Survey Research Methods (2023)
Vol. 17, No. 2, pp. 177-191
doi:10.18148/srm/2023.v17i2.8087
European Survey Research Association

Harmonizing data from open-ended and closed-ended quantity questions with observed score equating

Ranjit K. Singh and Matthias Roth
GESIS Leibniz Institute for the Social Sciences

Many surveys ask respondents about manifest quantities, such as income, age, weight, or their height. Surveys often either use open-ended questions, where respondents report the quantity directly as an integer value (e.g., "56"), or closed-ended quantity questions where respondents select from a set of discrete interval response options (e.g., "51 to 100"). Quantity data gathered with different response schemes thus becomes hard to compare or to harmonize to be used in integrative analyses. We compare two approaches to harmonizing quantity question data. Firstly, the widely used middle of category (MOC) interpolation. Secondly, Observed Score Equating in a Random Groups Design (OSE-RG). OSE-RG is originally an approach to harmonize measures for latent constructs. However, the equipercentile OSE-RG algorithm lends itself well to quantity questions. To test the performance of both algorithms, we gathered experimental data (N = 3484) on the number of books possessed as an example quantity, where we varied the quantity-question response scheme. We show that OSE-RG often outperforms or at least matches MOC when harmonizing closed-ended questions towards an open-ended format, or when harmonizing different closed-ended response formats amongst each other. Notably, OSE-RG is also less susceptible to response biases induced by different close-ended interval response schemes.

Keywords: data harmonization; observed score equating; FAIR data; comparability

1 Introduction

Survey questions about manifest, objective quantities are a staple of many survey programs: Income and age, number of people living in a household, or participants' physical characteristics such as height or weight. Such quantity questions appear very straightforward. After all, their object of interest is something concrete and observable unlike subjective questions about attitudes or values for example. Nonetheless, capturing reliable and valid numerical estimates of quantities in surveys is a complex endeavor that has engendered many different question designs. The goal is usually to balance different desirable features of quantity questions, such as facilitating accurate memory recall, reducing socially desirable responding, or reducing respondent burden (Tourangeau et al., 2000). In our paper, we want to tackle the challenge of how to harmonize data on the same quantity that were measured with differently designed questions. Specifically, we address comparability and harmonization between quantity questions with open-ended numeric response format as well as several versions of closed-ended close-ended quantity questions with different numeric intervals as discrete response options.

Harmonization of existing data from different surveys (i.e., ex-post harmonization) is becoming more and more popular, because it allows us to answer research questions that were harder or impossible to answer with the separate data sources individually (Dubrow & Tomescu-Dubrow, 2016). For example, harmonized datasets allow us to synthesize longer time-series, to increase our sample sizes, or to fill gaps in regional or conceptual coverage. Several past and ongoing harmonization projects in the social sciences document the need for ex-post harmonization (Durand et al., 2021; May et al., 2021; Schulz et al., 2022; Slomczynski & Tomescu-Dubrow, 2018).

Researchers looking to compare or combine data from instruments measuring quantities encounter two types of challenges. The first challenge is different information content. It occurs when asking respondents about quantities with discrete, interval response options. Imagine a response option "30 to 50". Which quantity best represents respondents who choose this response option? And how can we relate this response option to an overlapping but not identical response option in another question (e.g., "20 to 40")? This challenge of different information content is especially pronounced in the highest response interval, which is usually open towards infinity (e.g., "100 or more"). The second challenge are response biases. It is well established that responses to quantity questions are sensitive to biases such as socially desirable responding, but also biases induced by different response for-

Contact information: Ranjit K. Singh PLEASE PROVIDE POSTAL ADDRESS OF CORRESPONDING AUTHOR (E-mail: ranjit.singh@gesis.org)

mats. In other words, we cannot be certain that respondents' true quantities always fall within the verbatim boundaries defined by the response options.

A prominent method to harmonize close-ended quantity questions is the middle of category (MOC) interpolation (Von Hippel et al., 2016). MOC uses the mid-point of a response option as the most representative value for the respondents choosing a given response option. However, harmonizing quantity questions with MOC comes with challenges if a response option is open-ended (e.g. "more than 500 €") and fails to consider response biases in the harmonization process. In this paper we suggest an alternative method: Observed score equating in a random groups design (OSE-RG). OSE-RG is a psychometric method used to align measurement units across different instruments measuring the same latent construct (Kolen & Brennan, 2014). While well established in psychometric educational testing, applying OSE-RG to latent concepts in the social sciences is a recent development (Singh, 2022). In this paper we now argue and empirically demonstrate that OSE-RG can also be applied to harmonizing manifest quantities. Apart from a validation of OSE-RG as a method to harmonize quantity questions, we also demonstrate that OSE-RG has a crucial advantage over MOC. While MOC treats response options labels verbatim, OSE-RG takes the empirical distribution into account as well. This means it can mitigate response biases between different instrument versions. Apart from that, if its preconditions are met, OSE-RG also requires less effort to perform than MOC.

In the following, we will first establish the background of our study, describing the types of quantity questions we will harmonize and discuss the harmonization challenges posed by the different question types. Second, we discuss two approaches to harmonizing quantity questions: The conventional MOC interpolation and the proposed OSE-RG approach. Third, we will present our research design and methods to validate OSE-RG and compare its results to the MOC approach. In the results section we will, fourth, present the empirical results of our validation experiment. Finally, the discussion will lay out the implications for harmonization practitioners seeking to make existing survey data on quantities more comparable.

2 Background

2.1 Quantity questions

In our paper, we define quantity questions as survey instruments which measure a manifest quantity by asking respondents directly about that quantity. Examples are questions about age, income, number of children in the household, bodyweight, or height. Quantity questions have two crucial components that may vary across instruments. Firstly, the question text, which instructs respondents on which quantity to report. Secondly, the response format, with which respondents can indicate their quantity response. In our paper, we focus on differences in such response formats while assuming that the question text is the same across the instruments. We will investigate two response formats: Open-ended quantity questions and close-ended quantity questions. Open-ended quantity questions give the respondents the opportunity to directly report their numerical response. In close-ended quantity questions numeric ranges (e.g., 0–20, 21–50, etc.. . .) are presented as discrete response categories or "bins".

2.2 Comparable measurement units

The many possible ways to construct an interval response format pose a crucial challenge if we want to harmonize questions for the same quantity. At its core, this challenge is about establishing comparable units of measurement. Manifest quantities usually have clear units associated with them: Income measured in units of the local currency, age measured in years, or children measured in integer numbers. However, by measurement units we mean the relationship between measurement values (i.e., scores) in our data and the true quantities they are supposed to represent. And those measured values are fundamentally distinct from true quantities.

Challenge one: Different information content

Different measurement instruments retain a different amount of information about the measured quantity. Close-ended quantity questions only retain the information in which interval respondents fall (e.g., 31 to 50), but discard respondents' intra-interval position (e.g., 34). This information loss poses a hurdle for two common cases in harmonization. The first case is harmonizing a close-ended quantity question with an open-ended quantity question. Here we can either aggregate the open-ended quantities into the interval format, or we can interpolate continuous values for each interval. Aggregating open-ended responses into categories is easy but introduces massive information loss. Interpolating interval quantity categories might thus be preferable but poses the challenge that we need to estimate the expected true quantity for respondents who chose a certain interval. If we had access to the true quantities of respondents who chose an interval, this would be the average quantity of respondents in an interval. However, since the true quantities are unknown, we need to rely on assumptions on the intra-interval distribution to estimate the expected quantity (Von Hippel et al., 2016).

