit into the female-male dichotomy (i.e., intersex persons). The sample was more highly educated than the national average. More than half of respondents (54%) had at least passed a university entrance exam, in contrast to only 34% in the adult German population (Statistisches Bundesamt (Destatis), 2020).

Regarding the experimental procedure: The web-survey was composed of modules from different empirical studies. However, I omit describing the other components, because the experiment used in the present paper was the first and thus not influenced by the other modules. Respondents were greeted and briefed about their privacy protection rights. The survey was anonymous and did not save respondents IP-addresses. Next, respondents answered a number of socio-demographic questions. Then, respondents either saw one or the other of the two instruments for political interest that are to be harmonized (i.e., the a5 and i4 instruments described in the next section). Then, the questionnaire continued with elements not relevant for the present paper.

3.2 Measurement Instruments for political Interest

The paper focusses on two single-question measurement instruments for political interest. Unlike psychometric scales, the instruments have no formal names. Meanwhile, naming them directly after the surveys they are from is misleading, because some are used in several surveys, while the same survey might use different instruments over time. Hence, I use short codes for each of the two instruments: a5 and i4.

a5 instrument. The a5 (ALLBUS 5-point) instrument is the political interest measure that is in use in the ALLBUS since the surveys first wave in 1980. It is still used today and has an almost unbroken time series with the sole exception of 1988, where an alternative instrument had been used (Baumann & Schulz, 2018). The a5 instrument is a single question with a five-point response scale. The question wording was "How strongly are you interested in politics?"[1]. The five response options were "very strongly", "strongly", "middling", "not very", and "not at all"[2]. The response options are unipolar. The instrument was scored with integer values from 1 ("very strongly") to 5 ("not at all"). The instrument, as is custom in the ALLBUS, did not offer an explicit non-substantive response option.

i4 instrument. The i4 (ISSP 4-point) instrument was used in three waves of the ISSP so far. Two times in the Citizienship modules I and II in 2004 (ISSP Research Group, n.d.) and 2014 (ISSP Research Group, 2016) respectively. And one time in 2007 in the Leisure Time & Sports module (ISSP Research Group, 2009). Please note that the i4 instrument is not the only political interest instrument used by the ISSP.

The i4 instrument is a single question with a four-point response scale. The question wording was "How interested would you say you are in politics?"[3]. The four response options were "very interested", "fairly interested", "not very

[1] German original: "Wie stark interessieren Sie sich für Politik?"

[2] German original: "sehr stark", "stark", "mittel", "wenig", and "überhaupt nicht"

[3] German version: "Was würden Sie sagen, wie sehr sind Sie an Politik interessiert?"

interested", and "not at all interested"[4]. The response options are unipolar. The instrument was scored with integer values from 1 ("very interested") to 4 ("überhaupt nicht interessiert"). The ISSP, unlike the ALLBUS, offers an explicit non-substantive response option: "Cannot choose"[5]. Although it should be noted that this option is very seldomly chosen. In the ISSP of 2014, only seven respondents (0.4%) chose the "Cannot choose" option. In our web-experiment, only eight respondents (0.7%) of respondents chose this option.

3.3 Analysis and Software

R. All data transformations and analyses were conducted in R (R Core Team, 2021) using RStudio (RStudio Team, 2022). All original datasets were in SPSS format and read into R using haven (Wickham & Miller, 2018). The tidyverse package collection (Wickham, 2017) was used for data transformation and data visualization. Skewness was calculated using the moments package (Komsta & Novometsky, 2021). Correlations were compared using the cocor package (Diedenhofen & Musch, 2015). All equating operations were performed using the *equate* package (Albano, 2016). For linear and equipercentile observed score equating in random groups design, the package uses the formulas and algorithms from Kolen and Brennan (2014), which I had summarized in the theory section.

Linear Stretching. Linear stretching was performed using a custom R function. Unlike the simplified function described earlier, the i4 and a5 instruments are scored with a minimum score of one. Thus, for instruments X and Y with minimum scores J and maximum scores K the formula becomes:

$$\text{stretch}(x) = \frac{(x - J_X)(K_Y - J_Y)}{K_X - J_X} + J_Y \tag{12}$$

This formula sets the minimum response options equal to each other, the maximum response options equal to each other, and then distributes all responses in between with the same distance (Cohen et al., 1999; de Jonge et al., 2017).

