Score ReLiability? Exploring the role of score reliability for effect size heterogeneity

Beinhauer, L.J., Fünderich, J.H., Renkewitz, F.

In light of discussions surrounding the replicability of psychological results, meta-analytic heterogeneity is understood as a crucial parameter that may inform some aspects of (non-)replicability. Heterogeneity in itself helps understand why some replications of a single phenomenon may be successful, while other are not. If heterogeneity grows, meaning the phenomenon’s effect size varies more strongly for unexplained reasons, the probability of observing an effect size around zero or even in the negative space grows larger as well. If we know the true size and heterogeneity of a phenomenon’s effect size, we would theoretically be able to establish a baseline of expected replication rate, given no presence of publication bias or QRPs (reference).

Similarly, it has been argued that heterogeneity in effect sizes is an indicator of the theory’s “completeness” surrounding the phenomenon. Linden and Hönekopp argue that “low (as opposed to high) heterogeneity reflects a more advanced understanding of the subject matter being studied” (2021, p.2). Similarly, Schuetze and von Hippel (2023) argue that heterogeneity in effects is an indicator of a vague, poorly specified theory. With these perspectives, heterogeneity in effect sizes can be understood as an indicator of the extent, to which the science surrounding a phenomenon has evolved to a sufficient understanding of that phenomenon.

Re-analyses of multi-site direct replications have demonstrated that heterogeneity tends to be the norm, if a psychological effect overall manages to replicate. Large-scale attempts of direct replications, using identical protocols, such as the Many Labs studies or Registered Replication Reports (references), for the first time allow researchers to estimate heterogeneity undistorted by typical experimental factors. In re-analyses of these studies, Olsson-Collentine et al. (2020) identify a strong correlation between a studies effect size and its heterogeneity. Similarly, van Erp et al. (2017) or Stanley et al. (2018) estimate strong degrees of heterogeneity across psychological replications in general. In a separate re-analysis of large scale direct replications, Renkewitz et al. (2024) identify substantial heterogeneity in almost all projects where a non-zero effect could be identified. This aligns with the correlation found by Olsson-Collentine et al. (2020).

In both the initial reports of large scale replication attempts, as well as the re-analyses by Olsson-Collentine et al. (2020), standardized effect sizes (ES), such as Cohen’s d or Hedge’s g were used. For the remainder of the article, Cohen’s d, defined in equation [1](#eq-SMD0), will be used as an exemplary estimate of ES, as its used across a wide range of contexts and well understood by a broad audience.

Here, refers to the difference in means between the two groups of interest, while is the total pooled standard deviation.

As is widely known, score reliability affects such ES. In the context of classical test theory (CTT) score reliability is defined as the ratio of true to observed score variance, as defined in equation [2](#eq-alpha).

In this equation refers to the true variance in the sense of CTT, meaning the actual variance of the variable, undistorted by measurement error. In the same sense, refers to the total variance of scores, including both the true variance and the random error variance. In line with Hunter & Schmidt (2014), score reliability is denoted with . Lower score reliability in a data-set leads to a smaller ES, opposed to a similar data-set with identical but higher score reliability (references). In the meta-analytic context, as score reliability is an aspect of a measuring instrument applied to a population, score reliability may not be identical across replications. If score reliability varies across replications, meaning it also carries heterogeneity, it clearly should affect heterogeneity in observed ES as well.

Previous discussions of heterogeneity in score reliability have exclusively discussed it as a parameter that inflates heterogeneity in observed ES, implying that, if score reliability was perfect across all replications, the actual heterogeneity would have been lower. In their discussion, Wiernik and Dahlke claim that „*Measurement error variance will impact the results of meta-analyses in three ways: by (a) biasing the mean effect size toward zero, (b) inflating effect-size heterogeneity and confounding moderator effects, and (c) confounding publication-bias and sensitivity analyses*” (p. 3, 2020). Additionally, more clearly, they state that “*If the studies included in a meta-analysis differ in their measures’ reliabilities, heterogeneity estimates will be artifactually inflated, erroneously suggesting larger potential moderator effects*” (p. 4, Wiernik & Dahlke, 2020). They base their claims largely on he work done by Hunter and Schmidt (2014), who claim that “*Variation in reliability across studies causes variation in the observed effect sizes above and beyond that produced by sampling error.*” (p. 302). Overall, both references imply that differences in score (un)reliability inevitably lead to an inflated heterogeneity in ES.

