MISS Measurement Instruments for the Social Sciences

PsychOpen GOLD

Advances in Methodology

Check for updates

Biased Bivariate Correlations in Combined Survey Data Measured With Different Instruments

Ranjit K. Singh [1]

[1] *Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany.*

Measurement Instruments for the Social Sciences, 2024, Vol. 6, Article e11217, https://doi.org/10.5964/miss.11217

Received: 2023-01-25 • **Accepted:** 2024-06-03 • **Published (VoR):** 2024-08-14

Handling Editor: Beatrice Rammstedt, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author: Ranjit K. Singh, GESIS – Leibniz-Institute for the Social Sciences. Department: Survey Design and Methodology, P.O. Box 12 21 55, 68072 Mannheim, Germany. E-mail: ranjit.singh@gesis.org

Supplementary Materials: Code, Data [see Index of Supplementary Materials]

Abstract

Social scientists increasingly form composite datasets using data from different survey programs, which often use different single-question instruments to measure the same latent construct. This creates an obstacle when we want to run analyses using the combined data, since the scores measured with different instruments are not necessarily comparable. In this paper, we explore one consequence of such comparability problems. Specifically, we examine the case where instruments measuring the same construct have different item difficulties. This means if we applied the instruments to the same population, we would get different mean responses. If such mean differences are not mitigated before combining data, we introduce a mean bias into our composite data. Such mean bias has direct consequences for analyses based on the combined data. In data drawn from the same population, mean bias introduces error variance. In data drawn from different populations it would bias or even invert true population differences. However, in this paper I demonstrate that mean bias can also bias bivariate correlations if one or both variables in a composite dataset are subject to mean bias. If differences in item difficulty are not mitigated before combining data, we introduce a variant of Simpson's paradox into our data: The bivariate correlation in each source survey might differ substantially from the correlation in the composite dataset. In a set of systematic simulations, I demonstrate this correlation bias effect and show how it changes depending on the mean biases in each variable and the strength of the underlying true correlation.

Keywords

harmonization, comparability, survey methodology, integrated data analysis, Simpson’s paradox, pseudocontingencies

Surveys in the social sciences often use single-question instruments to measure latent constructs, such as attitudes, values, interests, or emotions (Tourangeau et al., 2000). Furthermore, the same construct is often measured with different instruments in different surveys (Tomescu-Dubrow & Slomczynski, 2016). Instruments might differ in their wording, response option labels, number of response options, or other design aspects. This instrument diversity is challenging when we want to combine data from different survey sources to be used in a joint analysis (Singh, 2020; Tomescu-Dubrow & Slomczynski, 2016). And such so called ex-post harmonization projects research projects are becoming increasingly common: From international comparative research (Dubrow & Tomescu-Dubrow, 2016; Durand et al., 2021; May et al., 2021), to integrative meta-analyses (Hussong et al., 2013), to research projects integrating national data on specific substantive topics (Schulz et al., 2022).

To mitigate comparability issues due to instrument diversity, such research projects must employ ex-post harmonization techniques (Granda et al., 2010). In other words, researchers have to carefully select, prepare, transform, and combine source data to create a homogeneous target data set. For example, when researchers aim to harmonize single-question instruments for the same latent construct, they have to ensure that the instruments do in fact measure the same construct and they have to assess the reliability (i.e., robustness against random error) of each instrument, to avoid introducing bias through attenuation into their harmonized dataset (Kolen & Brennan, 2014; Singh, 2022).

However, even if two instruments were perfect, error-free measures of the same construct, the researchers still face the challenge of aligning the units of measurement (i.e., scales) of their different instruments (Kolen & Brennan, 2014; Price, 2017). Instruments tend to differ in the numerical scale with which they represent a construct in the source data. This is not an error, per se. Latent constructs have no natural units and we can use arbitrary scales to represent latent construct intensity (i.e., respondents’ positions on a latent dimension) numerically (Price, 2017). This is easiest to see when we compare instruments with a different number of response options (i.e., scale points). If we measure the same population with a four-point scale or an eleven-point scale, we will most likely measure a higher average response and standard deviation with the eleven-point scale. This is because we scale (or map) the same construct intensities onto a different numerical scheme. However, the number of scale-points is only one factor of many. The measurement units also depend on the question wording, the response labels, the visual layout or any number of other design characteristics (Price, 2017; Tourangeau et al., 2000).

This paper aims to demonstrate how insufficient harmonization efforts can cause substantive and complex bias in our subsequent analyses using the combined data.

Specifically, we explore the last link in a chain of biases. If we do not properly align measurement units (i.e., scales) of different measurement instruments before combining the data, we often incur a mean bias in the combined data (Kolen & Brennan, 2014). With mean bias I mean that two instruments applied to the same population would result in systematically different average response scores. This is problematic, because it might introduce spurious population differences and needless error variance in analyses with the combined data. However, in this paper I set out to demonstrate that such mean biases have another, less intuitive consequence: They also bias correlations based on the combined data. Using a three-dimensional matrix of simulated bivariate correlations with varying degrees of mean bias, the paper sounds out the extent of this bias. Since the matrix of simulation varies mean bias in both variables separately, as well as the underlying unbiased correlation, we can also explore how the resulting mean bias depends on these three factors. In sum, the result of these simulations informs harmonization practitioners about the potential extent and shape of this often-overlooked form of bias in combined (survey) data.