Item difficulty. Descriptive item difficulty was calculated using the formula of (Dahl, 1971) reported in (Moosbrugger & Kelava, 2012, p. 81), which is applicable to items with Likert scales (as opposed to difficulty for dichotomous response scales). For an instrument X with a minimum score J and maximum score K, we can calculate the item difficulty as:

$$P_X = \frac{\sum_{i=1}^{n}(x_i - J_X)}{n(K_X - J_X)} \dot{1}00 \tag{13}$$

This form of item difficulty can be interpreted as a measure where the average response lies along the range of a scale. A difficulty of zero means that all respondents chose the first response option. A difficulty of 100 means that all respondents chose the highest possible response option. Values in between can be interpreted as a percentual position between those two extremes.

Modified Cohen's *d* for mean bias (i.e., Glass' Δ). To report mean bias in an easily interpretable format, Cohen's *d* is used as a measure of mean difference. Cohen's *d* expresses mean differences relative to the standard distribution. However, which standard distribution? If two groups have the same variance, the issue is moot. If they differ in variance, then we can either pool the variances or choose the variance of one group. In this paper, I want to compare mean bias fairly. Thus, I standardize Cohen's *d* using the standard deviation of the target scale (a5). This is important because the harmonization approaches might also introduce biases into the standard deviation of the transformed source instrument (i4). By using only the target standard deviation, the comparison isolates the mean bias issue. As a second issue, please note that I present absolute mean difference values, because the direction of the mean bias is irrelevant for this paper. Technically, standardizing with the standard deviation of only one group means I use Glass' Δ, not Cohen's *d*. However, since the delta could easily be confused with a mere unstandardized difference, I use *d* as the more well-known concept (Hedges & Olkin, 1985).

4 Results

4.1 Construct validation of the a5 and i4 instruments

Equating should only be applied when two instruments measure the same construct. Unfortunately, we cannot formally demonstrate that two single-question instruments are congeneric tests of the same construct. However, we can at least explore how the two instruments correlate with related constructs. In our example, we can use the combined ALLBUS-ISSP 2014 data in which both instruments (i4 and a5) were used alongside some related constructs. In table 1 below, we see how the two instruments correlate with (1) interest in TV news, (2) interest in political TV shows, (3) Respondents self-assessed understanding of important political issues facing Germany, and (4) how often respondents discuss politics with others. Note that all instruments are coded in the same direction as political interest: Lower scores represent higher interest (or understanding or frequency). The sample was $N = 1704$, including only respondents who answered the ALLBUS questionnaire and the ISSP 2014 questionnaire.

Both instruments result in almost identical correlations with the related constructs. None of the correlation differences were significant, according to Fisher (1925) *r*-to-*z* transformation. While this is not a formal proof of con-

[4] German version: "sehr interessiert", "einigermaßen interessiert", "eher nicht interessiert", and "überhaupt nicht interessiert"

[5] German version: "Kann ich nicht sagen"

Table 1
Correlation of the a5 and i4 instrument with related constructs

	r_{a5}	r_{i4}	Δr	p
Interest in TV news	0.37	0.38	−0.00	0.934
Interest in political TV shows	0.61	0.58	0.03	0.132
Understanding of the important political issues facing Germany	0.54	0.57	−0.03	0.245
How often do you discuss politics	0.56	0.59	−0.02	0.346

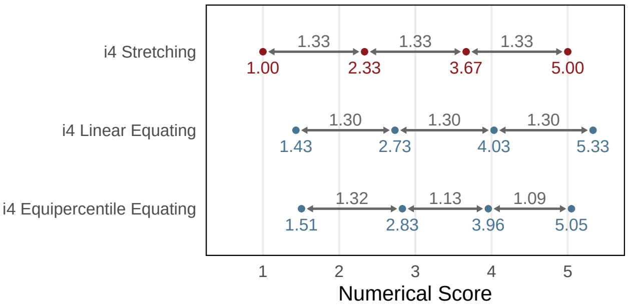

Figure 3. Equivalent values of the i4 responses in terms of the target instrument a5

generic tests, it does support the assumption that both instruments capture very similar concepts. The very comparable correlations also imply that the reliability of both instruments is similar, because otherwise, correlations of the instrument with lower reliability would be consistently lower due to attenuation.

4.2 Observed Score equating: Proof of principle

Here and in all following parts, the a5 instrument (five-point scale first used in the ALLBUS) will serve as the target instrument. The i4 instrument (four points, used in the ISSP) is the source instrument which we want to align to the a5 instrument. To illustrate the challenge and to compare the efficacy of linear stretching against that of observed score equating, I first harmonized scores of the i4 instrument towards the a5 instrument with three different approaches.