However, if information concerning score reliability is available, it is possible to correct the individual ES for its unreliability. This process is also known as attenuation correction and, while not without its criticisms (reference), is a widespread practice in Psychology (reference). If the claims above, made by Wiernik and Dahlke (2020) or Hunter and Schmidt (2014), were to be true, it would imply that corrections of individual observed ESs for their unreliability, any subsequently performed meta-analysis should generate lower estimates of heterogeneity. Hunter & Schmidt even claim to have derived an equation which demonstrates how heterogeneity in score reliability inflates heterogeneity in ESs. “*If the level of reliability is independent of the true effect size across studies, then, to a close approximation*:” (2014, p. 309)

Here, refers to the “true” ES, undistorted by sampling error but not corrected for measuring error, while refers to the “true” ES, undistorted by sampling error and measurement error.

## Analytical Arguments

However, we propose that this understanding of how differences in score reliability affect the heterogeneity of observed ESs is incomplete. Hunter and Schmidt’s implication, that heterogeneity in score reliability inflates ES heterogeneity does not follow from equation [3](#eq-HS). If equation [3](#eq-HS) were to be true, it would imply that heterogeneity in is essentially a function in which heterogeneity and mean value of are weighted by mean score reliability and its heterogeneity. In this scenario, it is easy to construct cases where heterogeneity in is inflated or deflated by heterogeneity in . In table 1, two such cases are demonstrated. All else is held equal, only the mean score reliability is changed from .7 to .8. True heterogeneity is .2, at this leads to a smaller , while at this leads to a larger .

Parameter combinations of ES and score reliability, according to Hunter & Schmidt (2014)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 0.2 | 0.7 | 0.1 | 0.198 |
| 1 | 0.2 | 0.8 | 0.1 | 0.228 |

Hunter & Schmidt point out that equation [3](#eq-HS) is only valid “*If the level of reliability is independent of the true effect size across studies […]”* (2014, p. 309). However, both parameters, ES and score reliability in original notation can not, by definition, be independent parameters. ES is the “true” standardized mean difference, unaffected by sampling or measurement error. The ES is standardized using the total standard deviation, which is inflated by measurement error, see equation [-@eg-SMD0]. The ES on the other hand is standardized using the true standard deviation in the sense of CTT.

In equation [4](#eq-SMD), is the true standard deviation, the square root of the true score variance found in equation 2. The fact that the true standard deviation parameter is found both in equation 4 and equation 2 demonstrates that ES and score reliability can not be independent variables. Therefore, the basic premise of equation [3](#eq-HS) is not fulfilled, as the assumption of variable independence could not be fulfilled.

Alternatively, assuming that both the mean differences and pooled standard deviations are no constant values across replications, meaning both carry non-zero heterogeneity, we can describe observed ESs as a random variable following a ratio distribution. Such a ratio distribution is the probability distribution, constructed by dividing one random variable by a second variable: (reference). If the distribution of the components is known, first order taylor approximation may be used to generate estimates of mean and variance of the ratio variable (reference). Using some additional assumptions, we use the Taylor-estimator of the ratio’s variance in [5](#eq-TaylorZ) to demonstrate how differences in score reliability may affect heterogeneity in ES.

Assuming that the random variable are uncorrelated, simplifies equation [5](#eq-TaylorZ) to equation [6](#eq-TaylorZs)

Substituting for ES, unstandardized mean difference MD and pooled standard deviation SD leads to equation [7](#eq-TaylorES0)

Correcting ES for score reliability is technically done by estimating true score standard deviation from total score standardd eviation : .

Since is always between zero and one, in the case of imperfect score reliability, will always be smaller than total . Additionally, if the variance in was introduced by differences in score reliability, these would be removed by the attenuation correction. However, in total, there are two parameters that may introduce variance in across replications.

