Manuscript

# Score ReLiability – Can we explain effect size heterogeneity using score reliability?

Large-scale collaborative replication efforts have sparked discussions surrounding the replicability of psychological phenomena. However, for attempts to estimate a phenomenon’s replicability or to predict whether a future replication will replicate successfully, heterogeneity is a crucial parameter that needs to be assessed. In the meta-analytic context, heterogeneity describes the variation in effect sizes, free of sampling error, if these are not stable or constant across replications. Therefore, the presence of heterogeneity implies that some replications of a single phenomenon may be successful, while others 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 mean size and heterogeneity of a phenomenon’s effect size, we would theoretically be able to establish a baseline of expected replication rate (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.

Re-analyses of multi-site direct replications have demonstrated that, if a psychological effect size can be distinguished from zero, heterogeneity is present across most phenomena (ref). 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 differences in experimental designs. In re-analyses of these studies, Olsson-Collentine et al. (2020) identify a positive correlation between a phenomenon’s effect size and its heterogeneity. Similarly, van Erp et al. (2017) or Stanley et al. (2018) estimate strong degrees of heterogeneity across conceptual replications in Psychology. In a separate re-analysis of large scale direct replications, Renkewitz et al. (2025) 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).

## Effect sizes and score reliability in meta-analysis

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 (from here-on abbreviated as 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-d), will be used as an exemplary estimate of ES, as it is used across a wide range of contexts and well understood by a broad audience.

Here, MD refers to the difference between two groups of interest, while is the total pooled standard deviation. Since constitutes an observed effect size it is affected by sampling error. In terms of heterogeneity, we are not interested in the variance of , but the variance of the true, underlying ES . The standard error of Cohen’s , which is used to weight individual estimates in a random-effects meta-analysis (reference) is computed as defined in [Equation 2](#eq-SEd).

As is widely known, score reliability affects ES. In the context of classical test theory (CTT) score reliability is defined as the ratio of true to total score variance, as defined in [Equation 3](#eq-rel).

In this equation refers to the true score variance in the sense of CTT, meaning the actual variance of the variable, undistorted by random measurement error. In the same sense, refers to the total variance of scores, including both the true variance and the random error variance . This implies that, if true score variance is assumed to be constant across replications, a lower score reliability can only occur due to a larger error score variance. Total score variance would also be larger, therein leading to a smaller ES, as opposed to a similar data-set where a higher score reliability is achieved through a smaller error score variance. However, score reliability is an aspect of a measuring instrument applied to a population. In the meta-analytic context, as studies tend to be replicated in different populations, neither true nor error score variance can be expected to always be identical across replications. This most likely leads to heterogeneity in score reliability, which, as discussed, is bound to affect ES heterogeneity as well.

Previous discussions of heterogeneity in score reliability have exclusively discussed it as a parameter that inflates heterogeneity in 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 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.

As both sources Hunter and Schmidt (2014) and Wiernik and Dahlke (2020) will be referred to repeatedly, we abbreviate these references as H&S and W&D respectively.

### Attenuation correction

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). [Equation 4](#eq-dc) describes a simple attenuation correction procedure.

Here, describes the attenuation-corrected estimate of , which is corrected by dividing it by the estimate of score reliability . As described in W&D or Lord & Novick (references), this is essentially the same, as if the corrected estimate of was constructed using an estimate of true score standard deviation, as in [Equation 5](#eq-dct).

can be estimated by simply rearranging the terms from [Equation 3](#eq-rel). Thereby, an estimate of ES is generated, under the premise that score reliability would have been perfect (). As the corrected estimate of Cohen’s is larger and requires an additional unknown parameter, an estimate of score reliability, the larger uncertainty in this parameter should be reflected in a larger standard error. The standard error for is defined in [Equation 6](#eq-SEdc).

The claims made by W&D imply that corrections of individual observed ESs for their unreliability should lead to lower estimates of heterogeneity in meta-analyses of corrected ES. If heterogeneity in score reliability adds to ES heterogeneity, attenuation correction procedures should eliminate that additional variation, as all corrected ES come with identical (perfect) score reliability.

Re-analyses of the collaborative multi-site replication efforts have uncovered differences in score reliability across replications of the same phenomena (ref McShaw etc.). The claims made in W&D and H&S imply that the ES heterogeneity uncovered in these projects therefore could potentially be explained by these differences in score reliability. The same argument has been proposed by both original authors of phenomena that did not replicate or carry substantial heterogeneity (references) as well as the authors of the multi-site replication attempts (refs). Over the following pages, we assess for the first time, whether the differences in score reliability do indeed explain, and thereby reduce, the ES heterogeneity observed across the different multi-site replication efforts. Even though only a limited number of phenomena allow for such an analysis, we find no evidence for a reduction in ES heterogeneity after taking score reliability into account. Subsequently, we discuss why the expectation formulated in H&S does not allow for such claims. Alternatively, we explore to what extent defining ES as a random ratio variable helps understand how ES heterogeneity is affected by differences in score reliability. We close with a discussion on what future meta-analysts can expect, when dealing with issues of (un)reliability in assessments of meta-analyses of ES.

## Re-analysis of archival data

We have attempted to collect all openly available multi-site direct replications on psychological phenomena (Fünderich et al. 2024). The data-sets of about 50 phenomena , largely stemming from the efforts of the Many Labs studies, Registered Replication Reports or the Psychological Science Accelerator, are made openly available in a standardised format at [osf.io], the DRIPHT repository.

### Data

The ManyLabs projects were collaborative efforts to replicate several psychological phenomena across different research sites, employing identical protocols. From five published projects, the data for the first three and the fifth Many Labs projects was made publicly available when the DRIPHT repository was set up (references). The Registered Replication Reports are similar collaborative efforts, with the sole distinction that for each report a single phenomenon was replicated several times. The Registered Reports 3-10 were added to the DRIPHT repository, as they employed experimental designs with at least two groups. Lastly, some projects from the Psychological Science Accelerator were added to the DRIPHT repository. Different to Many Labs or Registered Replication Projects, these collaborative efforts, while also distributed across the globe, do not formally attempt to perform direct replications, but focus on original research or conceptual replications. However, as for each project, across all sites the same protocol is used, and the subjects are distributed across different countries, a data structure similar to that of a multi-site replication study emerges. What makes all these collaborative efforts valuable for this manuscript is the fact that all phenomena collected in the DRIPHT-repository are making use of two-group designs and are therefore easily assessed using standardised effect sizes.