Figure 3 shows the resulting transformed scores. Below, in figure 4, we see the resulting transformed scores. In each row, the first point represents the first i4 response option, the second point the second response option and so on up to the maximum i4 score of 4.

If we consider figure 4 more closely, we see the logic of the different approaches clearly illustrated. All three approaches preserve the ordinal structure of the instrument scores. However, we see that linear stretching is a very rigid approach. It completely ignores any response distribution information and instead only takes the scale points of both instruments into account. The transformed responses are bounded between the minimum and maximum score and responses in between are treated as equidistant. Linear equating, in contrast, does take aspects of the response distributions of both instruments into account. Linear equating can shift transformed scores left or right to mitigate mean bias. It can also accommodate different standard deviations by stretching or compressing the range of transformed responses. However, the linear equating solution is still equidistant. Finally, equipercentile equating is the most flexible approach. Aside from shifting left or right and stretching or compressing the range of values, equipercentile equating is also no longer bound by equidistance. Note how the distance between the first two response options (1.32) is markedly greater than the distance between the last two response options (1.09).

Mean bias mitigation. The question remains, however, if these transformations have indeed helped solve the comparability issue. In the ALLBUS-ISSP 2014 data, we can assess this by comparing the distribution of the transformed i4 scores to the distribution of the a5 scores. Both i4 distributions and the a5 distribution are based on the same population and thus should be very similar after harmonization. First, with regard to mean bias mitigation, figure 4 shows the mean of the i4 and a5 instrument on the left and the mean bias compared to the a5 instrument on the right.

Not transforming the i4 instrument results in a mean bias of $|d| = 0.69$. Linear stretching mitigates this somewhat, but a substantial mean bias of $|d| = 0.38$ remains. This is unsurprising, because the two instruments have different item difficulties of $P_{\text{a5}} = 43$ and $P_{\text{i4}} = 33$. In other words, average respondents position themselves at different positions along the scale ranges of a5 and i4. Observed score equating, meanwhile, removes mean bias effectively. Linear equating, results in a perfect mean match, reducing the mean bias to $|d| = 0$. This is hardly surprising, given that linear equating directly aligns the mean responses of two instruments. More interestingly, the rather less straightforward equipercentile equating algorithm also resulted in an almost perfect mitigation of the mean bias with $|d| = 0.04$.

Higher distribution moments. Next, we consider the higher distribution moments. While reducing mean bias is important, harmonization should ideally also align other aspects of the response distribution shape, such as the standard deviation and the skewness. Table 2 lists the standard deviation and skewness values as well as their distance from the a5 target instrument.

Unsurprisingly, the untransformed i4 instrument has a markedly lower standard deviation than the a5 instrument. In our example, meanwhile, all harmonization approaches mitigate the standard deviation differences effectively. However, that linear stretching works well here is a coincidence that only occurs if the standard deviations of two instruments are similar in relation to the range of response options (i.e., max-

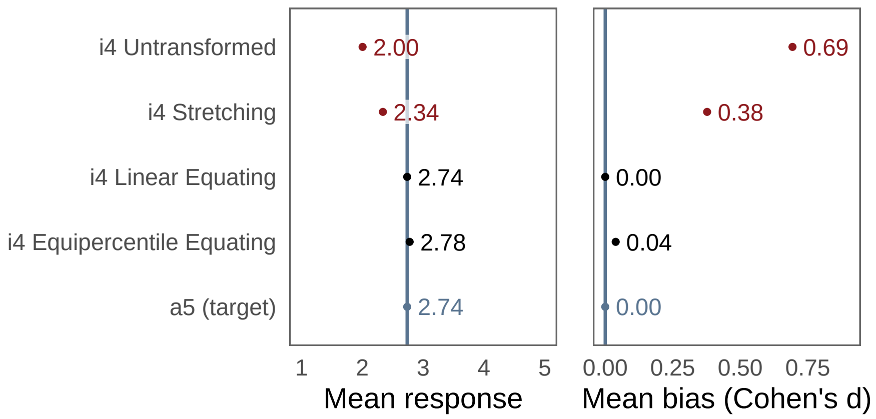

Figure 4. Mean response and mean bias by harmonization procedure. *Note:* Mean bias is calculated as absolute Cohen's d using the standard deviation of the a5 targetinstrument.