The second case is harmonizing two close-ended quantity questions which have different interval response formats. We thus cannot easily harmonize responses of intervals from different questions where the interval ranges overlap, but are not identical (e.g., 21 to 40 versus 31 to 50). Since intra-interval information has been discarded, we cannot easily determine which portion of respondents falls into the intersection, and which do not. Again, we need to rely on assumptions on the intra-interval distribution to estimate the expected quantity (Von Hippel et al., 2016).

Challenge two: response bias

The second comparability problem arises from response biases. It is tempting to assume that respondents reliably choose the objectively correct response option. However, it is well established that responses can by systematically biased (Tourangeau et al., 2000). Respondents may intentionally choose to misreport their true quantity, but even if they intend to answer truthfully their responses may be subject to unconscious response biases (Tourangeau et al., 2000). There are many different biases established in the literature. As a well-known example, consider socially desirable responding (Paulhus, 2002). For example, respondents often underreport their weight (Polivy et al., 2014) and overreport their physical activity level (Rzewnicki et al., 2003). The key issue here is that response bias means we cannot trust that a respondents true quantity falls into the interval boundaries of their chosen response option.

In our paper, we focus on another response bias: Respondents sensitivity to close-ended quantity questions interval boundaries (Schwarz et al., 1985). Questions with response boundaries that emphasize higher quantities can cause respondents to overreport quantities on average. Vice versa, response boundaries that emphasize lower quantities can cause respondents underreport quantities. This specific response bias was chosen for two reasons. First, it can be manipulated experimentally, which allows us to introduce the bias in a controlled manner into our study. Second, in an experimental design it can be empirically demonstrated.

Establishing comparability

To our mind, the harmonization of quantity questions has three goals: (1) Retaining as much information of the source instruments as possible. (2) Avoiding introducing bias through an inadequate harmonization procedure. (3) Reduce differences between questions that are the result of response biases.

First, retaining information means that we should avoid discarding information by unnecessary aggregation or performing irreversible (i.e., asymmetrical) transformations. In other words, differences in information content should ideally not be solved by reducing the information content of the more finely grained instrument in favor of the more granular instrument. For this reason, approaches such as the lossy aggregation that we briefly discuss in the next section, are suboptimal.

The second and third issue are, in fact, two aspects of the same harmonization issue. In a perfect harmonization, we would like to establish the same relationship between true quantities and measured quantities across the harmonized instruments. As a basic intuition, this would mean that after perfect harmonization, people with a certain true quantity would be represented by the same value across different instruments. However, this ideal is unobtainable because we do not have access to the true quantities of our respondents. This means we cannot compare each respondent's response to their true quantity. It also means that the intra-interval information discarded by interval response formats cannot be easily regained.

Instead, we can borrow an idea from observed score equating. If we apply two instruments to the same population of respondents, we want harmonization to ensure that measurements (i.e., scores) with both instruments to have the same distribution shape. More formally, assume that we applied instruments X and Y to random samples of the same population. What we want is some harmonization function that transforms scores of instrument X towards scores of instrument Y: $h_Y(x)$. After this transformation, the cumulative distribution of the transformed scores of X, $G^*(h_Y(x))$, should be identical with the cumulative distribution of the scores of Y, $G(y)$ (Kolen & Brennan, 2014).

$$G^*(h_Y(x)) = G(y) \tag{1}$$

This observed score equity property might seem abstract at first, but it has very desirable properties (Kolen & Brennan, 2014). If such a harmonization function exists, it would mean that we would get quantity measurements with the same mean, standard deviation, skewness, and percentiles for the same population across different instruments. Differences in measurement units and systematic bias have been aligned. In other words: Measurements are not necessarily free of systematic bias, but biases are at least aligned so that there is no differential bias depending on the instrument used.

3 Harmonization approaches

What are potential solutions for the problems discussed above? We will discuss the MOC approach as a conventional method to harmonize quantity questions and present OSE-RG as a novel approach. It should be noted that there is another common way to harmonize quantity questions that we term lossy aggregation: Aggregating response options to a lowest common denominator between two quantity questions (Esteve & Sobek, 2003; Rolland et al., 2015). If instrument A has the response options 0 to 10, 11 to 20, ... and instrument B has the response options 0 to 20, ..., the first two response options of instrument A could be combined to cover the same interval. However, this approach irreversibly discards information in the process and is only viable if there are matching inner boundaries and passes response biases into the harmonized dataset. Projects using this approach thus often provide the discarded information in separate variables or as separate code digits (Esteve & Sobek, 2003). In the following we will instead focus on MOC and OSE-RG.

3.1 Middle-of-category interpolation (MOC)

An important conventional approach is the middle-of-category interpolation (MOC). As the name implies, MOC focuses on interpolation. Specifically, it aims to assign each response interval a single, continuous quantity value that best represents the average true quantity of respondents who chose this interval (Von Hippel et al., 2016). Unfortunately, we do not have access to the true quantities. Instead, MOC makes certain assumptions (Von Hippel et al., 2016):

1. Respondents with a certain true quantity τ choose, on average, an interval (or "bin") B with upper and lower boundaries $[l_B, u_B]$ so that $l_B \leq \tau \leq u_B$. In other words, respondents answer truthfully and unbiased.

2. Respondents' true quantities in each interval are uniformly distributed as $U_{[l_B,u_B]}$ if $u_B \neq \infty$.

3. Respondent's true quantities in an interval $[l_B, u_B]$, if $u_B = \infty$ are distributed according to a function that has to be defined based on domain knowledge of the measured quantity.

Assumption two is why it is called the middle of category interpolation. The average response lies in the middle of the two category boundaries, because this is where the expected value of the assumedly uniformly distributed quantities lies. An income category "501€ to 1000€", for example, would be replaced with the value "750.50€".

$$\overline{x}_{[l_B,u_B]} = E\left(U_{[l_B,u_B]}\right) = \frac{1}{2}(l_B + u_B) \qquad (2)$$

Assumption three, however, remains a challenge. The middle of a category bounded on one side by infinity is infinity. Instead, practitioners must make assumptions about the shape of the cumulative distribution of the true quantity in the surveyed population (Von Hippel et al., 2016). Based on that assumed distribution shape, we can infer a most representative value for the highest interval. Often, a Pareto distribution is used for this purpose (Von Hippel et al., 2016). The specifics of this process will be demonstrated in the methods section, where we describe the MOC interpolation in the context of our empirical example. In summary, MOC is a plausible approach for harmonization if the assumptions are met. However, assumption one and three are easily violated in an empirical setting.

Assumption one can be violated by response biases which can cause respondents to choose a response interval which does not encompass their true quantity. For example, if we combine data from close-ended quantity questions overestimating the quantity with data from close-ended quantity questions underestimating quantities, then we bake these spurious differences as methodological artifacts into our MOC harmonized data. Assumption three is violated if the true distribution of quantities does not follow the assumed distribution type or if the parameters differ. We require domain specific knowledge to choose an adequate distribution type and to fine-tune plausible parameters. This adds qualified manual work, increases researchers' degrees of freedom, and makes the approach hard to generalize across different quantities or populations.