While replication data on a decent number of psychological phenomena is available in the DRIPHT-repository, to demonstrate how incorporating score reliability affects heterogeneity in ES, selected phenomena need to fulfil three conditions: (1) As we focus on the use of ES , the phenomenon needs to be assessed using a two-group design. This is the case for all projects in the DRIPHT repository; (2) Score reliability estimates can be derived, this implies that the phenomenon needs to be measured using either several indicators forming a single scale or by repeatedly measuring across several timepoints; (3) Score reliability can only attenuate effect sizes that are not zero in the first place. Therefore, we focus on phenomena where meta-analytic mean ES can be statistically distinguished from zero.

As all phenomena in the collection fulfil the first condition, 50 phenomena have been catalogued where the effect is studied by comparing the outcome across two separate groups. While the majority of designs make use of some form of control-treatment manipulation, randomly assigning participants to a condition (20 phenomena), some phenomena discuss the effect of pre-existing differences, e.g. biological sex (2 phenomena). From the collection of 50 phenomena, 22 fulfil the second condition, so that estimates of score reliability can be derived. For these phenomena, the dependent variable used to construct the effect size is measured by employing more than a single indicator. At the same time, however, this implies that the remaining 28 phenomena did not make use of more than a single indicator to measure the dependent variable. Therefore, for the majority of phenomena it is not possible to assess in how far measurement quality is sufficiently high in terms of e.g. internal consistency or low random error variance.

From the 22 phenomena that employed more than a single indicator, 19 made use of Likert-style items, where respondents would indicate their agreement to some kind of statement. The scale-length varied from 3 to 10-points. The items measuring the remaining 3 phenomena were coded dichotomously, as responses were right or wrong (2 cases), or expressed agreement vs disagreement dichotomously (1 case). Subsequently, those phenomena need to be identified where the meta-analytic mean effect size is statistically significant different from zero. This will be done as a first step of this manuscript’s analysis procedure. Detailed information concerning the different phenomena and how they were measured can be found at [osf-link], while a brief summary can be found in Table A.

To assess whether ES heterogeneity can be reduced by correcting for score reliability, estimates of meta-analytic heterogeneity are compared across two situations: first, heterogeneity of raw ES, as computed in equation 1 is assessed. Secondly, ES are corrected for imperfect reliability. Computing the heterogeneity of these corrected ES allows for an assessment, whether the heterogeneity did indeed shrink compared to the first situation, as predicted in W&D and H&S.

### Methods

For each replication of a phenomenon, Cohen’s is computed as an estimate of raw ES, according to [Equation 1](#eq-d), including its standard error ([Equation 2](#eq-SEd)). Subsequently, a random-effects meta-analysis is performed using metafor version X.XX. To estimate meta-analytic mean ES and heterogeneity, the REML-estimator is used (ref). To identify which phenomena pass criterion (3), using a Wald-type significance test we assess for which phenomenon the meta-analytic mean ES is significantly different from zero. For this hypothesis test, as all other hypothesis tests in this manuscript, a significance level of .05 is used.

Additionally, estimates of score reliability are derived. While not without criticism (references), Cronbach’s Alpha is used to estimate score reliability. It is often noted that tau-equivalence is an unrealistic assumption to hold against real-world data, implying that other estimates of score reliability, such as McDonald’s Omega (reference), Guttman’s Lambda 2 or 4 (reference), or the Greatest Lower Bound (reference) might be better suited. However, all scales employed in this project are scored by computing the simple mean or sum of responses across items for each individual. Computing the mean or sum of several items implicitly assumes tau-equivalence, as each item contributes equally to the individual’s test score. Therefore, Cronbach’s Alpha is the better choice to estimate score reliability, as it avoids a mismatch in assumptions between how the score is computed and how score reliability is estimated. Using these estimates of Cronbach’s Alpha, the individual estimates of Cohen’s d are corrected, according to [Equation 4](#eq-dc), including the corrected standard errors ([Equation 6](#eq-SEdc)). Thereby, estimates of ES are computed, which are corrected for imperfect reliability. Again, a random-effects meta-analysis is run using metafor, generating an estimate of heterogeneity in corrected ES. Different indicators of heterogeneity are readily available, such as , or the coefficient of variation. However, the descriptions found in H&S and W&D discuss the absolute amount of heterogeneity in terms of variance () or standard deviation (). These parameters discuss the variability of the standardized ES in the population in terms of variance or standard deviation. Therefore, we also make use of absolute heterogeneity here.

Subsequently, in order to assess whether ES heterogeneity was indeed reduced by the attenuation correction procedure, the estimates of heterogeneity in uncorrected ES and corrected ES are compared. Using Cochran’s Q-test, we assess whether the estimates of heterogeneity are statistically significantly different from 0. Only for those projects where we have sufficient confidence that the individual estimates of heterogeneity are larger than zero can we actually begin to interpret whether there was any change.

To identify whether a reduction in ES heterogeneity is indeed accompanied by differences in score reliability, the estimates of score reliability are assessed meta-analytically as well. To do so, a Reliability Generalization Meta-Analysis is performed (references). This entails that individual estimates of score reliability are adequately transformed using Bonett’s transformation: . This transformation has variance-stabilising properties and therefore allows for adequate inferences concerning heterogeneity in score reliability (references). Cochran’s Q-test is used to identify statistically significant heterogeneity in transformed score reliability. Estimates derived from a Reliability Generalization Meta-Analysis using these transformations can be back-transformed to the original score reliability scale (references):

These procedures are followed separately for each phenomenon. The statistical programming language R, Version X.XXX (reference) is used for all data manipulation and statistical analysis.