Table 2
Comparison of standard deviation and skewness across approaches

Approach	SD	|ΔSD|	Skewness	|ΔSkewness|
i4 Untransformed	0.81	0.24	0.60	0.45
i4 Stretching	1.09	0.03	0.60	0.45
i4 Linear Equating	1.06	0.00	0.60	0.45
i4 Equipercentile Equating	0.98	0.08	0.37	0.22
a5 (target)	1.06		0.16	

imum score minus minimum score). The a5 instrument has a range of 4, the i4 instrument a range of 3. The standard deviation divided by the range then happens to be very similar for both instruments: 0.26 for a5 and 0.27 for i4. Both observed score equating procedures, in contrast, can also mitigate standard deviation differences that are not due to the scale range (Kolen & Brennan, 2014).

Skewness, meanwhile, is left unchanged by linear stretching and linear equating. Only equipercentile equating mitigates the skewness difference because its transformed scores need not be equidistant. The skewness bias is not fully eliminated but is half as strong after equipercentile equating as compared with all other approaches.

4.3 Equating with nonprobability samples

Of course, a data structure as in ALLBUS-ISSP 2014 is rare. Instead, ex-post harmonization with equating will often require a separate source for data in a random group design. Here, I explore a straightforward approach: Using data from a split-half experiment conducted in an affordable nonprobability online access panel. The approach is as follows: With data from the split-half experiment featuring the i4 and the a5 instruments, both linear and equipercentile equating are performed. The resulting recoding tables are then applied to the ALLBUS-ISSP 2014 data to transform the i4 scores there. Then, the quality of this linking with external data is assessed just like in part two.

First, we look at the mean bias mitigation. Figure 5 below gives an overview that includes linear stretching as a worst case, direct equating via ALLBUS-ISSP 2014 data, and now the equating via nonprobability sample.

As we can see, the mean bias is reduced substantially. Instead of the mean bias of $|d| = 0.38$ in linear stretching, the bias is reduced to $|d| = 0.10$ in linear and $|d| = 0.07$ in equipercentile equating. Table X meanwhile gives an overview of the higher distribution moments as well. Using the recoding table from the nonprobability sample underestimates the standard deviation slightly by 0.07 scale points in linear equating and 0.16 scale points in equipercentile equating. The skewness mitigation of equipercentile equating is markedly lower in the nonprobability sample case (difference from target = 0.40) than in the direct equating case (difference from target = 0.20), but still better than linear stretching or equating (difference from target = 0.45). In sum, a very good mean bias mitigation with a slight trade-off towards less variance. Table 3 lists all standard deviation and skewness comparisons.

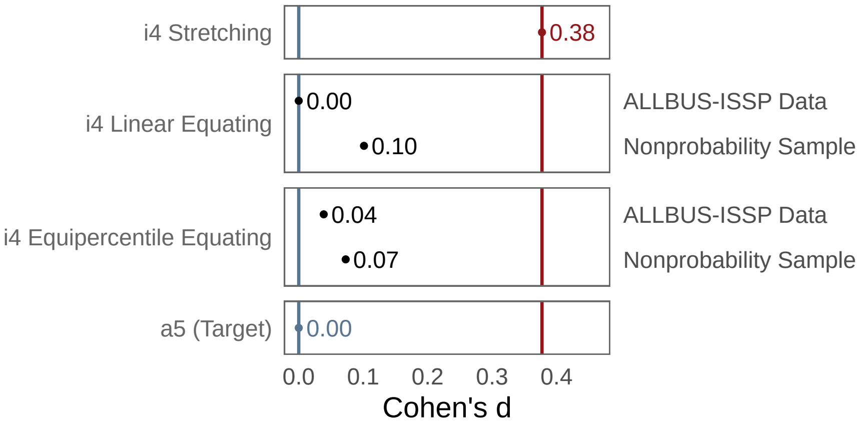

Figure 5. A comparison of mean bias in i4 responses by harmonization approach and data source

Table 3
SD and Skewness comparison by harmonization approach and data source

Instr.	Harmonization Approach	Data Source	SD	|ΔSD|	Skewness	|ΔSkewness|
a5	target		1.06		0.16	
i4	linear equating	ALLBUS-ISSP Data	1.06	0.00	0.60	0.45
i4	linear equating	Nonprob. Experiment	0.99	0.07	0.60	0.45
i4	equipercentile equating	ALLBUS-ISSP Data	0.98	0.08	0.37	0.22
i4	equipercentile equating	Nonprob. Experiment	0.90	0.16	0.56	0.40

4.4 Equating across different probability survey programs

Lastly, I demonstrate that observed score equating can also be performed with data from different surveys with random samples of the same population. Specifically, equating is performed between data from the GLES and the ALLBUS surveys, which both use the a5 instrument. It is also important to match survey waves temporally. Otherwise, changes in the construct over time might bias equating. Thus, GLES data from 2017 is equated with ALLBUS data from 2016 and 2018 combined for form an interpolated ALLBUS 2017.