Mean Bias

A substantive problem caused by incomparable and insufficiently harmonized measurement units is mean bias (Kolen & Brennan, 2014; Singh, 2022). Mean bias is best understood when we consider the following thought experiment. Imagine applying two instruments X and Y to sufficiently large random samples A and B of the same population. If ex-post harmonization was successful, we would expect the mean responses to be approximately equal: $\overline{X} \approx \overline{Y}$. This is because the average true score should be the approximately the same in two random samples of the same population: $\overline{\tau}_A \approx \overline{\tau}_B$ (Kolen & Brennan, 2014; Singh, 2020). However, without adequate ex-post harmonization, we might find that the average response to two congeneric instruments for the same construct differs by some constant d: $\overline{X} = \overline{Y} + d$ (Price, 2017; Raykov & Marcoulides, 2011). In other words, combining data across the two instruments introduces a mean bias d. This can easily occur, if two instruments for the same construct have different item difficulties (Moosbrugger & Kelava, 2012). Respondents may find one instrument wording easier to agree to then the alternative and thus for the same population of respondents would choose a higher mean response on one instrument than the other.

In practical terms this means that even after (insufficient) harmonization, an average respondent for our measured population would be represented by different numerical scores in our combined data (Kolen & Brennan, 2014). Of course, mean bias can also occur if instruments X and Y are applied to different populations. However, in such cases we cannot easily isolate the bias for single-question instruments, because the difference between $\overline{X}$ and $\overline{Y}$ is then a composite of the true construct difference $\overline{\tau}_A - \overline{\tau}_B$ and the bias. Again, in practical terms, this means the mean differences between the populations are either over- or underestimated by an amount proportional to the mean bias (Kolen & Brennan, 2014).

For a concrete example, consider two very similar measures of political interest. In Germany, the International Social Survey Programme (ISSP) is fielded together with the German General Social Survey (ALLBUS). In 2014, both asked the respondents about their level of political interest. The ALLBUS asked “How strongly are you interested in politics” with a five-point scale (GESIS-Leibniz-Institut Für Sozialwissenschaften, 2018), and the ISSP asked “How interested would you say you are in politics?” with a four-point scale (ISSP Research Group, 2016). Due to the different number of response options alone, we would expect a different mean response. And indeed, the mean response differed by Cohen’s $|d| = 0.69$ (Singh, 2022). To align these differences in scale points (and thus scale range), harmonization practitioners then often apply linear stretching (Cohen et al., 1999; de Jonge et al., 2017; Durand et al., 2021). This means that the scale ranges (i.e., maximum score minus minimum score) are aligned, by setting the minimum responses and the maximum responses as equal across instruments and then stretching all scores equidistantly in between. In our example, we would linearly stretch the scores 1, 2, 3, 4, of the four-point ISSP instrument to 1, 2.33, 3.67, and 5. However, after this transformation we still find that the average responses differed by $|d| = 0.38$ between the two instruments (Singh, 2022). As it turns out, the two instruments differed in more than their scale range. Additionally, we find that both instruments have different item difficulties of $P = 43$ for the ALLBUS instrument and $P = 33$ for the ISSP instrument (Singh, 2022). In other words, the average respondent chose a score that was 43% along the range from 1 to 5 in the ALLBUS instrument but chose a score that was only 33% along the range from 1 to 4 in the ISSP instrument (Moosbrugger & Kelava, 2012; Singh, 2022). However, linear stretching only aligns the scale ranges but not the position of the average response within the scale range. Thus, differences in item difficulty between two instruments remain untouched and can cause a substantive remaining mean bias when data from the two instruments are combined.

Such differences in difficulty between single-question instruments can be mitigated if we apply more suitable harmonization method than linear stretching. One example is observed score equating in a random groups design: OSE-RG (Kolen & Brennan, 2014; Singh, 2022). Alternatively, such difficulty differences can also be mitigated via multiple imputation. However, in this paper, we want to explore what happens if we fail to mitigate mean bias. Or in terms of our example, what would have happened if we had only used linear stretching. After all, harmonization practitioners might be unaware of the limitations of linear stretching. Or they may find the more suitable harmonization techniques unfeasible in their projects. After all, both approaches have data requirements that are not always easy to meet. OSE-RG, for example, requires random groups data; that is samples for both instruments drawn randomly from the same population. Harmonizing two instruments for the same construct with multiple imputation, meanwhile, requires a calibration sample (Siddique et al., 2015): That is a sample in which each respondent answered both instruments, but in a way that does not lead to question order effects.

The first consequence of mean bias in composite data is straightforward: Under mean bias, scores derived from different instruments are biased by an additive constant. In data drawn from the same population, this introduces error variance. And even more worrisome, in data from different populations measured with different instruments, the mean bias is mingled with true population differences. Thus, we can no longer be certain that if two populations differ on average, that this is a true population difference. Instead, it could be a methodological artifact.