### Results

[Equations 1](#eq-d), [2](#eq-SEd), [4](#eq-dc), and [6](#eq-SEdc) were used to generate estimates of standardized ES, both uncorrected and corrected, with their corresponding standard errors. Generally, this leads to larger (absolute) ES and larger standard errors. As an example from the 12 selected phenomena, the results from this procedure on Nosek’s Explicit Art sex differences phenomenon are displayed in the forest plot in Figure 1. Here, grey dots represent the uncorrected ES for each sample, while black dots represent the corrected ES in each sample. The bars surrounding these dots show the respective 95%-Cis.

In [Figure 1](#fig-forest) it becomes apparent that the reliability attenuation procedure leads to an increase in the individual absolute effect size, as the black dots are moved further away from zero. Simultaneously, the standard errors grew larger after the attenuation correction, which leads to larger 95%-CIs. This is most easily observed for the mturk-sample, which already had a rather large confidence interval to begin with. Additionally, in the diamond at the bottom of the figure, we can see that the meta-analytic estimate mean ES is also larger after the attenuation correction took place. In [Table 1](#tbl-meanES), the estimates and tests on the meta-analytic average ES, both corrected and uncorrected, can be found.

|  |
| --- |
| Figure 1: Forest Plot Nosek\_Explicit\_Art |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Meta-analytic average ES   | MASC | mu\_str\_raw | pval\_raw | mu\_str\_cor | pval\_cor | | --- | --- | --- | --- | --- | | Albarracin\_Priming\_SAT | .127 (.053) | .016 | .158 (.066) | .016 | | Alter\_Analytic\_Processing | -.155 (.040) | <.001 | -.174 (.044) | <.001 | | Carter\_Flag\_Priming | .018 (.029) | .538 | .026 (.036) | .465 | | Caruso\_Currency\_Priming | -.019 (.026) | .457 | -.022 (.029) | .459 | | Finkel\_Exit\_Forgiveness | -.050 (.049) | .314 | -.056 (.056) | .316 | | Finkel\_Neglect\_Forgiveness | -.051 (.055) | .355 | -.059 (.063) | .347 | | Giessner\_Vertical\_Position | .026 (.023) | .248 | .027 (.025) | .267 | | Hart\_Criminal\_Intentionality | .170 (.078) | .030 | .184 (.085) | .030 | | Hart\_Detailed\_Processing | .082 (.058) | .157 | .096 (.067) | .147 | | Hart\_Intention\_Attribution | -.011 (.058) | .856 | -.021 (.080) | .795 | | Husnu\_Imagined\_Contact | .116 (.032) | <.001 | .129 (.035) | <.001 | | Nosek\_Explicit\_Art | .360 (.047) | <.001 | .383 (.050) | <.001 | | Nosek\_Explicit\_Math | .401 (.029) | <.001 | .415 (.030) | <.001 | | PSACR001\_anxiety\_int | .251 (.022) | <.001 | .263 (.023) | <.001 | | PSACR001\_behav\_int | .035 (.020) | .079 | .040 (.023) | .088 | | PSACR002\_neg\_photo | -.663 (.040) | <.001 | -.747 (.047) | <.001 | | Shnabel\_Willingness\_Reconcile\_Rev | -.533 (.101) | <.001 | -.582 (.109) | <.001 | | Shnabel\_Willingness\_Reconcile\_RPP | -.585 (.062) | <.001 | -.655 (.067) | <.001 | | Srull\_Behaviour\_Hostility | -.051 (.029) | .079 | -.062 (.034) | .071 | | Srull\_Ronald\_Hostility | .063 (.029) | .028 | .074 (.033) | .026 | | Tversky\_Directionality\_Similarity1 | -.111 (.034) | .001 | -.121 (.037) | <.001 | | Zhong\_Desirability\_Cleaning | -.031 (.024) | .194 | -.042 (.036) | .240 | |

[Table 1](#tbl-meanES) demonstrates that what we observed in Figure 1 holds across all 22 phenomena. All meta-analytic effect sizes are larger after applying an attenuation correction procedure. Similarly, the uncertainty quantified in the estimates’ standard error is larger. Additionally, [Table 1](#tbl-meanES) highlights which phenomena pass criterion (3) – the effect size must be statistically significantly distinguishable from zero. 12 phenomena pass this criterion, highlighted in boldface. It may be remarked that the attenuation correction procedure, in this data, appears to have no influence on whether a phenomenon passes the criterion. The differences in p-value are rather small and lead to no difference in conclusion, regarding a significance level of .05.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Tests for ES Heterogeneity   | MASC | tau\_raw | QE\_raw\_str | QEp\_raw | tau\_cor | QE\_cor\_str | QEp\_cor | | --- | --- | --- | --- | --- | --- | --- | | Albarracin\_Priming\_SAT | 0.000 | 5.987 (8) | 0.649 | 0.000 | 6.055 (8) | 0.641 | | Alter\_Analytic\_Processing | 0.025 | 15.706 (20) | 0.735 | 0.018 | 15.701 (20) | 0.735 | | Hart\_Criminal\_Intentionality | 0.171 | 18.96 (11) | 0.062 | 0.185 | 18.81 (11) | 0.065 | | Husnu\_Imagined\_Contact | 0.087 | 47.581 (35) | 0.076 | 0.094 | 47.351 (35) | 0.079 | | Nosek\_Explicit\_Art | 0.178 | 62.771 (35) | 0.003 | 0.190 | 63.074 (35) | 0.003 | | Nosek\_Explicit\_Math | 0.000 | 39.169 (35) | 0.288 | 0.000 | 39.107 (35) | 0.29 | | PSACR001\_anxiety\_int | 0.085 | 74.655 (48) | 0.008 | 0.088 | 74.251 (48) | 0.009 | | PSACR002\_neg\_photo | 0.220 | 219.709 (36) | <.001 | 0.260 | 237.115 (36) | <.001 | | Shnabel\_Willingness\_Reconcile\_Rev | 0.233 | 23.367 (7) | 0.001 | 0.253 | 23.182 (7) | 0.002 | | Shnabel\_Willingness\_Reconcile\_RPP | 0.073 | 8.472 (7) | 0.293 | 0.071 | 8.244 (7) | 0.312 | | Srull\_Ronald\_Hostility | 0.046 | 24.756 (21) | 0.258 | 0.053 | 24.593 (21) | 0.265 | | Tversky\_Directionality\_Similarity1 | 0.001 | 59.434 (60) | 0.496 | 0.001 | 59.102 (60) | 0.509 | |

Cases incompatible with claims made in W&D and H&S are highlighted using boldface.