Equating the a5 instrument with itself seems counterintuitive. However, this offers an ideal test for equating across different survey programs. After all, equating an instrument with itself should result in an identity relationship, meaning a 1 should be linked to a 1, a 2 to a 2, a 3 to a 3 and so on. Any deviation from that identity relationship would then indicate bias introduced into the equating process by survey characteristics, such as sampling strategy or survey mode. Figure 6 shows the result of both the linear and equipercentile equating of the a5 scale in GLES 2017 towards the same a5 scale in ALLBUS 20016 and 2018 combined. The linear and equipercentile equating results fall almost perfectly onto the grey diagonal of an identity relationship. Numerically, the equating solution differed from the ideal identity solution only by an average of 0.02 scale points in linear equating and anaverage of 0.03 scale points in equipercentile equating. The highest difference overall was 0.07 scale points from the deal identity solution. In practical terms, this implies that survey differences, sampling issues, and pooling the two ALLBUS waves did not bias the equating process.

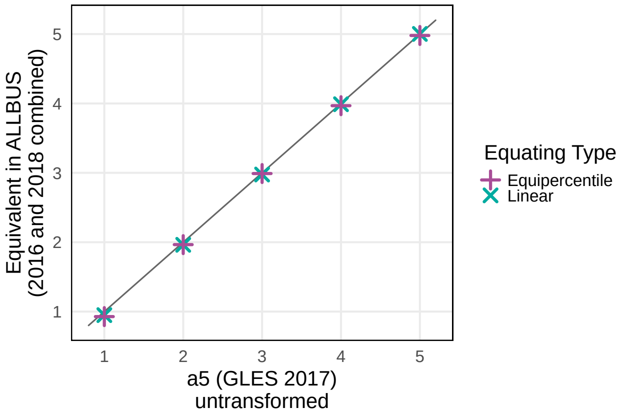

Figure 6. a5 scores in GLES 2017 and their equivalent values in the format of a5 in ALLBUS 2016 and 2018 combined

Next, I applied the recoding tables from equating to the ALLBUS 2014 data we used earlier. To be precise, I transformed the a5 responses in the ALLBUS from 2014 with the equating recoding tables derived from the comparison of GLES 2017 and ALLBUS 20016 and ALLBUS 2018. Again, since we equate an instrument with itself, the ideal outcome would be a perfect match between the distributions of the linearly and equipercentile equated a5 scores and the distribution of untransformed a5 scores. Any substantial differences would imply bias introduced by survey characteristics. Figure 7 shows that such bias was almost non-existent.

Numerically, both linear and equipercentile equating missed the mean by a mere $|d| = 0.03$ standard deviations. The standard deviation difference from the target instrument was a mere 0.01 for both. There was no skewness difference for linear equating and a difference of merely 0.05 for equipercentile equating.

In summary, equating with data from the two probability survey programs GLES and ALLBUS resulted in an almost perfect match with the theoretically ideal outcome. Nothing indicates that the survey and sample characteristics of the two surveys introduced a bias to the equating process.