Firstly, as discussed, differences in score reliability may introduce variation to . Secondly however, the underlying latent variable we attempted to measure may have been not identically distributed across replications. If the populations where the replications are collected from are differently spread out in terms of the variable of interest, true differences in . Re-analyses of the ManyLabs and Registered Replication Reports have shown that such differences in true variance across replications are the norm, not the exception. In that case, some variance in would remain, albeit smaller than the variance in . We reformulate equation [7](#eq-TaylorES0) in terms of

The implication of this equation is, again, not straightforward. The expected value of is guaranteed to be smaller than the expected value of uncorrected . Similarly, the variance in is guaranteed to be smaller than the uncorrected variance in . The expected value is always placed in the denominator of the function, while the variance is placed in the numerator of the second term in equation [8](#eq-TaylorES). This implies that the variance in uncorrected ES can only be larger than the variance in , if the reduction in mean due to the attenuation correction increases the total equation result faster than the reduction in variance of decreases the total equation result. To illustrate this point, we consider some parameter combinations in table 2.

*Table 2*

Parameter combinations, leading to varying degrees of heterogeneity in ES\_c, based on the Taylor approximation to the variance of a ratio distribution

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | .2^2 | 1 | .2^2 | 0.8 | .1^2 | 0.08 | 0.087 |
| 1 | .2^2 | 1 | .2^2 | 0.8 | .05^2 | 0.08 | 0.069 |
| 1 | .2^2 | 1 | .2^2 | 0.8 | 0 | 0.08 | 0.062 |
| 1 | .2^2 | 1 | .2^2 | 0.9 | .1^2 | 0.08 | 0.065 |
| 1 | .2^2 | 1 | .2^2 | 0.7 | .1^2 | 0.08 | 0.123 |
| 1 | .2^2 | 1 | .2^2 | 0.5 | .1^2 | 0.08 | 0.320 |

Table 2 demonstrates that in scenarios, where either score reliability was very high to begin with ( in this case), leading to very little change in to (row 5), or where the remaining variance in was comparably low (rows 3 and 4), the attenuation correction led to a reduction in ES heterogeneity. In most other scenarios, where score reliability was lower (.7 and .5 in rows 6 and 7) or remaining variance in sufficiently large (row 2), the attenuation correction actually increased the heterogeneity in ES. Note that in the table, we did not manipulate expected value or variance of or .

Reformulating the approximation of variance in a ratio variable demonstrates how correcting for differences in reliability might affect the heterogeneity in ES.

Using analytical arguments from a reformulation of equation X we have demonstrated that differences in measuring quality often do not inflate estimates of heterogeneity. Instead, assuming that and are uncorrelated, we argue that in many circumstances differences in measuring quality deflate heterogeneity in standardized mean differences. This could be understood as a form of masking, where differences in measuring quality lead to a seemingly lower heterogeneity than we would have observed without any measurement error. Clearly, these arguments are in stark contrast with claims made by Hunter & Schmidt or Wiernik & Dahlke. To further illustrate this point, we turn towards an exemplary data-set, demonstrating that we can observe the deflation in heterogeneity in empirical data.

## Empirical Example

An exemplary data-set needs to fulfil the following conditions: i) it needs to contain several (direct/close or conceptual) replications of something that can be described as a two-group experimental effect; ii) the dependent variable needs to be measured using an identical scale across administrations, with multiple indicators/items/responses, so that estimates of score reliability can be derived; iii) there needs to be statistically significant heterogeneity in the standardized effect sizes; and iv) there needs to be statistically significant heterogeneity in the score reliability across administrations.

While multi-site direct replication attempts, such as the Many Labs studies or Registered Replication Reports initially spring to mind, none of the studies replicated in these projects fulfil all four conditions. However, the Psychological Science Accelerator (PSA), even though their projects don’t have meta-analytic goals, provide data that may be transformed to fulfil all four conditions sufficiently. Specifically, we examined the data from PSACR002, one of the PSA’s studies concerning interventions and responses to the COVID-19 pandemic (reference). In the PSACR002 study, the authors assessed the effectiveness of reappraisal interventions on negative and positive emotion in participants grouped into four categories (reconstrual intervention, repurposing intervention, passive control, active control). The four groups may be collapsed into a single treatment group (reconstrual and repurposing) and a single control group (passive and active control). In total, Wang et al. (2021) assessed 21,644 individuals from 87 countries/regions. These countries/regions served as the study-level grouping variable for our meta-analyses.

Wang et al. (2021) used several measures which may be promising for this study. Out of four measures they use to assess negative and positive emotions, three should fit conditions ii), ii) and iv); positive and negative emotional responses to ten photographs shown in the experimental trial, positive and negative state emotion after viewing the ten photographs and positive and negative anticipated emotions following the study. Over the following paragraphs, we will briefly describe the study design and measures used that are of interest to us. Details regarding the protocol, stimuli or additional measures used may be found at […].