However, what [Figure 1](#fig-forest) and [Table 1](#tbl-meanES) can not inform us about is in how far the heterogeneity has changed after correcting for (un)reliability. In [Table 2](#tbl-tauES), the estimates and tests of heterogeneity concerning uncorrected and corrected ES on the 12 remaining phenomena, which passed all criteria (1) – (3), can be found. In the table, the estimate of heterogeneity , with the accompanying QE-test statistic, degrees of freedom and its p-value are reported, separately for uncorrected and corrected ES.

Most importantly, [Table 2](#tbl-tauES) demonstrates that of 12 phenomena assessed, for seven phenomena we find patterns incompatible with claims made in W&D and H&S. For those seven phenomena, the reliability attenuation correction led to an increase in absolute ES heterogeneity . For the remaining five phenomena, estimates of ES heterogeneity were zero before and after the correction procedure, leading to no change at all. This leaves three phenomena where the attenuation correction did indeed lead to a reduction of heterogeneity. However, it is crucial to note that the hypothesis tests identified statistically significant heterogeneity in only four out of 12 phenomena and for none of those a reduction in heterogeneity took place. In all four cases of statistically significant heterogeneity, the attenuation correction procedure led to an increase in heterogeneity, contrary to the W&D and H&S predictions.

To make sure that actual differences in score reliability exist, heterogeneity in score reliability is assessed, making use of Reliability Generalization Meta-Analysis (reference). [Table 3](#tbl-RGMA) summarises the results across all 12 phenomena. Overall, scales for almost all phenomena appear have produced scores with an average score reliability larger than .7. Only two phenomena (Albarracin\_Priming\_SAT and Alter\_Analytic\_Processing) come with lower score reliability. However, for those phenomena we could not identify statistically significant heterogeneity in effect sizes in [Table 1](#tbl-meanES) in the first place. From the phenomena where the ES heterogeneity grew larger as a result from the attenuation correction procedure, for six out of seven phenomena, we observe statistically significant heterogeneity in score reliability. Only the Shnabel\_Willingness\_Reconcile\_Rev project produced scores where no heterogeneity in score reliability was identified. However, for this project, it is crucial to note that only a small number of replications (8), was available for analysis. Since the power of Cochran’s Q-test for heterogeneity largely depends on the number of replications, this result is inconclusive.

Despite this one inconclusive result, for the majority of phenomena where ES heterogeneity is larger after applying an attenuation correction, a lack of differences in score reliability can not be made responsible for results contrary to claims made in W&D and H&S. All seven phenomena came with an average score reliability from .74 to .9 and for six out of seven phenomena we identified statistically significant heterogeneity in score reliability.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Reliability Generalization Meta-Analysis   | MASC | mu\_alpha\_str | tau\_alpha | QE\_str | QEp | | --- | --- | --- | --- | --- | | Albarracin\_Priming\_SAT | 0.61 [0.557:0.657] | 0.057 | 26.019 (8) | .001 | | Alter\_Analytic\_Processing | 0.64 [0.604:0.673] | 0.056 | 38.384 (20) | .008 | | Hart\_Criminal\_Intentionality | 0.849 [0.811:0.879] | 0.054 | 66.17 (11) | <.001 | | Husnu\_Imagined\_Contact | 0.814 [0.802:0.825] | 0.023 | 66.389 (35) | .001 | | Nosek\_Explicit\_Art | 0.881 [0.869:0.891] | 0.029 | 142.551 (35) | <.001 | | Nosek\_Explicit\_Math | 0.936 [0.932:0.939] | 0.007 | 59.43 (35) | .006 | | PSACR001\_anxiety\_int | 0.903 [0.89:0.914] | 0.041 | 494.822 (48) | <.001 | | PSACR002\_neg\_photo | 0.791 [0.774:0.806] | 0.047 | 350.852 (36) | <.001 | | Shnabel\_Willingness\_Reconcile\_Rev | 0.832 [0.818:0.845] | 0.000 | 5.053 (7) | .653 | | Shnabel\_Willingness\_Reconcile\_RPP | 0.768 [0.737:0.796] | 0.032 | 17.535 (7) | .014 | | Srull\_Ronald\_Hostility | 0.739 [0.716:0.76] | 0.047 | 91.722 (21) | <.001 | | Tversky\_Directionality\_Similarity1 | 0.854 [0.841:0.866] | 0.039 | 164.933 (60) | <.001 | |

## Alternative discussion of how score reliability affects ES heterogeneity

In this section, we explore some facets on why we disagree with claims stated in W&D and H&S. Additionally, we explore some alternative explanations that might help understand why correcting ES for score reliability does not lead to an inevitable reduction in heterogeneity.

### Misinterpreted equation in Hunter & Schmidt (XXXX)

Hunter & Schmidt 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:” (p. 309)

We adjusted [Equation 9](#eq-HS) using the notation used throughout this text. Here, refers to the ES, undistorted by sampling error but not corrected for measuring error, refers to the ES, undistorted by sampling error and measurement error and ρxx‘ refers to score reliability.