Biased Bivariate Correlations due to Mean Bias

In this paper, however, I will demonstrate with systematic simulations that bivariate correlations between two variables in a harmonized dataset can be biased as well, if one or both variables are subject to mean bias. This might seem surprising, because Pearson product-moment correlations are unaffected by linear transformations of variable scores. Specifically, an additive constant *d* would not change the correlation coefficient *r*, because adding a constant to each score changes the arithmetic mean with the same constant (adapted from Gill, 2008):

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i + d\right) = \overline{x} + d$$

Thus, the Pearson product moment correlation formula (Gill, 2008) remains unchanged by an additive constant applied to all values of x (or y):

$$r_{xy} = \frac{\sum_{i=1}^{n}\left[\left(x_i + d\right) - \left(\overline{x} + d\right)\right]\left(y_i - \overline{y}\right)}{s_x s_y} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{s_x s_y}$$

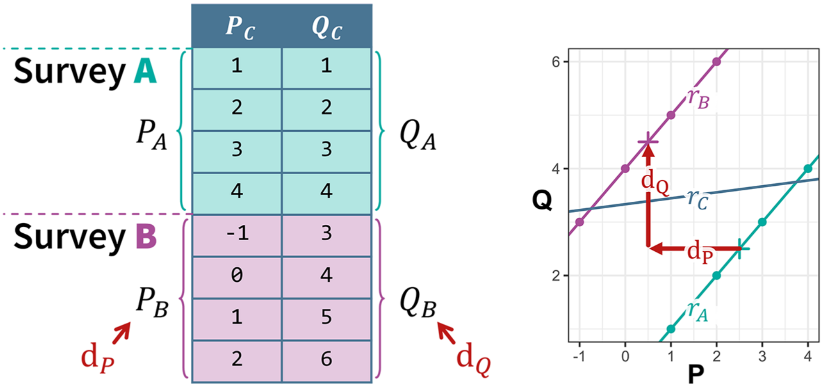
However, this intuition is misleading, because in harmonization we combine data from different sources. In our combined dataset, a vector of responses for a construct would thus be composed of scores derived from different instruments. Consequently, mean bias does not add a constant to the whole variable, as above. Instead, mean bias adds different additive biases to different segments of the combined variable. The argument above obviously no longer holds if we add a constant d to some x_i, but not all.

Imagine the following simplified combined dataset with data from surveys A and B. Both surveys measured the constructs P and Q. However, both surveys used different instruments for each construct, thus leading to a total of four instruments. If we combine these data, we arrive at the following combined dataset structure: a dataset with two variables, one for construct P and one for construct Q. Crucially, each construct variable (i.e., vector) is a composite of values from two surveys and thus two instruments. In summary, there are thus three vectors for construct P: P_A from survey A, P_B from survey B, and P_C as the composite vector in the joint dataset with data from two different instruments combined. Analogously, for construct Q there are then the vectors Q_A, Q_B, and Q_C.

This composite structure of variables in the combined dataset is crucial for understanding the impact of mean bias. After all, mean bias introduced by instrument differences does not affect the whole composite variable. Instead, it would be as if we added a constant to only half of the composite variable. Figure 1 shows the data structure schematically on the left.

Figure 1

Schematic Overview of a Composite Dataset and the Resulting Simpson's Paradox



P_C	Q_C
1	1
2	2
3	3
4	4
-1	3
0	4
1	5
2	6

In this example, we assume that the scale of instrument P_B assigns scores that are on average $d_P = -2$ lower than the scores that instrument P_A would assign for respondents with the same true score in construct P. Instrument Q_B, meanwhile, assigns scores that are on average $d_Q = 2$ higher than the scores that instrument Q_A would assign for respondents with the same true score in construct Q.

If we now plot this combined dataset but differentiate by source surveys (and thus source instruments) we observe a surprising pattern in Figure 1 on the right. Within the data from each survey, constructs P and Q are perfectly correlated with $r_A = r_B = 1.0$ (pink and green trend lines). However, if we correlate P_C and Q_C across the combined data, the correlation drops to a mere $r_C = .11$ (blue trend line).

What has happened? The conditional correlations in two groups are identical, but the correlation across both groups is substantially different. Through the mean bias in P and in Q, we have introduced a version Simpson's paradox into our combined data (Rücker & Schumacher, 2008). In general, Simpson's paradox describes an empirical pattern where we observe the same relationship between two variables in two (or more) groups separately, but a substantially different relationship in analyses across the groups. Of course, Simpson's paradox is not an actual paradox. Instead it is "a form of bias, resulting from heterogeneity in the data [that has not been] accounted for" (Rücker & Schumacher, 2008). If we decompose the problem, we see that the overall correlation r_C is a composite of the individual source survey (and thus source instrument) correlations, r_A and r_B, on the one hand, and of the spurious correlation introduced by the mean biases (d_P and d_Q in red). To visualize this mean biased induced competing correlation as a trendline through the two bivariate group means in surveys A and B (the plus signs on the trend lines of surveys A and B). In other words, if we do not account for mean bias, ideally by removing it with a suitable harmonization procedure, we bias correlations by $\Delta r = r_C - r_u$.

This simple example showed that, in principle, mean bias between source instruments can result in biased correlations. However, under what circumstances does this correlation bias occur and with which intensity? In this paper, I set out to map the landscape of this bias with a series of systematic simulations. Specifically, the simulations will demonstrate that the correlation bias due to mean bias in composite data is determined by the interaction of three factors. First, by the mean bias in the composite variable P_C. Graphically, in Figure 1 above, we would shift the data of survey B left or right. Second, by the mean bias in the composite variable Q_C. Graphically, we would shift the data of survey B up or down. Third, by the strength of the unbiased correlation r_u between the constructs P and Q. By unbiased correlation, I mean the correlation we would expect in the absence of mean biases: $r_u = r_C | d_P = d_Q = 0$. Graphically, lower or higher values of r_u would mean that the datapoints of A and B would fluctuate more or less around the diagonal trendlines for each survey.