After randomization to one of four conditions, participants in the two intervention conditions received instructions on how to “change one’s thinking to change one’s emotions”, with more detail pertaining to the respective intervention. In the active control condition, participants were asked to reflect on their emotions, while in the passive control condition participants were asked to react naturally. Afterwards, the participants completed practice trials where they viewed two negative photographs in relation to the COVID-19 pandemic (e.g. people in hazmat suits, exhausted doctors), rated those photos on a Likert-scale from 1 to 5 on two questions: “how negative did the photo make you feel” and “how positive did the photo make you feel” (anchors: 1 = not at all, 5 = extremely). Depending on the condition the individual was randomized to, reminders of the strategy to be used were displayed.

In the experimental trials, the participants viewed 10, similarly negative, photographs of the COVID-19 pandemic and rated their emotions on each photograph using Likert-style responses to the positive/negative questions. After the experimental trials, individuals responded to five negative items from the Differential Emotions Scale on fear, anger, sadness, distrust, and stress. Similarly, they responded to five positive items from the same scale, on hope, gratitude, love, inspiration, and serenity. Averages of the five positive and negative items respectively were used as measures of state emotions after the experimental trials. Similarly, negative and positive anticipated emotions were assessed, by prefacing a question of “In the next week, to what extent, if at all, do you think you will feel each of the following?” and following up with the same items used to assess state emotions. All items or the measures used may be found in the appendix XXX, or at [url…].

*Table 3*

Meta-analytic results concerning uncorrected and corrected standardized effect sizes across 3x2 measures of emotional response

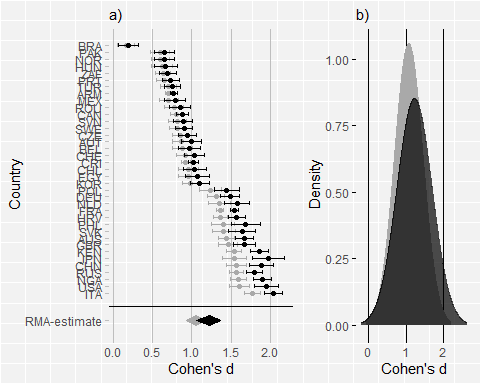
| Measure | Valence | ES | X..tau\_.ES.. | X.ES\_c. | X..tau\_.ES\_c.. |
| --- | --- | --- | --- | --- | --- |
| Photograph Response | Positive | 1.065 (0.065) | 0.377 (sq. 0.036) | 1.219 (0.08) | 0.468 (sq. 0.055) |
| Photograph Response | Negative | -0.663 (0.04) | 0.22 (sq. 0.014) | -0.747 (0.047) | 0.26 (sq. 0.019) |
| Emotional State | Positive | 0.213 (0.018) | 0.052 (sq. 0.003) | 0.234 (0.019) | 0.054 (sq. 0.003) |
| Emotional State | Negative | -0.203 (0.02) | 0.071 (sq. 0.003) | -0.221 (0.022) | 0.079 (sq. 0.004) |
| Anticipated Emotion | Positive | 0.159 (0.015) | 0 (sq. 0.001) | 0.172 (0.016) | 0 (sq. 0.002) |
| Anticipated Emotion | Negative | -0.146 (0.02) | 0.073 (sq. 0.003) | -0.156 (0.022) | 0.079 (sq. 0.004) |

In table 3 the results of meta-analyses on uncorrected and correct standardized effect sizes are summarised. In line with previous analytical arguments and the exemplary table of conditions, we observe that for effects of substantial size, meaning the absolute values of uncorrected ES for both photograph response measures, the heterogeneity in ES is increased by about 25% for the positive measures and about 20% for the negative measures after correcting for attenuation. For all other measures, where the difference in treatment and control groups is less strongly pronounced in terms of mean ES, heterogeneity also increases after correcting for attenuation, but a lot less strongly.

In figure 1 a), the meta-analytical results for the positive photograph response are visualised in a forest plot. Grey dots represent the uncorrected ESs for each country, with accompanying confidence interval represented by the bars surrounding the dots. The overlying black dots on the other hand represent the corrected ESs, with additional confidence intervals.  The position of the grey and black diamonds on the bottom display the meta-analytical estimate of mean ES, uncorrected (grey) and corrected (black), whereas the width of the diamonds represent the respective confidence interval. The forest plot shows that, as expected, the attenuation correction leads to larger ES and larger individual confidence intervals. Similarly, the meta-analytic estimate of mean ES is also larger, if score unreliability was corrected for, just like its accompanying confidence interval.