[Equation 9](#eq-HS) implies that heterogeneity in is essentially a function where heterogeneity and mean value of are weighted by mean score reliability and its heterogeneity. While it may not be self-evident in the equation itself, it is easy to construct cases where heterogeneity in is inflated or deflated by heterogeneity in , contradicting Hunter & Schmidt’s interpretation of this equation. In [Figure 2](#fig-HS), the results of employing [Equation 9](#eq-HS) to compute the heterogeneity (variance) in corrected ES is highlighted. All else held equal, only the individual level of mean reliability is varied from .5 to .8. Mean corrected ES is held constant at 1, its variance held constant at .04 and the variance of score reliability held constant at .02.

|  |
| --- |
| Figure 2: Implications by Hunter & Schmidt’s [Equation 9](#eq-HS) |

In [Figure 2](#fig-HS), the dashed red line corresponds to the variance of corrected ES.

If everything else is held constant, the variance in uncorrected ES can be a function only of the mean level of score reliability , according to H&S equation. In that case, Figure 2 demonstrates that correcting for imperfect score reliability does not inevitably lead to lower heterogeneity, as the line representing heterogeneity in uncorrected ES does cross the red line, representing the heterogeneity in corrected ES. This implies that, as long as mean score reliability is below .72, an attenuation correction would lead to a reduction in ES heterogeneity. As mean score reliability grows larger than .72, however, the heterogeneity in uncorrected ES grows larger than the heterogeneity in corrected ES - an attenuation correction would therefore only increase the extent of ES heterogeneity. This demonstrates that, depending on the mean level of score reliability, heterogeneity in uncorrected ES can be higher or lower than heterogeneity in corrected ES. As heterogeneity in corrected ES is not always smaller, the interpretationsof H&S concerning this equation is erroneous.

However, [Equation 9](#eq-HS) is additionally flawed, as it violates a crucial definition. As Hunter & Schmidt point out, [Equation 9](#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 ρxx‘, can not be independent parameters by definition. 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. Equation 1, demonstrating how Cohen’s d is computed demonstrates how the mean difference is standardized in this case. As demonstrated further above in @eg-dct, ES is standardized using the true standard deviation in the sense of CTT. The fact that the true standard deviation parameter is found both in [Equation 5](#eq-dct) and [Equation 3](#eq-rel) demonstrates that ES and score reliability can not be independent variables. Therefore, the basic premise of [Equation 9](#eq-HS) is not fulfilled, as the assumption of variable independence could not be fulfilled.

### Describing ES as a random ratio variable

However, realising that an ES, as defined in [Equation 1](#eq-d), is actually a ratio, allows for a different analytical description of how differences in score reliability can affect ES heterogeneity. If the phenomenon is heterogeneous and score reliability varies across replications, it seems sensible to assume that both numerator and denominator (MD and ) from [Equation 1](#eq-d) vary across replications. In that case, an ES can be described as a random variable stemming from a ratio distribution. A ratio distribution is a probability distribution constructed by dividing one random variable by a second random 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). Here we use the Taylor-estimator of the ratio’s variance to demonstrate how differences in score reliability may affect heterogeneity in ES.

Assuming that the random variables are uncorrelated, simplifies the equation to

Substituting for ES, unstandardized mean difference MD and pooled standard deviation leads to

As stated in [Equation 5](#eq-dct), correcting ES for score reliability can be described through a correction of the pooled standard deviation: . Since is always between zero and one, in the case of imperfect score reliability, will always be smaller than total . Additionally, if the heterogeneity in was introduced purely by differences in measurement precision – the error score variance , this would be removed by the attenuation correction. However, alternatively, the underlying latent variable we attempted to measure may not be distributed identically across replications. In that case, true score variance would vary across replications, leading to differences in both score reliability, corrected ES and raw ES. Importantly, this implies that even if measurements of perfect score reliability were taken across the different replications, heterogeneity in ES would still persist.

In that case, some heterogeneity in would remain, albeit heterogeneity found in was reduced. [Equation 12](#eq-ratio_d0) essentially demonstrates how mean and heterogeneity of MD, paired with mean and heterogeneity of total score standard deviation can be used to estimate heterogeneity in raw ES . Similarly, we can adjust Equation [Equation 12](#eq-ratio_d0) to describe how, instead of total score standard deviation, mean and variance in true score standard deviation affect heterogeneity in corrected ES .

The differences between the two [Equations 12](#eq-ratio_d0) and [13](#eq-ratio_d) demonstrates how ES heterogeneity is affected by an attenuation correction procedure. According to the quotes in H&S and W&D, [Equation 12](#eq-ratio_d0) should lead to a larger value of , compared to the estimate of from [Equation 13](#eq-ratio_d). However, the equations shown throughout this text do not agree with this conclusion.

Instead, what is guaranteed is that the expected value of is smaller than the expected value of . Similarly, the heterogeneity in is guaranteed to be smaller than the uncorrected heterogeneity in . However, since the expected value of either standard deviation is placed in the denominator of the function and the variance of either is placed in the numerator of the function, [Equation 9](#eq-HS) is not sufficient to explain how heterogeneity in ES changes due to the attenuation correction. Instead, we introduce two additional metrics.

describes the relative size of mean true score variance, compared to the mean observed score variance. This informs us, how much of the mean total score variance can be attributed to mean true score variance. Therefore, this metric is close to the average score reliability. Even though formally and the meta-analytic mean score reliability are not the same value, in practice we observe little to no differences between the two parameters. Large values indicate that correcting the individual score variances (or standard deviations for that matter) using the attenuation correction procedure should lead to very little change, as most of the variation observed were due to genuine differences in true scores. Small values indicate the opposite, the observed score variance was due to a larger amount of random error score variance, leading to a stronger correction.

The metric on the other hand describes, in how far the attenuation correction procedure has successfully reduced heterogeneity in score variance . A value of 1 indicates that the heterogeneity in true score variance , the score variance after the attenuation correction, is essentially just as large as the heterogeneity initially observed. Smaller values indicate how much heterogeneity is “left”, after applying an attenuation correction, relative to the initial heterogeneity. This means that a value of .7 indicates that about 70% of heterogeneity in score variance remains, even after correcting for differences in score reliability. As only 30% of score variance heterogeneity could be removed, this would imply that differences in error score variance were responsible for less than a third of the heterogeneity in score variances found. On the contrary, more than two thirds of heterogeneity could be attributed to actual differences in how the underlying true scores are distributed across samples.