5 Discussion

The paper explored the usefulness of observed score equating for the ex-post harmonization of single-question survey instruments for the same latent construct; here specifically, political interest. First, the problem of different instruments was demonstrated and observed score equating was applied to mitigate bias. Even two very similar instruments resulted in a substantial mean bias even after attempting to align them with linear stretching ($|d| = 0.38$). Applying observed score equating with a random groups design to harmonize the two instruments directly in the ALLBUS and ISSP data from 2014 resulted in almost perfect numerical comparability. However, this case where the data we want to harmonize is already structured in a way that allows for equating is seldom. Consequently, third, the possibility of collecting data explicitly for the purpose of equating was explored. If you will recall, equating in random groups design requires samples of both instruments randomly drawn from the same population. As one possible solution, I collected data on both instruments in an experiment in a nonprobability online access panel. The equating solution resulting from this experimental data was then applied to the ALLBUS and ISSP 2014 case. This resulted in a slightly less perfect, but still very acceptable ex-post harmonization of the two instruments: a $|d|$ of 0.10 and 0.07 for linear and equipercentile equating respectively. Furthermore, a second approach to obtaining data for equating was explored: Using data from two probability surveys (i.e., ALLBUS and GLES) covering the same population in the same year (here: the adult German population). To clearly isolate potential biases introduced by sample differences and other survey characteristics, I equated the same instrument with itself. The result was an ideal identity relationship, implying that the survey characteristics of ALLBUS and GLES did not bias the equating process. Furthermore, transforming the ALLBUS 2014 responses and comparing them to the untransformed responses again yielded almost no bias at all.

The bottom line is that observed score equating with random groups design can result in valid and largely unbiased ex-post harmonization results. Both using equating data from nonprobability split-half experiments as well as from probability national surveys to gain data for instrument equating seems promising. As for the specific equating algorithm, both linear and equipercentile equating worked well with no compelling advantage of one over the other. However, the two instruments had few response options (five and four respectively) and more or less normal response distributions. With stark distribution differences and more response options, equipercentile equating might turn out to be generally more applicable to survey data.

5.1 Points to consider

Of course, harmonizing instruments using equating is no panacea for all ex-post harmonization challenges. Hence, I would like to stress a few boundaries and aspects that need to be considered when applying the approach.

First, equating does not mitigate differences in content (i.e., different constructs. Differences in content should be assessed before performing equating; for example, by applying validation techniques to both instruments. Unfortunately, a formal factor analytical assessment if two single-question instruments are congeneric measures of the same construct is not possible. However, we might apply techniques for criterion, content, and construct validation (Price, 2017). In this paper, for example, we correlated both instruments with related constructs.

Second, equating does not mitigate differences in measurement precision (i.e., different reliabilities, meaning different levels of random error in measurement). Ideally, the reliability of the data measured with both instruments should at least be known so that substantial reliability differences can be corrected for (e.g., *correction for attenuation;* Charles, 2005. However, estimating the measurement reliability of single-question instruments in surveys is a challenge; especially with the ease of calculating internal consistency measures for multi-item instruments. For an overview of approaches, see Tourangeau, Rips, and Rasinski (2000).

Third, the random groups design implies that OSE-RG primarily useful for harmonization within a language area and not across different languages. For the random groups design, we have to be able to sample from a common population, where respondents can understand both instruments similarly well. In cross-language contexts, this is hard to achieve since we would need costly samples of bilingual respondents. Furthermore, bilingual respondents may dif-

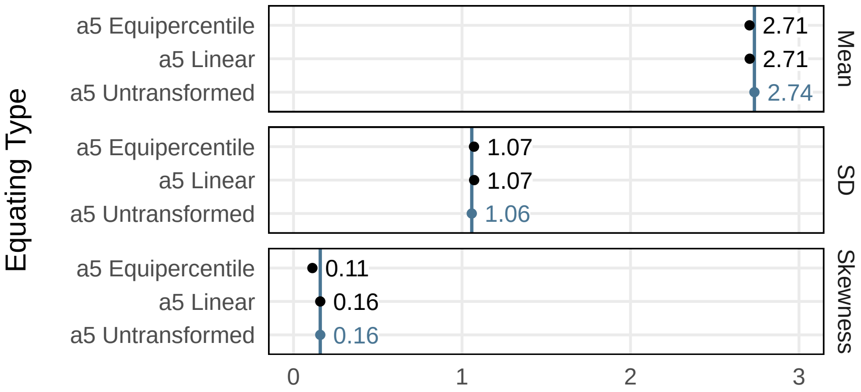

Figure 7. Distribution preservation after equating with data from two probability survey

fer systematically in their understanding from monolingual respondents, limiting the generalizability of harmonization results obtained from such samples (Sireci & Berberoglu, 2000).

Fourth, explicit non-substantive response options (e.g., an explicit "don't know"-option) are a potential complication. Specifically, a problem might occur if only one of the instruments offers a non-substantive response option. The basic problem is this: Equating in random groups design can only interpret ordinal, substantive responses (Kolen & Brennan, 2014). Non-substantive response options in one instrument thus can lead to what is in essence a drop-out from the equating data. This might break the assumptions of the random groups design that samples for both instruments represent the same population.