*Figure 1*

Forest plot and visualised underlying density for positive measures of Photographs



*Note:* In a), grey dots represent individual uncorrected ES with grey bars representing the accompanying 95%-confidence intervals; black dots represent individual corrected ES wit black bars representing accompanying 95%-confidence intervals. In b), the grey density in the background represents the implied underlying distribution of uncorrected ES, while the black density in the foreground represents the implied underlying distribution of corrected ES.

Part b) in figure 2 visualises the implied underlying “true” distribution of ESs, free of sampling error. Essentially, the meta-analytic estimates of mean ES and its heterogeneity are used to construct a density for a normal distribution with these specific parameters. In part b), again, the grey density represents the distribution of uncorrected ES, while the black density represents the distribution of ES corrected for score unreliability. The higher spike in the grey demonstrates that the heterogeneity in uncorrected ES is smaller than the heterogeneity for corrected ES. Generally, the distribution of corrected ES is positioned further to the right, as the average ES is larger, as can also be seen in the forest plot. Additionally, as the density is more “flat” and with a lower spike, it shows that the heterogeneity in corrected ES is larger as well.

## Discussion

Using both analytical arguments as well as an exemplary empirical dataset, we have demonstrated that differences in score reliability do not necessarily inflate heterogeneity in effect sizes. Inferences derived from the equation supplied in Hunter and Schmidt (ref) also do not warrant such a conclusion. Alternatively, describing standardised effect sizes in terms of a ratio distribution and approximating its variance, we have demonstrated that differences in score reliability may very well inflate or deflate variance in ES. In data supplied by the PSA (ref.) we have demonstrated that, across all measures, correcting observed ES for (un)reliability led to larger heterogeneity in ES than before correction.

These results fit in well with work recently published/preprinted by Olsson-Collentine et al. (2023). In a large simulation scheme, they demonstrate that differences in score reliability across administrations initially deflate heterogeneity in correlations. Only as the true heterogeneity in correlations grows larger, score reliability differences actually inflate heterogeneity as predicted by Wiernik and Dahlke (xxx). While correlations and standardized effect sizes are not identical, the way score (un)reliability affects these parameters is highly similar. Both correlations and ES are deflated by (un)reliability, implying that an attenuation correction increases these parameters.

Heterogeneity in ES even in direct replications, where experimental factors are held as constant as possible, is substantial (Renkewitz, Fünderich, & Beinhauer, 2024). Previous assumptions on the role of score reliability differences, as discussed in Wiernik & Dahlke (xxx), has been that they “inflate” heterogeneity. If these previous assumptions were true, it would imply that the true heterogeneity in such projects may be considerably smaller and less alarming.

However, so far we have demonstrated that in several scenarios, differences in score reliability may in fact be masking, or “deflating” ES heterogeneity. This implies that the already substantial heterogeneity across direct replications may very well be even larger, if ES were corrected for (un)reliability. Understanding heterogeneity as an extent to which a theory or phenomenon is understood, heterogeneity that was actually deflated implies that the theory or phenomenon is understood even worse than initially assumed. If the score reliability varies across administrations because the measuring quality is not identical across populations (Beinhauer, Fünderich & Renkewitz, 2024), these differences additionally reflect the poor understanding of phenomenon or theory.

The empirical arguments presented are based on a single study. While, based on the combination of analytical and empirical arguments, we are convinced that these differences in score reliability across replications oftentimes mask true heterogeneity, whether these results actually generalize beyond this data-set remains to be seen. Unfortunately, the behavioural sciences are in dire need of open-data that resembles multi-site replications. As the majority of results discussed over the last years (e.g. ManyLabs or Registered Replication Reports) employ single-indicator scales as dependent measures, score reliability can not be easily estimated in order to replicate our analyses.

The analytical arguments presented in equation X could be improved by properly disentangling the terms concerning variance of SD and expected value of SD. As it stands now, the polynomial structure of the equation makes singling out an individual term rather difficult. However, in order to more easily understand in what way differences in score reliability lead to an inflation or deflation of heterogeneity, solving the equation for both the variance and expected value of SD would be highly beneficial.