Both metrics and make use of parameters of the distribution of true and total score variance, not standard deviations. This choice was made to facilitate the role of score reliability in the subsequent discussion, as score reliability itself is defined by variances, not standard deviations (see [Equation 3](#eq-rel)).

In this section, we attempt to understand under which circumstances the heterogeneity in ES grows larger as the result of an attenuation correction procedure, contrary to claims made in W&D and H&S. To do so, we propose the following inequality, involving [Equations 12](#eq-ratio_d0) and [13](#eq-ratio_d).

Using the metrics and , defined in [Equations 14](#eq-R1) and [15](#eq-R2), we attempt to understand under which circumstances the inequality defined in [Equation 16](#eq-ineq_init) holds true. However, while the metrics make use of mean and heterogeneity of the score variances, [Equation 16](#eq-ineq_init) makes use of mean and heterogeneity of the standard deviations. In order to incorporate the metrics, it is necessary to reparameterise the equation. Using the delta method to approximate mean value and heterogeneity of variance from those of the standard deviations, we know that (ref):

While [Equation 17](#eq-delta_method) only explicitly contains parameters for the observed score variance/standard deviation, the same can be done using the true score variance/standard deviation. Using the approximates defined in [Equation 17](#eq-delta_method) for observed and true score variance, we arrive at a new inequality in [Equation 18](#eq-ineq_CV).

[Equation 18](#eq-ineq_CV) describes that the heterogeneity in corrected ES is larger than the heterogeneity in uncorrected ES, using parameters of the distributions of mean difference MD, true score variance and observed score variance . By introducing the metrics and to this inequality, we can begin to disentangle which conditions need to be fulfilled for the inequality to hold. Rearranging the terms in [Equations 14](#eq-R1) and [15](#eq-R2), and subsequently entering these terms into Equation 18 leads to the following

[Equation 19](#eq-ineq_R), again, describes the inequality that heterogeneity in corrected ES is larger than the heterogeneity in uncorrected ES, incorporating the metrics and . This equation alone is not sufficient to identify the relevant circumstances required for the inequality to hold. However, rearranging the terms from [Equation 19](#eq-ineq_R) leads to [Equation 20](#eq-ineq_fin):

The circumstances under which the inequality described in Equation 20 holds, are circumstances where the claims made in W&D and H&S are directly contradicted, as in those cases the ES heterogeneity is larger after applying an attenuation correction procedure. Concerning the terms in [Equation 20](#eq-ineq_fin), we know that all terms left-hand of the inequality, outside of the brackets (, and ), are bound to be positive. For the inequality to hold, we need the left-hand side of the equation to remain positive, larger than zero. We can distill two scenarios, under which this equation should hold: (a) one of the terms inside the brackets ( and ) is positive and large enough, so that the second term not containing that same bracket is positively dominated by the first term; or (b) both terms inside the brackets need to be positive.

Generally, we know that both and are bound to be positive, as both contain different, strictly positive parameters of the distributions of true and observed score variance. Additionally, we know that is equivalent to the average score reliability, and therefore bound between . Therefore, the term inside the first bracket is bound to be positive (). Similarly, , as the ratio of true and error score variance heterogeneity, is bound between . If is smaller than , then the term inside the second bracket turns negative. Generally, this means that the inequality can only be violated if the reduction in relative heterogeneity of score variance () is larger than the the average score reliability to the power of 3.

Concerning the two scenarios, for scenario (a) to hold true, since the term involving only in the first bracket can only be positive, the second term involving both metrics would need to be negative. For the first term to positively dominate the second term, however, we would need to make additional assumptions about the size of the parameters outside of the brackets, specifically about and . This is something we would prefer to avoid, as it involves additional assumptions about the individual raw effect sizes of the different phenomena. Additionally, it can be said that as long as the term in the second bracket involving both metrics is positive, scenario (b) is true by default. This occurs any time . For example, assuming the average score reliability is about .8, if is at least .512 or larger, the inequality holds true and ES heterogeneity grows larger as a consequence of the attenuation correction procedure, contrary to claims made in W&D and H&S. This means that for the exemplary score reliability of .8, a reduction in relative heterogeneity in score variance of 48.8% () or more is necessary for scenario (b) to no longer be true. The smaller the average score reliability, the smaller the -metric may be for the inequality to still hold true.

Keeping this in mind, even if is smaller than .512, it does not immediately imply that ES heterogeneity is smaller in . For example, it is unlikely that an leads to a term that is not positively dominated by the first term involving only . However, where exactly the cut-off is, where ES heterogeneity starts shrinking due to the attenuation correction procedure, can not be easily defined without additional assumptions about the distribution of .

Generally speaking, the attenuation correction leads to a reduction both in mean value of and in its heterogeneity. While a reduction in the former typically leads to an increase in heterogeneity of , the latter has the potential to compensate for that increase and thereby reduce that heterogeneity after the attenuation correction. This reduction in ES heterogeneity can only occur, if the reduction in score variance heterogeneity is substantially stronger than the reduction in mean score variance. An extreme version of such a situation might occur, if the true score variance is identical across all samples. This would imply that all heterogeneity in score variance identified can be attributed to differences in how much random noise is found in the scores across the samples. In such a case, would be zero and a reduction in ES heterogeneity due to the attenuation correction would be guaranteed. This also implies that the lower and the larger the average score reliability , the more likely it is that the attenuation correction actually does lead to a reduction in ES heterogeneity.

### Re-analysis of archival data

[Table 2](#tbl-tauES) showed that for all phenomena, where we identified statistically significant ES heterogeneity, the heterogeneity was even larger after the attenuation correction procedure. Subsequently, in Equation 16 we demonstrate that if the metric is larger than , ES heterogeneity is bound to be larger after applying said procedure. [Table 4](#tbl-R1R2) reports the meta-analytic estimates of observed and true score variances, including their heterogeneity and metrics and across all 12 phenomena. In [Table 4](#tbl-R1R2) we see that for ten out of 12 phenomena, the metric is larger than the metric. Additionally, we observe that across nine phenomena, is larger than .97. This implies that of the heterogeneity observed in score variances, 97% or more could not be explained by differences in random error introduced by the measuring instrument, but by genuine differences in the underlying distribution. Next to the two exceptional phenomena, where zero heterogeneity in score variance was found (Alter\_Analytic\_Processing and Shnabel\_Willingness\_Reconcile\_Rev), only for the Srull\_Ronald\_Hostility phenomenon some heterogeneity could be attributed to differences in random error variance, albeit less than 14%.

