frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 15 October 2025
DOI 10.3389/fpsyg.2025.1592658

OPEN ACCESS

EDITED BY
Jason W. Osborne,
Miami University, United States

REVIEWED BY
Nikolaus Bezruczko,
The Chicago School of Professional Psychology, United States
Ron Jonathan Pat-El,
Open University of the Netherlands, Netherlands

*CORRESPONDENCE
Povilas Karvelis
povilas.karvelis@camh.ca

RECEIVED 12 March 2025
ACCEPTED 05 September 2025
PUBLISHED 15 October 2025

CITATION
Karvelis P and Diaconescu AO (2025) Clarifying the reliability paradox: poor measurement reliability attenuates group differences. *Front. Psychol.* 16:1592658. doi: 10.3389/fpsyg.2025.1592658

Clarifying the reliability paradox: poor measurement reliability attenuates group differences

Povilas Karvelis[1]* and Andreea O. Diaconescu[1,2,3,4]

[1]Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, ON, Canada, [2]Department of Psychiatry, University of Toronto, Toronto, ON, Canada, [3]Institute of Medical Sciences, University of Toronto, Toronto, ON, Canada, [4]Department of Psychology, University of Toronto, Toronto, ON, Canada

Cognitive sciences are grappling with the reliability paradox: measures that robustly produce within-group effects tend to have low test-retest reliability, rendering them unsuitable for studying individual differences. Despite the growing awareness of this paradox, its full extent remains underappreciated. Specifically, most research focuses exclusively on how reliability affects correlational analyses of individual differences, while largely ignoring its effects on studying group differences. Moreover, some studies explicitly and erroneously suggest that poor reliability does not pose problems for studying group differences, possibly due to conflating within- and between-group effects. In this brief report, we aim to clarify this misunderstanding. Using both data simulations and mathematical derivations, we show how observed group differences get attenuated by measurement reliability. We consider multiple scenarios, including when groups are created based on thresholding a continuous measure (e.g., patients vs. controls or median split), when groups are defined exogenously (e.g., treatment vs. control groups, or male vs. female), and how the observed effect sizes are further affected by differences in measurement reliability and between-subject variance between the groups. We provide a set of equations for calculating attenuation effects across these scenarios. This has important implications for biomarker research and clinical translation, as well as any other area of research that relies on group comparisons to inform policy and real-world applications.

KEYWORDS

reliability paradox, test-retest reliability, individual differences, group differences, group effects, measurement reliability, effect size attenuation, clinical translation

1 Introduction

An influential paper by Hedge et al. (2018) has highlighted the "reliability paradox": cognitive tasks that produce robust within-group effects tend to have poor test-retest reliability, undermining their use for studying individual differences. Many studies have followed, demonstrating the prevalence of low test-retest reliability and emphasizing its implications for studying individual differences across various research contexts, including neuroimaging, computational modeling, psychiatric disorders, and clinical translation (Enkavi et al., 2019; Elliott et al., 2020, 2021; Fröhner et al., 2019; Nikolaidis et al., 2022; Kennedy et al., 2022; Nitsch et al., 2022; Blair et al., 2022; Zuo et al., 2019; Milham et al., 2021; Feng et al., 2022; Haines et al., 2023; Parsons et al., 2019; Hedge et al., 2020; Enkavi and Poldrack, 2021; Zorowitz and Niv, 2023; Gell et al., 2023; Rouder et al., 2023; Karvelis et al., 2023, 2024; Clayson, 2024; Vrizzi et al., 2025).