By systematically varying all plausible combinations of those three factors, we can map out plausible boundaries of the mean bias induced correlation bias Δr. With these simulations, I aim to provide practical insights into the following questions:

1. How large is the maximum potential bias for plausible mean biases $-1 \leq d \leq 1$?
2. How do different combinations of mean biases d_P and d_Q impact the correlation bias?
3. How different unbiased correlations r_u impact the correlation bias?
4. Under which conditions are absolute empirical effect size $|r_C|$ over- or underestimations of the unbiased absolute effect size $|r_u|$?
5. Can mean bias cause correlations to change direction?

The overarching goal of this paper is to clearly map out the extent and shape of a bias in survey data harmonization that practitioners might not have previously considered. Armed with this intuition, harmonization practitioners can better anticipate the risk of incurring a correlation bias in their analyses if substantive variables composed of different source variables are not adequately harmonized.

Method

Software

All simulations, analyses, and plots were done in R (R Core Team, 2021) within the RStudio IDE (RStudio Team, 2022). The tidyverse package collection (Wickham et al., 2019) was used for data transformation, automation, and data visualization. The pairs of correlated variables were simulated using the faux package (DeBruine, 2021).

Simulation

To answer the research questions, a three-dimensional matrix of simulations was computed. The matrix thus contains simulated estimates of Δr for different bias configurations: i.e., combinations of plausible values of the mean biases d_P, d_Q in the composite variables and unbiased correlations r_u. This bias configuration matrix allows us to systematically map out the resulting correlation biases for all combinations of mean biases and unbiased bivariate correlations. This is crucial because, as we will see in the results, the correlation bias changes drastically depending on these three factors. In the following, I first describe the basic assumptions, the process of generating a single correlation simulation for a single bias configuration, and then the structure of the whole bias configuration matrix.

Simulation Parameters

To clearly isolate the effects of mean bias, the simulation uses continuous, standard normally distributed pairs of variables (i.e., $\overline{x} = 0,\ s = 1$), and a predefined covariance and thus a predefined correlation r_u. These variables were randomly generated using the faux package (DeBruine, 2021). Each simulated vector had a length of 10,000 elements. Since each vector was then duplicated (see next section), each simulation created a combined dataset C with two variables and a sample size $N = 20{,}000$. Mean bias is introduced by adding a constant to every value of a variable. This creates pure mean bias in the sense that the mean is the only distribution parameter that changes. Furthermore, since all source variables have the same standard deviation of $s = 1$, raw mean differences can be directly interpreted as Cohen's d values.

Simulated Mean and Correlation Bias

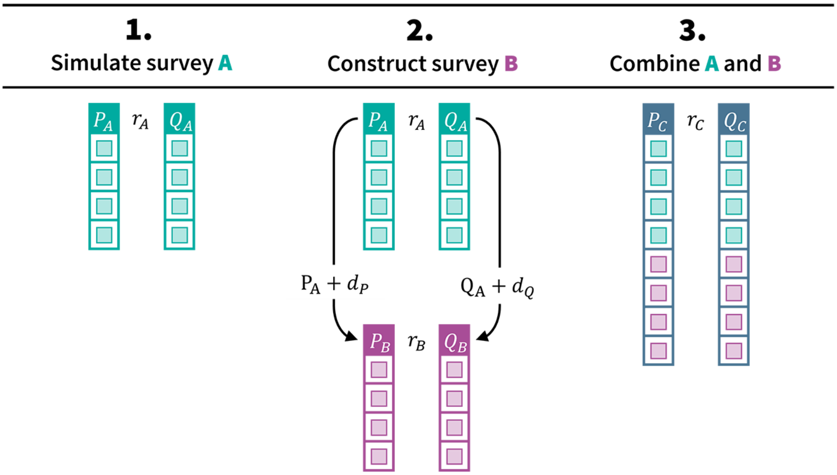
For this paper, I simulate many bivariate correlations with different parameters. Each correlation is determined by three parameters. First, the mean bias d_P in the composite variable P_C, second, the mean bias d_Q in the composite variable Q, and third the unbiased correlation coefficient between constructs P and Q, r_u. Based on those three parameters, a simple harmonized dataset is simulated analogous to the one presented in Figure 1.

PsychOpen GOLD

The algorithm works like this. First, two vectors of simulated data are created. Both vectors contain standard normally distributed continuous values. The vectors exhibit a bivariate correlation that reflects the unbiased correlation r_u that we aim for in this specific simulation. This setup represents data from survey A with variables P_A and Q_A, measured by survey A's instruments for the constructs P and Q. Second, we duplicate the simulated vectors for survey A and modify it with the mean bias parameters d_P and d_Q. This creates a simulated survey B with variables P_B and Q_B, that is identical to survey A except the different measurement unit: For example, a respondent who chose a response x in P_A would have chosen a response $x + d_P$ in P_B. This also ensures that the correlations in each survey are the same and equal to the unbiased correlation we aim for: $r_A = r_B = r_u$. Third, we combine the simulated data for survey A and survey B together to generate a simulated harmonized dataset C. This means we append P_B to P_A to form P_C and we append Q_B and Q_A to form Q_C. Figure 2 below summarizes this process.