This implies that for all measurements concerning these phenomena, the heterogeneity in observed score variances could not be substantially explained by differences in measurement quality. For no phenomenon, if differences in score variance could be found, was smaller than . As demonstrated in Equation 20, this implies that the ES heterogeneity grows even larger, as we correct for differences in score reliability.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Re-analysis concerning metrics and   | MASC | mu\_X | mu\_T | tau\_X | tau\_T | R1 | R2 | | --- | --- | --- | --- | --- | --- | --- | | Albarracin\_Priming\_SAT | 0.024 | 0.015 | 0.004 | 0.004 | 0.619 | 0.995 | | Alter\_Analytic\_Processing | 0.166 | 0.105 | 0.000 | NA | 0.632 | -Inf | | Hart\_Criminal\_Intentionality | 2.201 | 1.871 | 0.754 | 0.743 | 0.850 | 0.970 | | Husnu\_Imagined\_Contact | 3.324 | 2.712 | 0.472 | 0.469 | 0.816 | 0.987 | | Nosek\_Explicit\_Art | 0.927 | 0.822 | 0.189 | 0.188 | 0.887 | 0.992 | | Nosek\_Explicit\_Math | 1.732 | 1.621 | 0.109 | 0.108 | 0.935 | 0.983 | | PSACR001\_anxiety\_int | 1.149 | 1.040 | 0.251 | 0.247 | 0.905 | 0.966 | | PSACR002\_neg\_photo | 0.603 | 0.481 | 0.131 | 0.131 | 0.799 | 0.986 | | Shnabel\_Willingness\_Reconcile\_Rev | 1.051 | 0.874 | 0.000 | NA | 0.832 | -Inf | | Shnabel\_Willingness\_Reconcile\_RPP | 0.901 | 0.700 | 0.192 | 0.191 | 0.776 | 0.989 | | Srull\_Ronald\_Hostility | 1.877 | 1.387 | 0.328 | 0.305 | 0.739 | 0.865 | | Tversky\_Directionality\_Similarity1 | 4.498 | 3.858 | 1.259 | 1.252 | 0.858 | 0.989 | |

In [Table 2](#tbl-tauES) we also found some phenomena where the heterogeneity did in fact shrink after applying the attenuation correction procedure. This is not reflected by what we observe in [Table 4](#tbl-R1R2), as for almost all phenomena, which would imply that across all phenomena we expected ES heterogeneity to grow as a consequence of the attenuation correction procedure. However, it is crucial to realise that we did not find statistically significant ES heterogeneity for any of the phenomena where ES heterogeneity was reduced. This implies, that we actually could not statistically distinguish the ES heterogeneity from zero. Therefore, what seems to be a reduction in ES heterogeneity was most likely caused by estimation issues, as we could not accurately estimate the extent of ES heterogeneity, either before or after the attenuation correction.

Lastly, for two out of the 12 phenomena, we find a of 0. For these phenomena, it is therefore also impossible to compute the metric. While for the Alter\_Analytic\_Processing phenomenon we did not identify statistically significant ES heterogeneity in the first place, we did find some for the Shnabel\_Willingness\_Reconcile\_Rev phenomenon. At the same time, we did not identify statistically significant heterogeneity in score reliability. With zero heterogeneity in score variances, Equation 13 implies that the ES heterogeneity is bound to grow larger due to the attenuation procedure. Also, this phenomenon was assessed with a particularly low number of replications (8). Most likely, adequate power to detect heterogeneity in score reliability and observed score variance requires a larger number of replications than adequate power to detect heterogeneity. In that case, the results from Table 5, concerning the Shnabel\_Willingness\_Reconcile\_Rev phenomenon might alternatively be explained by power-issues. Similarly, while Table 3 does not imply specific expectations concerning the presence of score variance heterogeneity in the Alter\_Analytic\_Processing phenomenon, the low number of replications (20) makes it hard to distinguish whether the estimate of is actually sensible or can also be attributed to low power.

## Discussion

Across twelve exemplary archival datasets, we demonstrated that claims made in H&S and W&D do not hold in empirical observations. Even though removing the impact of score reliability on ES heterogeneity by means of attenuation correction procedures, they do not lead to a reduction in heterogeneity as claimed in H&S and W&D. Inferences derived from the equation supplied in H&S (ref), [Equation 9](#eq-HS), also do not warrant such a conclusion. Alternatively, describing standardised effect sizes in terms of a ratio distribution and approximating its variance, we demonstrated that differences in score reliability may both inflate or deflate variance in ES, depending on how the relative heterogeneity in score variance changes due to the attenuation correction procedure. The metric , compared with the metric or the average score reliability, informs us in how far score variance heterogeneity decreases due to the attenuation correction.

These results fit in 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 typically deflate heterogeneity in uncorrected correlations. Only as the true heterogeneity in correlations grows larger, score reliability differences actually inflate heterogeneity as claimed in W&D. 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 mean values of these parameters, while the impact on standard deviations used for standardisation purposes needs to be checked via .

The analytical arguments presented indicate that an attenuation correction procedure, even if no differences between score reliability exist, lead to larger ES heterogeneity. It appears that, as the meta-analytical mean ES grows larger due to the correction procedure, so does the ES heterogeneity, if homogeneity is present. If heterogeneity is present on the other hand and , still a reduction in ES heterogeneity is not inevitable. It can be argued that, for the differences in ES to grow smaller, the reduction in relative heterogeneity in score variance needs to be substantially large, so the increase in ES heterogeneity due to the larger mean ES can be “outpaced”. However, we did not observe this pattern in the 12 phenomena that allowed for such an analysis.