3.2 A novel approach: Observed score equating in a random groups design (OSE-RG)

As we have seen, there is a need for a new approach to harmonizing quantity data. First and foremost, the new approach should be able to address response bias. Furthermore, it would be ideal to have an approach that works across different quantities with little manual work and few researcher degrees of freedom. OSE-RG is a promising candidate that fulfills those criteria. OSE-RG is an ex-post harmonization method with a long tradition in psychometrics, specifically the harmonization of educational attainment measurements (Kolen & Brennan, 2014). However, OSE-RG is novel in the context of harmonizing survey quantity questions in two ways. First, equating in general is only recently being applied to single-item survey instruments (Singh, 2022), and second, equating is conventionally used to harmonize measurement instruments for latent constructs, not manifest quantities. However, the mechanisms that allow us to harmonize the unobservable (i.e., latent constructs) may also serve us well in harmonizing the unobserved (i.e., true manifest quantities). In the following, we describe OSE-RGs logic and make the case that it can be used to harmonize quantity questions. However, please note that none of the following formulas are applied by hand. Mature software and packages, such as the *equate* package for R, conveniently automate this process (Albano, 2016).

OSE-RG aims to align the (cumulative) score distributions of two instruments X and Y for the same population. The key idea is already implied by the qualifier "for the same population": OSE-RG uses the random groups design, in which we collect data for both instruments in samples randomly drawn from the same population (Kolen & Brennan, 2014). This is equivalent to a split ballot experiment often used in survey methods research. Through this experimental design, we ensure that there are no systematic differences in the true quantity distribution in both samples. However, the measured responses, the observed scores, will have different distribution shapes.

In some harmonization projects, the data already has the suitable format. However, if this is not the case, we can use different instances where the respective quantity question designs were used. Singh (2022) provides a proof of principle for using external data for performing OSE-RG: Either non-probabilistic experimental data in an online access panel to perform OSE-RG or two probabilistic samples of the adult

German population from different survey programs.

If suitable data is found, OSE-RG then simply transforms scores of X so that the response distributions align in shape (Kolen & Brennan, 2014, p. 11):

$$G^{*}\left(\mathrm{eq}_{Y}(x)\right) = G(y) \tag{3}$$

Formula (3) is the same as formula (1), with the only difference that the specific harmonization function $\mathrm{eq}_Y(x)$ is put in place of the general, placeholder harmonization function $h_Y(x)$. Here the bias correcting aspect of OSE-RG comes into play. Aligning the distribution shapes aligns bias, as the distribution shapes of observed scores are the product of both the true quantity distribution but also response bias. The resulting harmonized data is no longer differently biased across instruments.

OSE-RG intuition

While there are different algorithms to align distribution shapes, we focus on the equipercentile algorithm. Equipercentile equating is well suited to harmonizing non-normally distributed scores—such as quantities usually are. In equipercentile equating, we create two functions: (1) A percentile function $P(x)$ which transforms responses of instrument X into linearly interpolated percentile ranks. (2) An inverted percentile function $Q^{-1}(P^{*})$, which transforms percentile ranks P^{*} into their corresponding responses in instrument Y. Then we can express the equipercentile equating function harmonizing responses of X towards the format of instrument Y as (Kolen & Brennan, 2014, p. 36):

$$e_Y(x) = Q^{-1}(P(x)) \tag{4}$$

Where x are scores of instrument X, which are equated towards the scale of instrument Y via an equipercentile equating function $e_Y(x)$. In the following we will show how to use OSE-RG in three common use-cases. This sequence of use-cases also serves to build up the full set of formulas of the equipercentile equating algorithm step by step.

Use-case 1: Harmonization of two open-ended quantity questions

Let us consider this algorithm by harmonizing two open-ended quantity questions X and Y, where X is susceptible to overreported quantities through socially desirable responding, while Y is not. For continuous quantities x in an open-ended question, the percentile function $P(x)$ is nothing else than the cumulative frequency function $100 \cdot (F(x))$. Note that relative frequencies are bounded between 0 and 1, whereas percentiles are bounded between 0 and 100. The algorithm would first transform each reported quantity in X into a corresponding percentile rank. A reported quantity of "15" in instrument x would become $P(15) = 100 \cdot (F(15)) = 33$ meaning that 33% of respondents reported a quantity of 15 or lower on instrument X. Then we would transform this percentile rank $P^{*} = 33$ into the corresponding reported quantity in instrument Y. In other words, we would look for a quantity in y with a percentile rank of 33, using $Q^{-1}(33)$. Empirically, we might find that the response "12" has a percentile rank in instrument Y. We would thus transform a "15" in X into a "12" in Y. Since the true quantity distribution is the same in both experimental samples, we have mitigated the different levels of social desirability bias of X and Y by aligning the percentiles.

Use-case 2: Harmonization of a close-ended quantity question towards an open-ended quantity question

However, the main challenge that we want to address in this paper is harmonizing close-ended quantity questions, both towards open-ended questions as well as other close-ended quantity questions. And here we face a problem: What is the percentile rank of an interval "30 to 45"? Such an interval covers a range of quantities and thus also a range of quantity percentile ranks. To solve this, equipercentile OSE-RG uses linear interpolation. Specifically, it assumes that the percentile ranks for a given response interval are uniformly distributed. Please note that going forward, the formulas assume that we represent the response options with integer values starting with 0. For a given response interval, denoted by an integer score $x \in N_0$, the percentile can be calculated using the relative frequency of a score $f()$ and the cumulative relative frequency of a score $F()$. Formally, $f()$ is the discrete density function for $X = x$ and for our purposes it represents the proportion of respondents who chose a specific score x. $F()$ is the discrete cumulative distribution function, which represents the proportion of respondents who chose a specific score x or a lower score. Then we can calculate an interpolated percentile score using formula 5(adapted from Kolen & Brennan, 2014, p. 42).

$$P(x) = 100\left(F(x-1) + \frac{1}{2}f(x)\right) \quad \text{for } x \in N_0 \tag{5}$$

Imagine an interval "30 to 45", which happens to be the third response option in an instrument X. Thus $x = 2$, because scores start at zero. We can then calculate how many respondent percent chose a lower response option than "30 to 45" as the cumulative frequency of the second response option, $F(x-1)$. Then we add half the percent of respondents who chose the option "30 to 45", $\frac{1}{2}f(x)$. In other words, the response option "30 to 45" covers a percentile interval of $[100 \cdot F(x-1), 100 \cdot F(x)]$ and we choose the middle of that interval, because we assume that percentiles are uniformly distributed in each interval. In other words, we applied a middle-of-percentiles interpolation. As a side note: The actual formula is more complicated, because equipercentile equating can also work with non-integer scores. However, this added layer of complexity is not necessary here.

Use-case 3: Harmonization between two different close-ended quantity questions

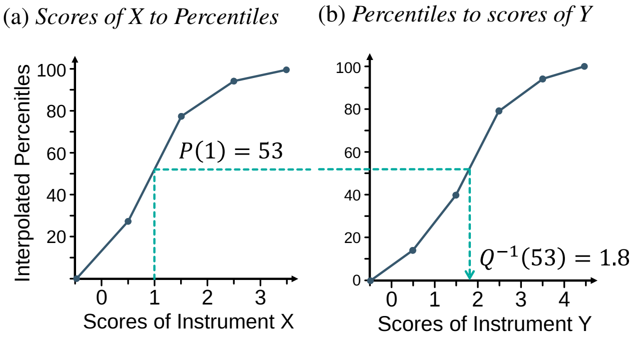
The percentile function $P()$ in equation 5 allows us to harmonize close-ended quantity questions towards open -ended quantity questions. However, if we want to use OSE-RG to harmonize one close-ended quantity question to another close-ended quantity question, we also require a new inverted percentile rank function. Specifically, we need an inverted percentile function $Q^{-1}()$ that can take arbitrary percentile ranks and find a corresponding linearly interpolated, "continuized" response in instrument Y (Kolen & Brennan, 2014). This is necessary, because the interpolated percentiles for the responses in X will not perfectly match the percentiles of the responses of Y.