Fifth, there is the issue of group invariance. Equating can be performed with one data and the derived equating relationship between the two instruments (i.e., a recoding table) can be applied to different data that we want to harmonize. The group invariance property of equating would then state that population differences in the equating data and the harmonized data do not matter. Unfortunately, observed score equating is not formally group invariant, although empirical research often finds observed score equating to be robust across different populations ((Kolen & Brennan, 2014). What does that mean in concrete terms for applying observed score equating to survey instruments? If we apply data from probability surveys to other probability surveys covering the same population, the problem does not occur. If we want to use equating for specific populations, or if we want to use nonprobability data to perform the equating, then a problem might occur. However, the underlying problem here is not equating itself. Instead, the problem is that for single-question instruments we do not know if they are measurement invariant across the populations (Putnick & Bornstein, 2016) meaning we do not know if both instruments measure fairly across relevant dimensions such as age, sex, education and so on. Equating, again, does not introduce a bias here, but it reproduces the bias already in the data. If such problems occur, a pragmatic solution is to equate important subpopulations separately, and then apply the resulting subpopulation specific recoding tables to the target data (Dorans & Holland, 2000).

Sixth, as a matter of context, I would like to address some other research designs for equating. In this paper, the focus was on the random groups design: Data for both instruments randomly drawn from the same population. This design is applicable to single-question instruments which characterize many surveys in the social sciences. It can also make use of the wealth of probability survey data we have available. However, other designs exist, and I would like to address them. Firstly, the *single group* design. It works much like the random group design, but this time we ensure that responses represent the same population by asking the exact same respondents twice; once with each instrument (Kolen & Brennan, 2014). This is, in fact, the ALLBUS ISSP 2014 structure. I just used it as if it was a random groups setup to clearly demonstrate that the equating result is, in fact, valid. Unfortunately, we seldomly encounter single group data in large survey programs. The ALLBUS ISSP single group case is a rare coincidence. If we collect equating data ourselves, then the single groups design might mean placing the two instruments are closer to each other. This incurs the risk of order and learning effects. In other words, it would matter which instrument was shown first and which second. To counteract this, the single group design is usually counterbalanced, meaning the order of instruments is experimentally randomized. Still, the advantage over the random groups design is minimal if the sample sizes are adequately large.

Another design is the non-equivalent groups with covariates design (NEC). This design is in essence the attempt to approximate a random groups design with data from different populations (Wiberg & Bränberg, 2015). Instead of drawing from the same population, population differences are measured with additional variables (i.e., covariates) and then the equating is adjusted for those measured differences. It is not different from the idea of applying adjustment weights to a nonprobability sample. However, the approach is at most a second best to a true random groups design. The covariates approach can only account for measured population differences (and thus observed heterogeneity). Respondent differences that were not measured can still bias the equating. Lastly, the *non-equivalent groups with anchor tests* design (NEAT), is a very common research design for multi-item instruments (González & Wiberg, 2017; Kolen & Brennan, 2014). It is thus not applicable for the use case discussed in this paper. Still, it bears mentioning, because it is a very powerful approach. It can be applied of two multi-item instruments share at least some items. Those shared items are the so-called anchor tests, which then help bridge the instrument differences. This works even if the data for both instruments is drawn from different populations.

5.2 In a nutshell

Harmonizing data for a latent construct measured with different single-question instruments is a hard challenge. However, I hope to have shown that observed score equating can be a powerful tool in this context. It does require some effort in finding or collecting adequate data. Fortunately, there is a good chance to find adequate data in the well-developed survey landscape of the social sciences. Furthermore, if no adequate existing data can be found, even nonprobability online samples might offer an acceptable second best. And once adequate data has been found, the process of equating is very straightforward and can be applied to many different instruments and constructs.