Heterogeneity in ES even in direct replications, where experimental factors are held as constant as possible, is already substantial (Renkewitz, Fünderich, & Beinhauer, 2024). 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, deflated estimates of heterogeneity imply that the theory or phenomenon is understood even worse than initially assumed.

The data-sets available to this project fall into two categories: direct multi-site replications of a specific phenomenon, employing identical protocols across samples; a large-scale collaborative effort to assess a single phenomenon globally, distributed across a large number of independent laboratories. For both categories, rather arbitrary sampling schemes were used. The majority of respondents are students at the respective universities, taking part in the studies voluntarily or because they were required as part of their study’s programme (at times compensation was awarded). This is not the same as random sampling. These samples are put together arbitrarily, it is unlikely that the underlying distribution of true scores is meaningful for a properly specified population. The between-studies randomness introduced by these undefined sampling schemes manifests itself in true score variance heterogeneity . Between-studies randomness introduces by issues of measurement quality manifest themselves in error score variance heterogeneity . This implies that it should be of little surprise that the differences in true score variance largely exceed those in error score variance.

Without sampling schemes that assure that the samples are constructed to be representative for some kind of population, we have no reason to assume that there should be few differences in the dispersion of true scores. Even though the measuring instruments are translated to different languages, which is bound to introduce differences in how much random noise is added to scores across samples, we have no reason to assume that these differences are larger than those introduced by employing no sampling schemes. Therein, we believe that, as the differences in score reliability are largely introduced by failure of ensuring appropriate sampling mechanisms, not by differences in how well the instrument captures the underlying construct, attenuation correction can not be expected to remedy ES heterogeneity.

However, this also means that results contrary to what we observed in this project are not unthinkable. Valid sampling schemes need to be considered, ensuring that the samples are representative for populations that can be meaningfully compared. In that case, we are convinced that true score variance should be more stable across samples and differences in the instrument’s measuring quality are bound to have a much larger impact. Not only would an attenuation correction in such a case actually have the potential of reducing ES heterogeneity, the variability underlying a psychological phenomenon could be meaningfully assessed across samples - something that would greatly help attempts to identify universal and non-universal patterns of psychological phenomena. As it stands, current approaches to multi-site direct replications and other collaborative efforts suffer strongly from the lack of appropriate sampling schemes.

The empirical arguments presented are based on a rather small number of non-representative data-sets. Based on the combination of analytical and empirical arguments, we are convinced that these differences in score reliability across replications oftentimes mask true heterogeneity. In realistic meta-analytic studies of conceptual replications, data stems from a variety of sources. Across different populations, different research settings or designs are used with varying or adjusted research instruments. It is unlikely that, across these replications, we can expect either the measuring quality or the true score variance to be stable. However, as demonstrated above, differences in score reliability and true score variance tend to lead to larger corrected ES heterogeneity, implying that observed ES heterogeneity was reduced by differences in score reliability. However, whether these results actually generalize beyond these data-sets remains to be seen. Unfortunately, the behavioural sciences are in dire need of open-data that resemble multi-site replications. As the majority of results discussed over the last years (in 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 following Equation 17 rested on the use of the delta-method, as Equation 17 contained parameters of the score standard deviations’ distributions, while the and metrics were designed to discuss score variances. The delta-method rests on a number of assumptions: (1) approximations derived from the delta-method require for the transformed variable to follow a marginal normal distribution (ref). Since the variable under discussion here is the standard deviation, which is bounded to be larger than zero and has a non-normal sampling distribution (ref), this assumption is most likely violated; (2) the derivative of the transformation function needs to be available. As the transformation function is essentially just the square root, this assumption should hold up; (3) as the delta-method approximates the parameters of the transformed variable using the first-order Taylor-expansion, this implies that higher-order terms are required to be negligible. In practice, this implies that the variance of the untransformed variable needs to be small. However, as this entire procedure discusses the influence of variance in standard deviations on ES heterogeneity, introduced by differences in score reliability, this assumption is probably violated as well. The violation of these assumptions likely introduce a bias in estimates derived from these approximations (ref). However, the inequality derived with use of the delta-method is not used to actually derive any estimates, but to discuss how an attenuation correction procedure can affect ES heterogeneity. Therefore, we believe that the violation of these assumptions is defensible in this case, as the analytical arguments derived from the inequality should still hold up, nevertheless. Additionally, we used the empirical examples available to assess in how far estimates derived using the delta-method approximations diverge from estimates derived without it. The differences identified are rather miniscule at the third decimal, reinforcing our belief that these assumption violations are defensible.

Similarly, all Equations following Equation 10 rest on the assumption that mean difference () and true or total score variance ( and ) are independent. While this assumption is in line with assumptions underlying CTT, it is by no means guaranteed that it holds in these data-sets. If this assumption was violated, whether the relationship was positive or negative, indicated by a positive or negative covariance in Equation 10, would most likely have an impact on how strongly ES heterogeneity is affected by the attenuation correction procedure, possibly even the direction of how it is affected. However, currently we have no reason to assume that there is a systematic relationship between mean difference and standard deviation. As this relationship would also affect the distribution of ES like Cohen’s d itself, strong violations of this assumption would go far beyond invalidating the claims we derived in Equation 16. Instead, assumptions of the meta-analytic tests concerning the size ES and the presence of ES heterogeneity would be violated, making any research on change in ES heterogeneity due to attenuation correction obsolete.

The work done by Hunter and Schmidt (2014) was crucial in informing and guiding methodology development in the field of meta-analyses. Similarly, Wiernik and Dahlke (2020) raise a number of important points that have been neglected in the application of meta-analytic research over the last decades. We strongly agree with their ideas that underappreciated (differences in un-)reliability introduce substantial biases in meta-analytic estimates and tests in Psychology, which need to be corrected. Future applications of meta-analyses on standardized ES need to take score reliability into account to arrive at correct estimates and inferences. However, the meta-analyst should not expect that correcting ES for score reliability will reduce the extent of heterogeneity identified.