Figure 2

Schematic Overview of the Simulation Process for a Specific Combination of d_P, d_Q, and r_u



Based on these combined vectors P_C and Q_C, we can calculate the biased correlation r_C. The correlation bias Δr can then be calculated as the difference between the correlation biased by mean bias in P and Q and the unbiased correlation: $\Delta r = r_C - r_u$. The correlation bias Δr can be interpreted as follows: A positive Δr values mean that the biased correlation is numerically higher than the unbiased correlation and negative Δr values mean that the biased correlation is numerically lower than the unbiased correlation. Please note that this is not the same as the absolute correlation effect size being stronger or weaker. For example, a positive Δr in the context of a strong negative unbiased correlation means that the correlation is weaker (e.g., $r_u = -1;\ r_C = -0.6; \Delta r = 0.4$).

Bias Configuration Matrix

The simulation described above covers one possible bias configuration, defined by the mean bias d_P in Variable P_C, the mean bias d_Q in variable Q_C, and the unbiased correlation r_u. For each such configuration, we get a correlation bias:

$$corrsim\left(d_P, d_Q, r_u\right) = \Delta r = r_C - r_u$$

To systematically demonstrate how the correlation bias depends on the interaction of these three simulation parameters, all three parameters were varied from -1 to 1 with 41 discrete steps (i.e., -1, -0.95, -0.90 … 0 … 0.90, 0.95, 1). This means the simulations in this paper cover the mean bias range $-1 \le d \le 1$ in variables P_C and Q_C and the unbiased correlation range $-1 \le r_u \le 1$. Then, all possible combinations of the three discrete parameter steps were formed. This resulted in a three-dimensional matrix of parameters with $41^3 = 68{,}921$ parameter combinations. Then this parameter matrix was populated with the correlation bias Δr by running the simulation described above for each of these parameter combinations. The end result then was a matrix of correlation biases where each specific simulated correlation bias was determined by a specific combination of the three parameters. Thus, the simulation parameters formed a coordinate system, where a specific bias estimate can be retrieved via its bias coordinate $\left(d_P, d_Q, r_u\right)$. Every data view reported in the results section is thus a specific subset of this matrix of correlation biases.

Results

Correlation Bias as a Function of the Combined Bivariate Mean Bias Strength

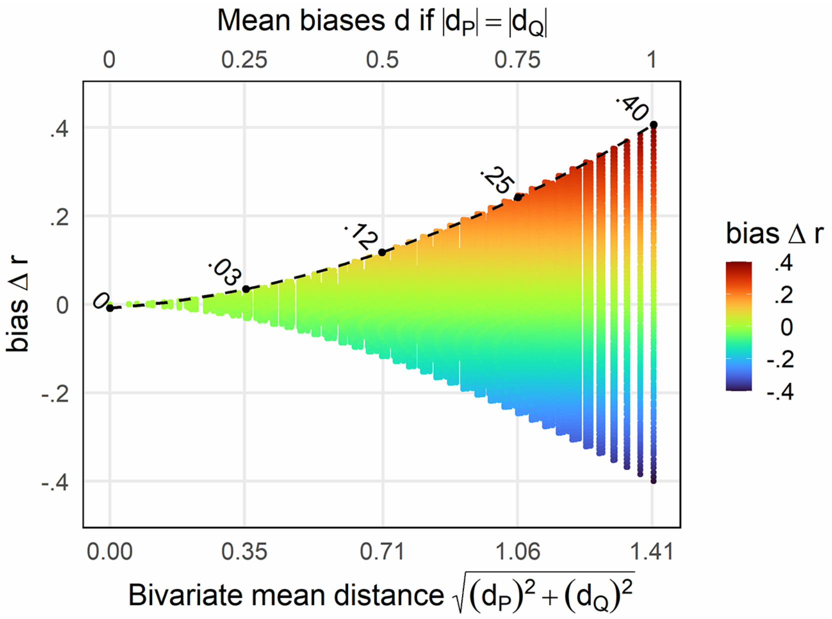
Our first question was: How large is the maximum potential bias for plausible mean biases $-1 \le d \le 1$? To answer this, we can calculate a measure of bivariate mean bias. Specifically, I calculated the Euclidian distance between two points defined by the means of the two variables in survey A and B. In Figure 2 above, this would be the distance between the green and pink plus-signs signifying the bivariate means in surveys A and B. Since the bivariate mean biases are two-dimensional, we can apply the Pythagorean theorem (Gill, 2008) to calculate the Euclidean distance as follows:

$$distance\left(\begin{pmatrix}\overline{P}_A \\ \overline{Q}_A\end{pmatrix}, \begin{pmatrix}\overline{P}_B \\ \overline{Q}_B\end{pmatrix}\right) = \sqrt{\left(P_B - P_A\right)^2 + \left(Q_B - Q_A\right)^2} = \sqrt{\left(d_P\right)^2 + \left(d_Q\right)^2}$$

If both composite variables are subject to a mean bias of $d_P = d_Q = 1$, then the distance would be $\sqrt{2} = 1.41$, for example. Meanwhile, if one mean bias is zero, then the distance is equal to the other mean bias. And, of course, all other combinations of mean biases work as well. For example, if $d_P = 0.5$ and $d_Q = 1$ then the distance is 1.12. Figure 3 below now shows all simulations plotted by their bivariate mean distance (x-axis) and their correlation bias Δr (y-axis). For easier interpretation, I have added two x-axis scales. The bottom scale shows the raw distance measure. The top x-axis, meanwhile, gives an example for the special case where both mean biases have the same absolute value. So a distance of 0.71 can, for example, be the result of the mean biases $d_P = d_Q = 0.5$.