In Figure 1, the process is explained visually. On the left side, we see the same process as before: A percentile function $P()$ transforming a score of instrument x into its corresponding, interpolated percentile rank of $P(1) = 53$. On the right side, we see the new, interpolated inverted percentile function $Q^{-1}()$. With it, we complete the equipercentile equating process by transforming the percentiles of scores in instrument X into their linearly interpolated equivalent scores in instrument Y. A percentile rank of 53, for example, is transformed into a score in instrument Y of $Q^{-1}(53) = 1.8$. The whole process of equipercentile equating thus is: $eq_Y(1) = Q^{-1}(P(1)) = 1.8$. Also note how in our example, instrument X has four interval response options, while Y has five. OSE-RG can not only harmonize across instruments with different interval boundaries but also across instruments with a different number of intervals.

Analytically, the inverted percentile function $Q^{-1}()$ in equipercentile equating looks like this (Kolen & Brennan, 2014, p. 43):

$$Q^{-1}(P^*) = \frac{\frac{P^*}{100} - F(y^*_U - 1)}{f(y^*_U)} + (y^*_U - 0.5) \tag{6}$$

First, we need to find y^*_U, which is the smallest response option with a cumulative frequency $F(y)$ larger than the percentile P^*. This is nothing else than the integer response option that is closes to the interpolated response that we will get as a result (e.g., if the result will be a score between 1.5 and 2.5, then y^*_U is 2). The formula seems daunting, but the logic is very straightforward. If y^*_U is a response option in instrument Y, then it covers percentiles in an interval $[100{\cdot}F(y^*_U - 1), 100{\cdot}F(y^*_U)]$. Thus, a percentile rank of $100 \cdot F(y^*_U - 1)$ corresponds to $y^*_U - 0.5$ and a percentile rank of $100\left(F(y^*_U)\right)$ corresponds to $y^*_U + 0.5$. This allows us to transform percentiles which do not directly match a specific response option in Y into decimal response scores. A response score of $Q^{-1}(P^*) = 1.5$ would mean that the percentile falls halfway between the percentiles of the second (1) and third (2) response option. Remember, in the formula, scores start at zero, which means 0 is the fist, 1 the second, and 2 the third response option. Lastly, please note that the equating process is perfectly symmetrical (Kolen & Brennan, 2014). Transforming scores of X into the format of Y is an arbitrary choice. We could just as easily transform scores of Y into the format of X.


(a) *Scores of X to Percentiles*

(b) *Percentiles to scores of Y*

Figure 1

Equipercentile OSE-RG from one close-ended quantity question to another. The figure illustrates on the left how scores of instrument X are transformed into interpolated percentiles with $P(X)$ and then how those interpolated percentiles are transformed into equated scores of instrument Y with $Q^{-1}(P(X))$. The solid lines visualize the transformation relationship between percentiles and scores. The dots on that line are the relative frequencies of scores 0, 1, 2 etc. The dots are placed at the 0.5 boundary between scores, since the relative frequency of a response option shows the highest percentile where respondents still most likely choose that option: in other words the upper bound of a response option and not its middle.

3.3 Study design

To validate OSE-RG and show its advantages compared to MOC, we designed a survey experiment. In an online access panel, we presented the same quantity question to respondents, but we randomly varied the response formats. Specifically, we presented respondents with a total of four response formats: an open-ended response format, or one of three different interval response formats. The three interval quantity response formats were designed so as to induce response bias. One interval format emphasized low quantities, one medium quantities, and one high quantities. This setup already allows us to demonstrate response bias by comparing the low, medium, and high interval quantity responses amongst each other as well as to the open-ended responses.

As a quantity to measure, we chose the number of (print and electronic) books respondents possessed. The number of books is not very sensitive, which avoids drop-out. We also wanted to focus on the close-ended quantity question induced response bias and not a global socially desirable responding bias. The number of books is also easily understood but not trivial to recall or estimate. This ensures conceptual comparability, while leaving room for bias in estimation.

We will first demonstrate response bias to illustrate why a new approach is necessary. Then we will apply OSE-RG to harmonize the three close-ended quantity question variants (low, medium, high) towards the open-ended quantity question format as a proof of principle. Here, we will also compare the OSE-RG solutions to the one MOC provides. Specifically, we aim to show that MOC retains the full response bias, whereas OSE-RG mitigates such differences in bias. Lastly, we will use OSE-RG to transform the low, medium, and high interval quantity not towards the open-ended format, but instead into each other. And again, we compare OSE-RGs performance with that of MOC.

4 Methods

4.1 Participants and procedure

We conducted an online experiment with a nonprobability sample recruited via the commercial online access panel of the respondi AG (Respondi AG, 2022). To ensure demographic variability we used quotas for sex and age. Please note that respondents choosing the intersex response option "divers" were added to the female quota. This is because we knew from earlier access panel samples that there would be a negligible number of cases (in this study, three).

To ensure the robustness of our findings derived from a nonprobability sample, we repeated each of the analyses in the results section for different subgroups. Regarding sex, we split the sample into male versus female (discarding the three intersex respondents). Regarding age, we split the sample into three age segments: Younger than 30, 30 to 59, and 60 or older. Regarding educational attainment, we split the sample into respondents with or without higher education entrance qualification (i.e., with or without "Fachabitur or Abitur"). Lastly, we split the sample into respondents from the old or new federal states (formerly West- and East Germany). In the results section, we only report analyses for the full sample, since the same pattern of results was found in all subgroups. Of course, the number of books reported differs between subpopulations, but the relative performance of the harmonization approaches is the same. We provide the subgroup analyses as an electronic supplement.

A total of 3497 respondents participated in the experiment. However, since 13 respondents failed to answer the quantity question, analyses are based on a sample of N = 3484. Of these respondents in our analysis, 49% reported their sex as female, 51% as male, and three individual respondents reported their sex to be "divers" (i.e., intersex in the terminology of the German statistics office). The mean age was 45 years with a standard deviation of 15. Ages ranged from 18 to 87. The sample was highly educated with 54% of respondents reporting some form of higher education entrance qualification.

Table 1

Percentiles of responses to the open-ended quantity question.

	Percentile						
	1th	5th	25th	50th	75th	95th	99th
Books	0	2	30	80	200	1000	2565

Respondents first read the study introduction, including information about what data we gather and to which purpose. After giving informed consent, respondents answered demographic questions about sex, age, education, and which federal state they live in. Then, respondents answered a question for another experiment. Specifically, they were asked about being annoyed by advertisements. Then, respondents were randomly assigned to one of our quantity question versions. The online questionnaire continued with other studies, but we omit describing them because they had no impact on our experiment. Respondents had completed our experiment after a median time of 97 seconds with an IQR of 65.

4.2 Quantity questions

We asked respondents about how many books they possess: "How many books do you possess? We mean both print books and e-books. Books which you share with other people in your household are also included."

Open-ended quantity question

In the open-ended quantity question condition, respondents could answer the question with a text input field: "I possess [____] books." The resulting answers exhibited a long tail of infrequently reported, very high quantities (Table 1). The maximum reported quantity was 6666 (further analysis of the open-ended response options results can be found in the appendix). In the results section, we will thus compare untrimmed and trimmed analyses. Where applicable, we also explore the median instead of the mean. However, please note that there is no sure way of knowing which responses are unrealistic. Some respondents may indeed possess very many e-books.