References

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*(8). doi:10.18637/jss.v074.i08

Baumann, H., & Schulz, S. (2018). *ALLBUS—Kumulation 1980-2016. Variable report*. GESIS Datenarchiv für Sozialwissenschaften. Köln.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*(1), 605–634. doi:10.1146/annurev.psych.53.100901.135239

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*(2), 206–226. doi:10.1037/1082-989X.10.2.206

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346. doi:10.1207/S15327906MBR3403_2

Dahl, G. (1971). Zur Berechnung des Schwierigkeitsindex bei quantitativ abgestrufter Aufgabenbewertung. *Diagnostica*, *17*, 139–142.

de Jonge, T., Veenhoven, R., & Kalmijn, W. (2017). Diversity in survey items and the comparability problem. In T. de Jonge, R. Veenhoven, & W. Kalmijn (Eds.), *Diversity in survey questions on the same topic: Techniques for improving comparability* (pp. 3–16). doi:10.1007/978-3-319-53261-5_1

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solutionfor the statistical comparison of correlations. *PLOS ONE*, *10*(4), e0121945. doi:10.1371/journal.pone.0121945

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*(4), 281–306. doi:10.1111/j.1745-3984.2000.tb01088.x

Durand, C., Peña Ibarra, L. P., Rezgui, N., & Wutchiett, D. (2021). How to combine and analyze all the data from diverse sources: A multilevel analysis of institutional trust in the world. *Quality & Quantity*. doi:10.1007/s11135-020-01088-1

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

GESIS-Leibniz-Institut Für Sozialwissenschaften. (2017). ALLBUS/GGSS 2016 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2016). Version Number: 2.1.0 type: dataset. doi:10.4232/1.12796

GESIS-Leibniz-Institut Für Sozialwissenschaften. (2018). ALLBUS/GGSS 2014 General Social Survey 2014). Version Number: 2.2.0 type: dataset. doi:10.4232/1.13141

GESIS-Leibniz-Institut Für Sozialwissenschaften. (2019). ALLBUS/GGSS 2018 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2018). Version Number: 2.0.0 type: dataset. doi:10.4232/1.13250

GLES. (2019). Pre- and post-election cross section cumulation (GLES 2017). Version Number: 3.0.1 type: dataset. doi:10.4232/1.13236

González, J., & Wiberg, M. (2017). *Applying test equating methods*. doi:10.1007/978-3-319-51824-4

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

ISSP Research Group. (2009). International social survey programme: Leisure time and sports—ISSP 2007. Version Number: 2.0.0 type: dataset. doi:10.4232/1.10079

ISSP Research Group. (2016). International social survey programme: Citizenship II—ISSP 2014. Version Number: 2.0.0 type: dataset. doi:10.4232/1.12590

ISSP Research Group. (n.d.). International Social Survey Programme: Citizenship—issp 2004. Version Nummer: 1.3.0 type: dataset.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). doi:10.1007/978-1-4939-0317-7

Komsta, L., & Novometsky, F. (2021). moments: Moments, cumulants, skewness, kurtosis and related tests. Retrieved from https://CRAN.R-project.org/package=moments

Link, G., Lumbard, K., Germonprez, M., Conboy, K., & Feller, J. (2017). Contemporary issues of open data in information systems research: Considerations and recommendations. *Communications of the Association for Information Systems*, *41*(1), 587–610. doi:10.17705/1CAIS.04125

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (2nd ed.). Berlin: Springer.

Price, L. R. (2017). *Psychometric methods: Theory into practice*. New York ; London: The Guilford Press.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. doi:10.1016/j.dr.2016.06.004.Measurement

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Respondi. (2021). Access panel. Retrieved from https://www.respondi.com/access-panel

RStudio Team. (2022). RStudio: Integrated development for R. Retrieved from http://www.rstudio.com/

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education*, *13*(3), 229–248. doi:10.1207/S15324818AME1303_1

Statistisches Bundesamt (Destatis). (2020). *Bildungsstand der Bevölkerung—Ergebnisse des Mikrozensus 2019*. Statistisches Bundesamt (Destatis). Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsstand/Publikationen/Downloads-Bildungsstand/bildungsstand-bevoelkerung-5210002197004.pdf

Tomescu-Dubrow, I., & Slomczynski, K. M. (2016). Harmonization of cross-national survey projects on political behavior: Developing the analytic framework of survey data recycling. *International Journal of Sociology*, *46*(1), 58–72. Publisher: Taylor & Francis ISBN: 0020-7659. doi:10.1080/00207659.2016.1130424

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

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non -equivalent groups with covariates design. *Applied Psychological Measurement*, *39*(5), 349–361. doi:10.1177/0146621614567939

Wickham, H. (2017). tidyverse: easily install and load the 'Tidyverse'. R package version 1.2.1. Retrieved from https://cran.r-project.org/package=tidyverse.

Wickham, H., & Miller, E. (2018). haven: import and export SPSS, Stata and SAS files. R package version 1.1.2. Retrieved from https://cran.r-project.org/package=haven