Figure 3

Correlation Bias as a Function of Bivariate Mean Bias



The graph illustrates that the range of possible correlation biases increases as mean biases increase. Specifically, the range of possible correlation biases increases quadratically. To show this, I have selected only the highest correlation biases for each distance and fitted a linear model which regressed the correlation bias on distance and quadratic distance. The black trendline shows the result. The numbers represent the maximum positive correlation bias at selected distances. If both variables have a mean bias of $d = 0.5$, then we would expect a bias range of $-.12 \leq \Delta r \leq .12$, or in other words a span of .22. If both mean biases are $d = 1$, then this range increases to $-.4 \leq \Delta r \leq .4$, or a span of .8.

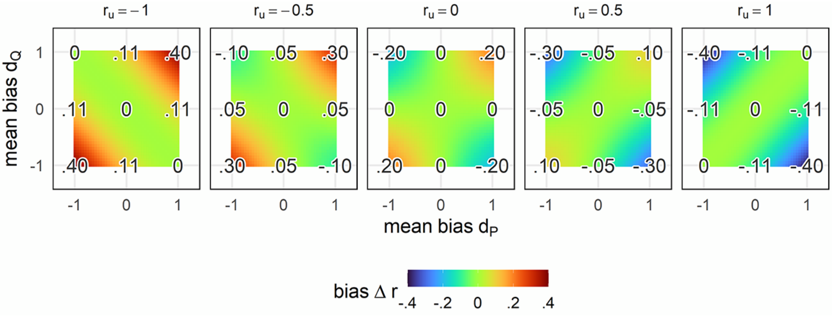
At the same time, the graph intuitively shows that the correlation bias depends on more than just the distance. In fact, even for the highest distance, there are still cases without any mean bias. This is because correlation bias depends on the direction of the mean bias in relation to the direction of the unbiased correlation r_u as well as on the strength of the unbiased correlation $|r_u|$. In the following sections, we will unravel these interactions step by step.

Correlation Bias as a Function of the Two Mean Biases Separately

To get a feeling for the interaction between the mean biases d_P and d_Q in both variables as well as the unbiased correlation r_u, let us consider a series of mean bias grids in Figure 4. Each grid varies mean biases systematically along the x and y axes. The colors meanwhile indicate the correlation bias for each mean bias combination. The five different panels, meanwhile, show the correlation bias pattern for a different underlying unbiased correlation r_u. From left to right, $r_u = -1; -.5; 0; .5; 1$. We can thus interpret each panel as a two-dimensional cross-section of the three-dimensional simulation matrix for a given value of r_u.

Figure 4

Correlation Bias as a Function of d_P and d_Q for Selected Values of r_u



Let us first consider the panel in the middle. Here, the two constructs are uncorrelated with $r_u = 0$. However, mean bias introduces spurious correlations. Since correlation bias is calculated as $\Delta r = r_C - r_u$, we see that if the mean biases are both positive or both negative, then we create a spurious positive correlation. In the extreme cases (the upper-right and lower-left corners), we see that $d_P = d_Q = 1$ and $d_P = d_Q = -1$ result in a spurious correlation of $r_C = .2$. In the opposite case, where $d_P = -1$ and $d_Q = 1$ or vice versa, we see a negative spurious correlation of $r_C = -.2$. In other words, the positive and negative mean bias diagonals are especially prone to creating correlation biases. In contrast, if either $d_P = 0$ or $d_Q = 0$, then no amount of mean bias creates a correlation bias. However, as we will see, this only remains true if $r_u = 0$.

Next, we consider edge cases where the constructs are perfectly correlated. First, the rightmost panel, where the unbiased correlation is perfectly positive with $r_u = 1$. Here we observe the strongest bias if the mean biases align diametrical to the correlation direction. In other words, we observe the strongest bias if $d_P = 1$ and $d_Q = -1$, or when $d_P = -1$ and $d_Q = 1$. At the same time, we observe a positive diagonal corridor that is completely free of correlation bias. If the bivariate mean bias moves along the positive diagonal, it merely aligns with the perfectly positive correlation. Second, leftmost panel, we see the same pattern, only mirrored. Bivariate mean bias along the positive diagonal creates the strongest correlation bias, bivariate mean bias along the negative diagonal results in no bias at all. We also find the same pattern as in the previous section. In the case of perfect correlations, we can realize positive or negative correlation biases of $|\Delta r| = .4$. In both cases, the bias means that the perfect correlation is reduced in strength.