Close-ended quantity questions

In the three close-ended quantity question conditions, respondents had four discrete response options to choose from. Responses started with the highest interval and ended with the lowest. For example, on the medium quantity condition, the options were in order: "more than 100", "51 to 100", "26

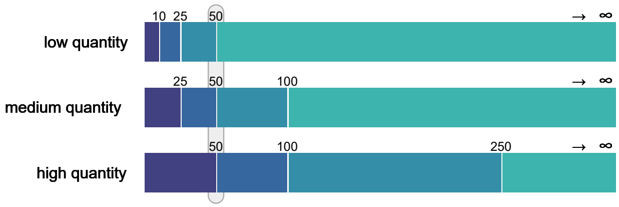


Figure 2

Boundaries of the close-ended quantity questions

Table 2

Distribution of respondents reporting to own 50 books or less by close-ended quantity question asked.

Question	Percentage of respondents with ≤ 50 books
Low quantity	45
Medium quantity	41
High quantity	33
Open question	44

to 50", and "25 or fewer". Response options were presented in a vertical layout. Please note that for ease of interpretation, we have inverted the scores in all following analyses so that 1 represents the lowest quantity interval and 4 the highest. In Figure 2, we see the interval boundaries of the three interval quantity conditions at a glance.

Please note that all three conditions intentionally share 50 books as one of their interval boundaries. This allows us to demonstrate response bias in an intuitive fashion by comparing the portion of respondents who report owning 50 books or fewer in each condition. Table 2 shows the results including a comparison with the open-ended quantity question in the last row.

We immediately see a striking bias between the close-ended quantity questions. In the low quantity condition, 11 percentage points more respondents claim to own 50 books or less compared to the high quantity condition. The medium quantity condition, meanwhile, is in between. This difference in percentages must be the result of bias, because respondents were randomly assigned to each condition. Thus, the true portion of respondents with 50 books or fewer should not vary between conditions. To quantify the bias analytically, we calculated Spearman's rank correlation between the ordinal close-ended quantity question conditions (low, medium, high) and a binary variable with 0 representing 50 or fewer reported books and 1 representing more than 50 reported books. The result, $\rho_{\text{Spearman}} = 0.32$; $p < 0.001$, shows a medium sized effect of the response scale on the number of reported books. This finding supports our claim that we cannot interpret the numerical boundaries defined by the response option labels verbatim.

4.3 Harmonization procedures

MOC

To perform MOC interpolation, we applied the formulas reported by Von Hippel et al. (2016). Each interval B (as in "bin") has a lower bound l_B, an upper bound u_B, and is populated by n_B respondents who chose interval B. The bounds are verbatim interpretations of the response option labels. For a given bound, we can calculate its MOC interpolated value as:

$$\text{MOC}(B) = \begin{cases} \frac{1}{2}(l_B + u_B) & \text{if } u_B \neq \infty \\ l_B \frac{\alpha}{\alpha - 1} & \text{if } u_B = \infty \end{cases} \tag{7}$$

In the latter case, $u_B = \infty$, we have to assume a distribution shape for the true quantities. We chose a Pareto distribution, since its cumulative distribution shape fits the empirical distribution of the open-ended quantity question well. While the formula is straightforward, it requires us to estimate a parameter $\widehat{\alpha}$. This is conventionally done by using the last two bins, i.e., B and $B - 1$.

$$\hat{\alpha} = \frac{\ln(n_{B-1} + n_B) - \ln(n_B)}{\ln(l_B) - \ln(l_{B-1})} \tag{8}$$

However, the formula often results in unsuitable α estimates with $\alpha \leq 1$. This leads to nonsensical MOC results. An α of one results in undefined values, and α lower than one result in negative quantities. If we apply the formula to our empirical example, we do indeed get such unsuitable α estimations of 0.47, 0.66, and 0.76 for the low, medium, and high quantity interval conditions. Thus, we follow the advice of Von Hippel et al. (2016) and use a plausible α of 2. This would mean that the highest category is interpolated with a value twice as high as its lower bound. A response option "50 or more" would be interpolated with 100, for example.

In Figure 3 we illustrated a point that Von Hippel et al. (2016) also stress: The parameter α and its estimation is crucial for MOC. As values of α approach 1 from above, the estimated quantities for the highest interval rise very quickly. In the area of $\alpha = 2$, the value we have chosen, the relationship between parameter and estimator is less volatile.

Equipercentile OSE-RG

OSE-RG was performed using the *equate* package (Albano, 2016). We chose the equipercentile equating algorithm, since it is well suited for non-normal distributed scores. This is crucial for quantities, with their distributions being compressed on one side by a lower bound of zero and stretched on the other side in a long tail towards rare but very high quantities. Note that the equate package uses the same formulas from Kolen and Brennan (2014), which we summarized in the theory section.

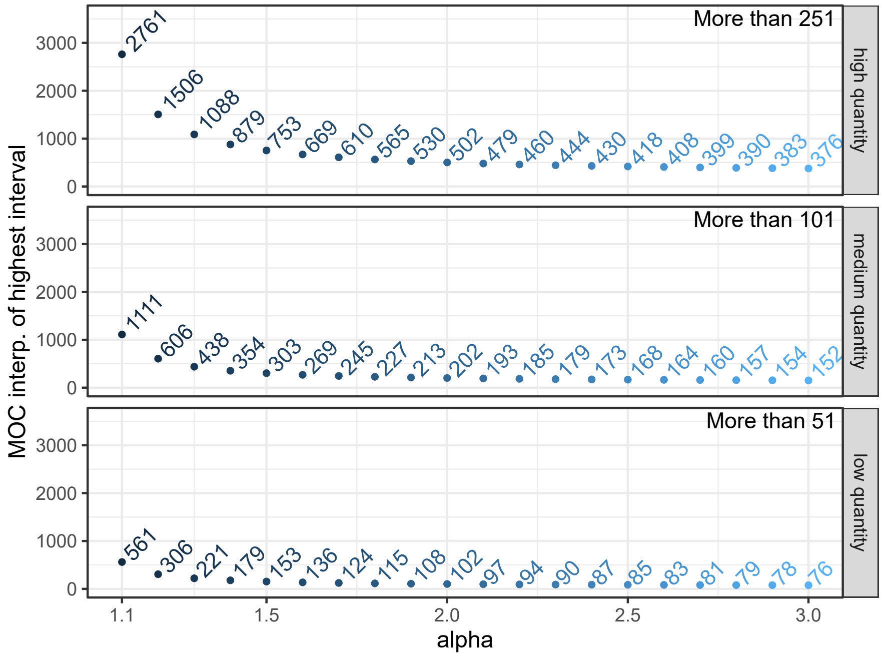


Figure 3

Effect of choosing different values of α when estimating the midpoint of the highest response interval using the pareto distribution.

Statistical software

All data transformations and analyses were conducted in R (R Core Team, 2021)using RStudio (RStudio Team, 2021). All original datasets were in SPSS format and read into R using haven (Wickham & Miller, 2021). The tidyverse package collection (Wickham, 2017) was used for data transformation and data visualization. Additional packages used for data manipulation and visualization are broom (Robinson et al., 2023), knitr (Xie, 2021), viridis (Garnier et al., 2021), Cairo (Urbanek & Horner, 2021), ggrepel (Slowikowski, 2021) and kableExtra (Zhu, 2021). Equating was conducted with the equate package (Albano, 2016).

5 Results

We will first compare the results of OSE-RG and MOC when harmonizing close-ended quantity questions towards an open-ended format and evaluate its harmonization performance. Then, we will harmonize different close-ended quantity questions amongst each other and evaluate the harmonization performance of OSE-RH and MOC.