Now consider the second and fourth bias grid, with more realistic unbiased correlations of $|r_u| = .5$. Here, the problem of mean bias creating a bivariate correlation bias reveals its full extent. The only mean bias configuration with a correlation bias of zero is the middle devoid of mean bias ($d_P = d_Q = 0$). Mean biases that move in the opposite direction as the unbiased correlation still results in stronger bias than mean biases that move in the same direction as the unbiased correlation. In these cases, the unbiased correlation is severely underestimated (|Δr|=.3). However, mean biases that move in the same direction as the unbiased correlation also led to bias; this time a bias that overestimates the correlation by $|\Delta \mathrm{r}| = 0.1$.

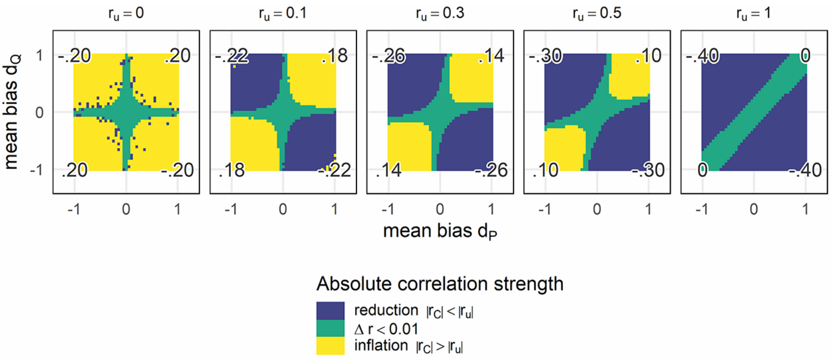
Lastly, we see that the correlation bias intensity seems to vary for different levels of unbiased correlations. However, this is not quite true. In fact, the overall correlation bias range remains a steady $\max(\Delta r) - \min(\Delta r) \approx .4$ along the whole range of unbiased correlations $-1 \leq r_u \leq 1$. It is just that the direction of the correlation bias shifts. Negative perfect unbiased correlations $r_u = -1$ only allow for positive bias values, positive perfect unbiased correlations $r_u = 1$ only allow for negative bias values, while an unbiased correlation of $r_u = 0$ allows for symmetrical biases in both directions. All other unbiased correlation values in between follow a linear pattern defined by these three selected cases. In the Supplementary Materials, I have plotted this dynamic in greater detail in Supplementary Figure A (see Singh, 2024b).

Is the Correlation Under- or Overestimated?

Now that we have gained a better understanding of the correlation bias dynamic, we can address two issues of practical relevance. First, practitioners might wonder which mean bias configurations inflate or reduce absolute correlations. In other words, whether absolute empirical effect sizes overestimate ($|r_C| < |r_u|$) or underestimate ($|r_C| < |r_u|$) the absolute unbiased effect size. In Figure 5 below we see the familiar mean bias grids, but this time colors represent areas of reduced effect sizes in blue, inflated effect sizes in yellow, and areas with negligible bias ($\Delta r < .01$) in green. The unbiased r_u values range from zero to one, because negative values are just mirrored along the x-axis. We see that when $r_u = 0$, then every corner inflates absolute correlations. However, as soon as r_u has a direction, then we have inflation in sectors where the mean bias direction aligns with the correlation direction, and reduction in sectors where the mean bias direction is opposite to the correlation direction. There are two further patterns worth noting. First, as r_u increases in strength, the inflated sector contracts, the reduction sector expands. Second, as r_u increases, bias that reduces the effect size gains in intensity and bias that inflates the effect size loses intensity.

Figure 5

Reduced or Inflated Absolute Effect Sizes



Can Mean Bias Change the Correlation Direction?

A last aspect of practical importance is if mean bias can invert the direction of correlations. Some researchers are more interested in the direction of effects then the absolute effect size. In such cases, it would be fatal if mean bias changed the direction of an effect. And indeed, there are simulations where the direction of r_u is inverted. The minimum absolute distance was 0.57. In other words, in the worst case, an inverted correlation can already occur if both variables are biased with $d = 0.4$. However, such a change in direction can only occur for unbiased r values $|r_u| < .25$ as long as the mean biases remain between $-1 \leq d \leq 1$. Supplementary Figure B shows the mean bias areas with inverted correlations in detail (see Singh, 2024b).

Discussion

When we combine data from different source surveys on the same latent constructs, we might incur a mean bias in our combined data. This usually happens if two instruments have different item difficulties, meaning they assign different positions along their scale range to average respondents for a specific population. In practical terms, if the two instruments were applied to the same population, we would get a higher mean value with one instrument than with the other. Such mean biases in combined data should ideally be removed by adequate harmonization techniques, such as equating or multiple imputation. However, if the combined data has not been fully harmonized, such as data that has only been linearly stretched, then mean biases might remain. Such mean biases are problematic in themselves, of course. If the instruments were applied in different populations, mean bias means that we might over- or underestimate the true population differences. However, the simulations in this paper demonstrate that mean bias can bias can also lead to biased bivariate correlations.

The simulations show that mean bias can lead to substantive correlation biases. For mean biases with $-1 \leq d \leq 1$, we observed correlation biases between $-.4 \leq \Delta r \leq .4$. For a specific unbiased correlation r_u, we observed a range of correlation biases of $|\max(\Delta r) - \min(\Delta r)| \approx .4$. And even if mean biases are weaker, this still implies a worrisome range of biases.