5.1 Harmonizing the close-ended quantity question towards the open-ended quantity question

Next, we harmonized the three close-ended quantity question versions towards the open-ended question format. Specifically, we interpolated the close-ended quantity questions with MOC ($\alpha = 2$) and equated the close-ended quantity questions towards the open-ended question. Figure 4 shows the results graphically.

Each panel shows the harmonized values for one of the three close-ended quantity question versions: blue plus for MOC and orange X for OSE-RG. Please note that the y-axes have different scales. The solid line in green, meanwhile, are the mean of the open-ended responses in each interval. For an interval $[51, \infty]$, this means selecting all respondents in the open-ended quantity question condition with at least 51 books and then calculating the mean. However, the mean for the open-ended question is susceptible to the influence of outliers. Thus, the dashed green line shows the average open-ended quantity response after trimming the top five percentiles. Lastly, the dotted green line shows the median open-ended question within each interval. It becomes clear that both the MOC approach and the OSE-RG approach work similarly well for the lower three intervals of all three close-ended quantity questions. Here it also does not matter whether we aggregate open responses within each interval with the mean, the trimmed mean, or the median. However, the solutions deviate considerably for the highest intervals, which are open to infinity. Both MOC and OSE-RG tend to underestimate the mean open-ended quantity in the highest intervals. This underestimation is reduced but persists when we trim the mean. When we aggregate open-ended re-

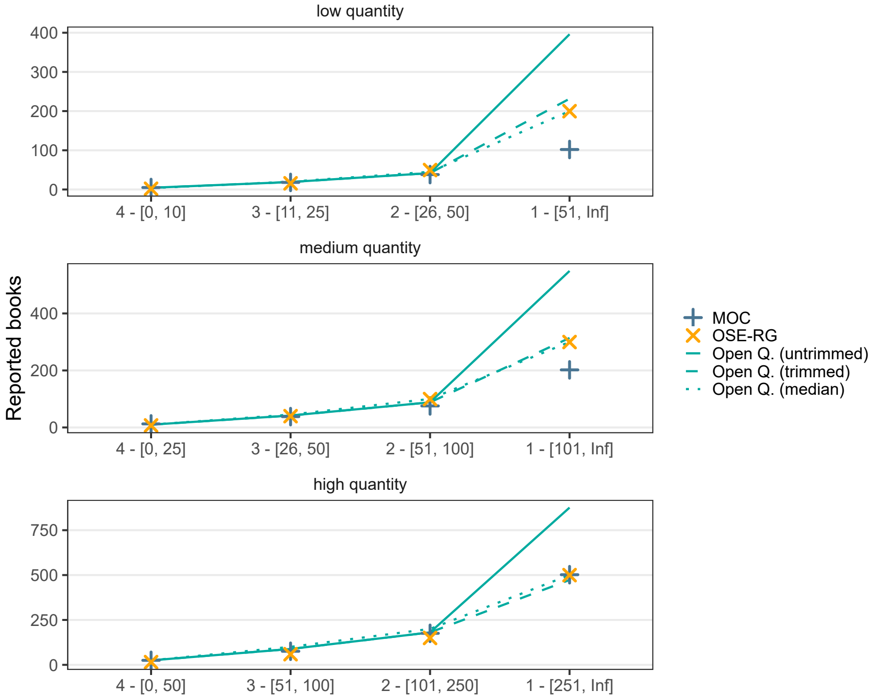


Figure 4

Harmonized number of books at each response option of the close-ended quantity questions.

sponses with the median, however, OSE-RG shows an almost perfect fit. This is unsurprising, because both equipercentile equating and the median are percentile based methods. If we compare MOC and OSE-RG, we see that OSE-RG performs better than MOC in two out of three conditions (low and medium quantity) when estimating the highest interval across all three approaches to aggregating the open-ended responses.

Evaluating harmonization performance

However, what does that mean for the overall quality of our harmonization efforts? To approach this question, we calculated the mean number of books estimated by MOC and OSE-RG for each close-ended quantity question condition. This means we replaced the close-ended response scores with their OSE-RG and MOC equivalent quantities. Then we calculated separate arithmetic means of those estimated quantities for each approach in each condition. We then calculated the difference of this estimated value to the average quantity measured with the open-ended question. In Figure 5, we see the results in two panels. The transparent trendlines behind these data points serve to illustrate the broader relationship across conditions. For better interpretability, we report this difference as a deviation from the open-ended question mean in percent. The upper panels show the results for untrimmed quantities. In the panel below, we again trimmed responses by removing the top five percentiles. It is important to note, that we did not just trim the open-ended questions, but also the close-ended quantity questions for a fair comparison.

The Figure illustrates several important points. First, depending on the approach used, the interval response format and whether or not data was trimmed, deviations can be very substantial. Harmonizing quantity questions is very sensitive to methodological choices. Second, trimming the data reduces the differences considerably. This is unsurprising, given the very long tailed open-ended quantity distribution. Third, OSE-RG approximated the open-ended quantity question mean more closely than MOC, except for the high quantity condition in the untrimmed data. Fourth, the overall trend-lines illustrate that the estimated mean quantities vary far more across close-ended quantity question conditions after MOC interpolation than after OSE-RG equating. This is consistent with the idea that OSE-RG is less susceptible to different close-ended quantity question formats than MOC. However, please recall figure 4 that clearly showed that all differences are most likely due to the highest intervals only.

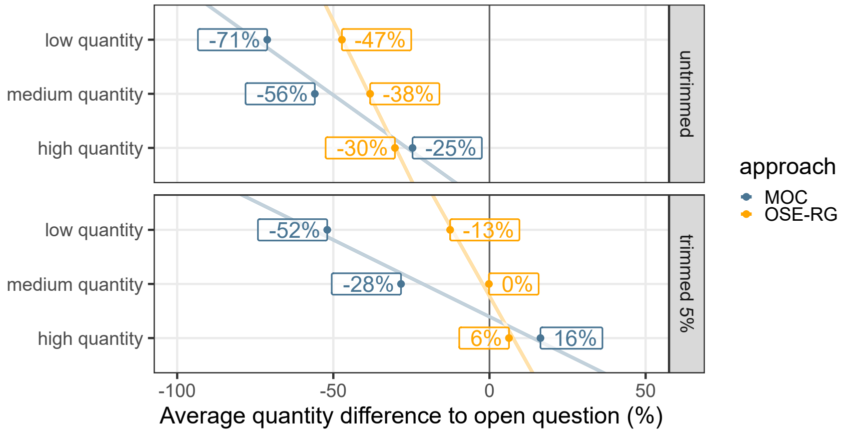


Figure 5

Comparing the approximated mean number of books measured by the close-ended quantity questions with the mean number of books in the open-ended quantity question. The difference to the mean is presented in percent of the mean of the open-quantity question. A lower absolute value represents a better harmonization. The analysis is shown with untrimmed data and with data where the top 5% of the data are trimmed.

6 Harmonizing interval to close-ended quantity questions

Next, we harmonize two close-ended quantity questions with each other. Specifically, we will harmonize the low quantity and high quantity versions towards the medium quantity version. Harmonizing close-ended quantity questions with each other using OSE-RG can be easily done. However, we need to consider the data structure to understand the OSE-RG outcome. In datasets, the different interval categories are represented as integer scores. For example, in the medium quantity interval condition, [0, 25] is a “1”, [26, 50] is a “2”, [51, 100] is a “3” and [101,∞) is a “4” in the dataset. If we harmonize one such close-ended quantity question towards the format of another close-ended quantity question, the result becomes interpretable in the format of the chosen reference instrument. This means the output are decimal equivalents in the format of the integer scores of the reference instrument. If we apply OSE-RG to then harmonize the low and high quantity condition responses towards the medium quantity question, we receive the following transformed values listed in Table 3.

The implications of these transformations become clear if we plot them. Figure 6 shows the intervals and their relative positions to each other after OSE-RG harmonization. Each box represents a response interval in one of the three close-ended quantity question conditions. In the middle, the medium quantity intervals, correspond exactly to the integer scores one to four, because it is the target scale. Left and right, we have the high and low quantity conditions. Note how OSE-RG assigns different scores to the same verbatim interval [51, 100] in the high quantity condition and the medium quantity condition. This is no mistake, but instead a correction for the percentile differences introduced by response bias.

Table 3

Numerical equivalents obtained by OSE-RG when harmonizing the low and high quantity question into the format of the medium quantity question.

Condition	Interval	Original score	OSE-RG towards medium quantity
low quantity	[0, 10]	1	0.75
	[11, 25]	2	1.33
	[26, 50]	3	2.19
	[51, ∞]	4	3.77
high quantity	[0, 50]	1	1.29
	[51, 100]	2	2.67
	[101, 250]	3	3.62
	[251, ∞]	4	4.21

We compared how well OSE-RG aligns the average quantity as compared to MOC. All three conditions represent random samples of the same population. Thus, after harmonization, we would expect no mean difference at all, since the mean true quantity should also be identical across conditions. Specifically, we transformed the data in all three conditions, twice. Once with OSE-RG and once with MOC. The we cal-

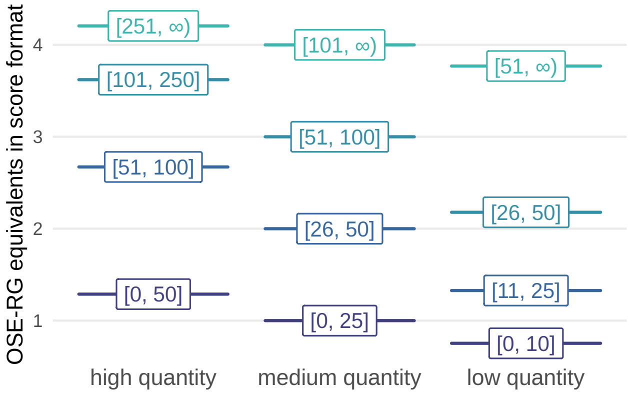


Figure 6

Numerical equivalents obtained by OSE-RG when harmonizing the low and high quantity question into the format of the medium quantity question. The medium quantity values are untransformed because the medium quantity instrument serves as the equating reference. The intervals in the boxes show the associated labels.

culated the mean and standard deviation for every condition with both approaches. Please note that the MOC results have to be interpreted in number of books, whereas the OSE-RG results have to be interpreted in the numerical score format of the medium quantity reference instrument. To compare harmonization success across these two diverse outcome formats, we calculate the standardized mean distance Cohen's d of the harmonized low and high conditions from the medium quantity condition. Since Cohen's d is standardized by the standard deviation of the medium quantity question, it is comparable across different units. Note that we use the standard deviation of the medium quantity reference instrument to standardize mean differences. Table 4 lists the results.

Note how OSE-RG has aligned both mean and standard deviation almost perfectly, whereas MOC results in distribution parameters that vary strongly across conditions. Note that after MOC interpolation, the low quantity condition underestimates the mean quantity by $d = -0.44$ and the high quantity condition overestimates the mean quantity by $d = 0.91$. OSE-RG, in contrast, has a negligible mean bias. Please also note that the choice of reference instrument did not change the outcome. If we choose the low quantity condition as our reference, the d mean biases in OSE-RG are 0.00, 0.01, and 0.04 for the low, medium, and high conditions respectively. If we chose the high quantity condition, the d mean biases in OSE-RG are 0.00, 0.00, and 0.00 for the low, medium, and high conditions respectively.

7 Discussion

We showed that Observed Score Equating in a Random Groups Design (OSE-RG) can be used to harmonize measurements of the same quantity but with different response formats. This includes harmonizing discrete close-ended quantity questions towards an open-ended quantity question format as well as harmonizing close-ended quantity questions amongst each other. We compared this novel approach to MOC which is commonly used. We found that one of the advantages of OSE-RG is that it can mitigate response bias differences between different response formats. We tested this using four response formats: An open-ended response format, and three interval response formats emphasizing low, medium, and high quantities respectively.

Our analyses show that both MOC and OSE-RG do approximate the open-ended question format. However, the approximation is only robust for response options with a finite upper bound. The last interval response options which are open towards infinity are challenging to estimate. OSE-RG performed better here than MOC. However, this also depends on the parameters chosen for the MOC distribution. It should be noted that both approaches underestimate the mean response to the open-ended question format. However, this is mainly due to a long tail of very high quantities reported in the open-ended question. If we truncate data by the top five percentiles, the harmonization becomes far better. And if we calculate the median, OSE-RG is an almost perfect fit. The analyses also demonstrated that OSE-RG was less influenced by the interval response format then MOC. This pattern emerged both when harmonizing close-ended quantity questions towards an open-ended format and when harmonizing different close-ended quantity questions amongst each other.

Note that the paper focuses on evaluating OSE-RG, not MOC. Thus, our examples were designed to draw out potential advantages of OSE-RG. This should not be taken to imply that MOC or interpolation techniques in general should be avoided. The main issue with MOC is the choice of distribution and parameters for estimating the highest response interval, which is open towards infinity. However, in many use cases, the distribution shape and its parameters can be estimated by drawing upon external data sources. Consider harmonizing a common quantity variable, such as income, in surveys with random samples of the adult population of a country. In such cases, we might supply distribution parameters based on official statistics for that country. And lastly, the response biases introduced by different response interval formats only occurs if the survey presented a close-ended quantity question to respondents. However, many instances of categorical quantity data are the result of data producers or archives synthetically binning responses from an open-ended format into a categorical interval scheme. This is often done to protect respondents' anonymity, for example.

Table 4

Comparing the results of OSE-RG and MOC harmonization when harmonizing into the format of the medium quantity question.

				Medium quantity reference		
Approach	Condition	Mean[a]	Std. Dev.	Mean	Std. Dev.	Diff.
OSE-RG	low quantity	2.79	1.16	2.76	1.17	0.03
	high quantity	2.76	1.16	2.76	1.17	0.00
MOC	low quantity	67.07	39.84	102.53	80.21	−0.44
	high quantity	175.32	183.02	102.53	80.21	0.91

[a] In the case of OSE-RG the target format is the score level (1-4), in the case of MOC it is the number of books given by the MOC harmonization.

In such cases, different interval formats do not introduce response bias. However, other biases such as socially desirable responding can still be an issue.

OSE-RG, meanwhile proved to be a harmonization approach at least on par with the more traditional MOC. In fact, OSE-RG may be less sensitive to interval response formats than MOC. Although, we stress that differences between the MOC and OSE-RG solutions only arose in the last response categories, which were intervals open to infinity. In projects where the preconditions of OSE-RG can be met (especially the random groups design), applying OSE-RG is easier than MOC and introduces fewer researcher degrees of freedom. After all, we neither have to choose a distribution type nor tune its parameters. Instead, the percentile-based approach of equipercentile OSE-RG can approximate any cumulative distribution shape.

Harmonization with OSE-RG has a straightforward workflow:

1. Obtain random samples of the same population for each differently designed quantity question (for example via split-ballot experiments, or via existing probability-based survey samples of the same population in the same timeframe).