Furthermore, the simulations demonstrate how complex the interaction between the mean biases in each variable and the unbiased correlation is. In some configurations, even very large mean biases lead to little if any correlation biases. However, these cases mainly occur either in cases where the unbiased correlation is zero or where it approaches a perfect negative or positive correlation. In more realistic correlation ranges, all substantive mean bias configuration lead to correlation biases. Even for cases, where only one of the two composite variables is subject to mean bias.

The simulations also revealed two additional patterns of practical interest. First, mean biases can both inflate or reduce the absolute effect size. If we obtain an empirical correlation from a harmonized dataset and suspect that mean bias might be present, this means that the unbiased correlation might be higher or lower than the empirical correlation. Second, if the unbiased correlation r_u is low, then its effect direction may be inverted by some mean bias configurations. Specifically, empirical correlations of $|r_C| < .2$ might misrepresent the unbiased correlation direction if we suspect mean biases up to a range of $-1 \leq d \leq 1$.

Conclusion

The paper has two main practical implications: (1) Wherever possible, mean bias should be mitigated by applying appropriate ex-post harmonization procedures, such as observed score equating (Singh, 2022) or multiple imputation using a calibration sample

(Siddique et al., 2015). (2) Where mean biases cannot be mitigated, correlations based on multi-source data should be interpreted with caution: The direction of small correlations may have been inverted and comparisons of the relative correlation strength across different instruments might be misleading.

Funding: The author has no funding to report.

Acknowledgments: I would like to thank Matthias Roth for showing me the helpful nest-unnest pattern for unfolding combinations of one or more sets into a data frame structure in R.

Competing Interests: The author shares the same institutional affiliation as the Handling Editor/Editor-in-Chief. However, the author declares that this affiliation had no influence on the peer review process or editorial decision.

Mastodon Account: https://fediscience.org/@R_K_Singh

Data Availability: The simulated data as well as all R scripts necessary to replicate the paper are available in the Supplementary Materials (see Singh, 2024a).

Supplementary Materials

For this article, the following Supplementary Materials are available:

- Simulated dataset (simulated_data.rds) with simulated correlations and correlation bias, R code and output (00_simulate_data.Rmd — Reruns the simulation, if desired; 00_simulate_data.html — R code and explanations regarding the simulation; 01_analyses.Rmd — Recreate the plots and some summary statistics; 01_analyses.html — R code and output) (see Singh, 2024a)
- Supplementary Figures A and B (see Singh, 2024b)

Index of Supplementary Materials

Singh, R. K. (2024a). *Supplementary materials to "Biased bivariate correlations in combined survey data measured with different instruments"* [Data, R code, output]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.15217

Singh, R. K. (2024b). *Supplementary materials to "Biased bivariate correlations in combined survey data measured with different instruments"* [Figures A and B]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.15218

References

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. https://doi.org/10.1207/S15327906MBR3403_2 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). Springer. https://doi.org/10.1007/978-3-319-53261-5_1

DeBruine, L. (2021). *faux: Simulation for factorial Designs* (Version 1.1.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.2669586

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*(4), 1449–1467. https://doi.org/10.1007/s11135-015-0215-z

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, 56*, 1755–1797. https://doi.org/10.1007/s11135-020-01088-1

GESIS-Leibniz-Institut Für Sozialwissenschaften. (2018). *ALLBUS/GGSS 2014 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2014)* (Version 2.2.0) [Data set]. GESIS Data Archive. https://doi.org/10.4232/1.13141

Gill, J. (2008). *Essential mathematics for political and social research* (Reprinted). Cambridge University Press.

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology, 9*(1), 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

ISSP Research Group. (2016). *International social survey programme: Citizenship II - ISSP 2014* (2.0.0) [Data set]. GESIS Data Archive. https://doi.org/10.4232/1.12590

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

May, A., Werhan, K., Bechert, I., Quandt, M., Schnabel, A., & Behrens, K. (2021). ONBound-Harmonization User Guide (Stata/SPSS) (Version 1.1). *GESIS Papers.* https://doi.org/10.21241/SSOAR.72442

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

Granda, P., Wolf, C., & Hadorn, R. (2010). Harmonizing survey data. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 315–332). John Wiley & Sons. https://doi.org/10.1002/9780470609927.ch17

Price, L. R. (2017). *Psychometric methods: Theory into practice.* Guilford Press.

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

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

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

Rücker, G., & Schumacher, M. (2008). Simpson’s paradox visualized: The example of the Rosiglitazone meta-analysis. *BMC Medical Research Methodology, 8*(1), Article 34. https://doi.org/10.1186/1471-2288-8-34

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)* (Version 1.0.0) [Data set]. GESIS Data Archive. https://doi.org/10.7802/2317

Siddique, J., Reiter, J. P., Brincks, A., Gibbons, R. D., Crespi, C. M., & Brown, C. H. (2015). Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. *Statistics in Medicine, 34*(26), 3399–3414. https://doi.org/10.1002/sim.6562

Singh, R. K. (2020). Harmonizing instruments with equating. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences, 6*(1), 11–18. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-68262-1

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

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. https://doi.org/10.1080/00207659.2016.1130424

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . .Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), Article 1686. https://doi.org/10.21105/joss.01686