2. Define one question as the target question (i.e., the reference format) and the other(s) as the source question(s).

3. Transform values of the source question(s) in such a way that the interpolated percentile ranks match across questions.

4. Derive a recoding table, which lists the original integer codes of the source question(s) and the corresponding transformed decimal equivalents in the format of the target question.

5. Apply that recoding table to recode source question values in the current dataset or other instances where the question(s) have been applied. (OSE-RG harmonizes instruments, and not just the current dataset, meaning that the RG data to perform OSE-RG does not have to be identical to the datasets we want to harmonize.)

Step three to five can be automated with software (e.g. the R package equate (Albano, 2016)). The only remaining entry hurdle of OSE-RG is the random groups design. We need samples of both instrument variations drawn randomly from the same population. This seemly restricts OSE-RG to a small number of use cases, where the data we want to harmonize happens to adhere to the random groups design. However, the actual restriction is less severe. OSE-RG makes a distinction between the equating sample and the harmonization samples. After all, OSE-RG was conceived to harmonize two instruments once (using an equating sample) and then to apply this harmonization result in many other instances where those two instruments were used. Thus, we only need some dataset that adheres to the random groups design. This dataset need not be identical to the data we want to harmonize for our research. Thus, could collect affordable equating data in a non-probability setting through experimental variation (as we did). Alternatively, even if our data of interest does not adhere to the random groups design, it is quite possible that the instruments were used elsewhere. Often, instrument designs are taken over from large-scale survey programs. If those instrument source surveys sample the same population, we can use their data to equate our instruments. This might mean using data from two national surveys, or to use the national subsamples of international surveys.

If data in a random groups design cannot be obtained with any of the described approaches, then you might consider equating approaches that attempt synthesize a common population by taking other variables into account as covariates. The idea is that if we have access to covariates explaining the systematic differences between the groups in the source datasets for both instruments, then we can use the covariates to relate the two instruments to each other (Bränberg & Wiberg, 2011). These newly emerging approachers are discussed under the term non-equivalent groups with covariates (NEC) design in the recent literature. There exist R packages which can accommodate NEC designs (González & Wiberg, 2017). However, NEC approaches depend on having access to relevant covariates (Bränberg & Wiberg, 2011) and introduces new assumptions. Specifically, NEC assumes that the conditional response distribution given the covariates is the same in both populations (Wiberg & Bränberg, 2015). In other words, the covariates have to relate similarly to the instruments of interest in both populations. This might preclude the use of NEC in cross-national settings, for example.

7.1 Limitations and future research

We made a first attempt to use OSE-RG as a harmonization approach for manifest quantities, and our proof-of-principle study was successful in this regard. However, our findings cannot yet be generalized across all quantities, quantity question designs, and populations. Future research is certainly necessary to explore the possibility space and ensure robustness. As a concrete next step, it might be fruitful to apply OSE-RG alongside MOC to quantity variables in existing survey programs which use probability-based samples. Ideally, this would involve quantities for which detailed population information exists. Applying OSE-RG in probabilistic data and for quantities where external distribution information exists also allows us to move beyond mere linear interpolation within each interval. With a known population distribution, we could instead use this distribution shape to transform intervals into percentiles and percentiles back into interpolated interval scores. A related point is to explore interval quantity schemes with more finely grained response options. It may well me that MOC performs more robustly when the intervals, and especially the last category, get narrower. Lastly, we should extend OSE-RG to quantity questions using vague, subjective quantifiers, such as “sometimes” or “often”. Here, OSE-RG might be of particular importance because subjective quantifiers preclude the use of MOC.

References

Albano, A. D. (2016). Equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*, 1–36.

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, *48*(4), 419–440.

Dubrow, J. K., & Tomescu-Dubrow, I. (2016). The rise of cross-national survey data harmonization in the social sciences: Emergence of an interdisciplinary methodological field. *Quality & Quantity*, *50*, 1449–1467.

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*, 1–43.

Esteve, A., & Sobek, M. (2003). Challenges and methods of international census harmonization. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *36*(2), 66–79.

Garnier, S., Ross, N., Rudis, R., Camargo, A. P., Sciaini, M., & Scherer, C. (2021). Viridis(lite)—colorblind-friendly color maps for r. sjmgarnier/viridis: CRAN release v0.6.2 (v0.6.2CRAN). https://doi.org/10.5281/zenodo.5579397

González, J., & Wiberg, M. (2017). *Applying test equating methods*. Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer New York. https://doi.org/10.1007/978-1-4939-0317-7

May, A., Werhan, K., Bechert, I., & Quandt, M. (2021). ONBound-harmonization user guide (Stata/SPSS), version 1.1. https://www.gesis.org/fileadmin/upload/dienstleistung/forschungsdatenzentren/IUP/ONBound/ONBound_Users_Guide_v1.1.pdf

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun & D. Jackson (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Routledge.

Polivy, J., Herman, C. P., Trottier, K., & Sidhu, R. (2014). Who are you trying to fool: Does weight underreporting by dieters reflect self-protection or self-presentation? *Health Psychology Review*, *8*(3), 319–338.

R Core Team. (2021). R: A language and environment for statistical computing [R Foundation for Statistical Computing]. https://www.R-project.org/

Respondi AG. (2022). Access panel. https://www.respondi.com/access-panel

Robinson, D., Hayes, A., & Couch, S. (2023). Broom: Convert statistical objects into tidy tibbles. https://broom.tidymodels.org/

Rolland, B., Reid, S., Stelling, D., Warnick, G., Thornquist, M., Feng, Z., & Potter, J. D. (2015). Toward rigorous data harmonization in cancer epidemiology research: One approach. *American journal of Epidemiology*, *182*(12), 1033–1038.

RStudio Team. (2021). RStudio: Integrated development for R. https://www.rstudio.com/

Rzewnicki, R., Auweele, Y. V., & De Bourdeaudhuij, I. (2003). Addressing overreporting on the International Physical Activity Questionnaire (IPAQ) telephone survey with a population sample. *Public health nutrition*, *6*(3), 299–305.

Schulz, S., Weiß, B., Sterl, S., Haensch, A.-C., Schmid, L., & May, A. (2022). Harmonizing and synthesizing partnership histories from different research data infrastructures: A model project for linking research data from various infrastructure (HaSpaD). https://doi.org/10.7802/2317

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public opinion quarterly*, *49*(3), 388–395.

Singh, R. K. (2022). *Harmonizing single-question instruments for latent constructs with equating using political interest as an example.*

Slomczynski, K. M., & Tomescu-Dubrow, I. (2018). Basic principles of survey data recycling. *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*, 937–962.

Slowikowski, K. (2021). Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'. R package version 0.9.1. https://rdrr.io/cran/ggrepel/

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

Urbanek, S., & Horner, J. (2021). Cairo: R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output. https://CRAN.R-project.org/package=Cairo

Von Hippel, P. T., Scarpino, S. V., & Holas, I. (2016). Robust estimation of inequality from binned incomes. *Sociological Methodology*, *46*(1), 212–251.

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

Wickham, H. (2017). Tidyverse: Easily install and load the "tidyverse". R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Wickham, H., & Miller, E. (2021). Haven: Import and export SPSS, Stata and SAS files. R package version 1.1.2. https://CRAN.R-project.org/package=haven

Xie, Y. (2021). Knitr: A general-purpose package for dynamic report generation in R. https://rdrr.io/cran/knitr/

Zhu, H. (2021). Kableextra: Construct complex table with kable and pipe syntax [manual]. https://rdrr.io/cran/kableExtra/