However, the studies on this topic tend to focus exclusively on how test-retest reliability affects correlational individual differences analyses without making it clear that it is just as relevant for studying group differences (although see LeBel and Paunonen, 2011; Zuo et al., 2019). Not only that, some studies incorrectly suggest that poor test-retest reliability is not problematic for studying group differences. For example: "*Low reliability scores are problematic only if we were interested in differences between individuals (within a group) rather than between groups*" (De Schryver et al., 2016); "*although improved reliability is critical for understanding individual differences in correlational research, it is not very relevant or informative for studies comparing conditions or groups*" (Zhang and Kappenman, 2024); "*On a more positive note, insufficient or unproven test-retest reliability does not directly imply that one cannot reliably assess group differences (e.g., clinical vs. control)*" (Schaaf et al., 2023); "*while many cognitive tasks (including those presented here) have been well validated in case-control studies (e.g., comparing MDD and healthy individuals) where there may be large group differences, arguably these tests may be less sensitive at detecting individual differences*" (Foley et al., 2024); "*The reliability paradox... implies that many behavioral paradigms that are otherwise robust at the group-level (e.g., those that produce highly replicable condition- or group-wise differences) are unsuited for testing and building theories of individual differences*" (Haines et al., 2020); "*Many tasks clearly display robust between-group or between-condition differences, but they also tend to have sub-optimal reliability for individual differences research*" (Parsons et al., 2019). Sometimes the opposite mistake is made by suggesting that poor reliability is equally detrimental for studying both between-group differences and within-group effects (e.g., see Figure 1 in Zuo et al., 2019).

An apparent common thread across these examples is the conflation of within-group effects and between-group effects, treating both simply as "group effects." However, within- and between-group effects are often in tension. If an instrument is designed to produce strong within-group effects (i.e., robust changes across conditions or time points), it will typically do so by minimizing between-subject variability – which in turn reduces its ability to reliably detect individual or between-group differences. This trade-off lies at the heart of the reliability paradox. The key insight here is that both group and individual differences live on the same dimension of between-subject variability and are, therefore, affected by measurement reliability in the same way.

The aim of this brief report is therefore (1) to clarify and highlight the relevance of the reliability paradox for studying group differences (2) to present simulation-based illustrations to make the implications of the reliability paradox more intuitive, and (3) to provide a set of mathematical formulae for effect size attenuation that cover different scenarios of group comparisons.

2 Methods

2.1 Simulated data

To simulate data, we sampled from a normal distribution

$$X \sim \mathcal{N}(\mu, \sigma_b^2 + \sigma_e^2) \tag{1}$$

by independently varying between-subject (σ_b^2) and error (σ_e^2) variances. To represent repeated measurements of task performance, we generated two distributions ("test" and "retest") with the same mean $\mu = 0$. To simulate one-sample effects, we simply generated another distribution that is shifted upward by a constant offset ($\mu = 2$). To simulate paired-sample effects, we generated two distributions (corresponding to Condition 1 and Condition 2) one of which was at $\mu = 0$ and the other at $\mu = 2$. Finally, to illustrate relationships with external traits, we generated additional datasets with fixed between-subject variance (σ_b^2) and no error variance ($\sigma_e^2 = 0$). We specified true population correlations of $r_{true} = 0.5$ and $r_{true} = 0.9$ to represent different levels of association between task performance and symptom/trait measures.

Note, while we refer to these data distributions as representing "task performance" and "traits/symptoms" to make this analysis more intuitive, these datasets are generated at a high level of abstraction and do not assume any specific data-generating process—i.e., we are not simulating trial-level or item-level data, we are simply generating distributions of individual-level scores.

Patients vs. controls groups were created by splitting the datasets such that 10% of the distribution with the highest symptom scores were assigned to the patient group while the remaining 90% were assigned to the control group. For creating high vs. low trait groups, we simply performed a median split across the datasets.

To achieve sufficient stability of the test-retest reliability and effect size estimates, we used a sample size of 10,000 for each combination of σ_b and σ_e, each of which was varied between 0.3 and 2 for most of the analysis. To further increase the stability of our effects, when investigating the relationship between true and observed effect size metrics as a function of reliability, we increased the sample size to 1,000,000. We also kept between-subject variance fixed at $\sigma_b = 0.5$, and only varied error variance in the range $\sigma_e \in [0.01\ 3]$. When comparing how the different statistical metrics fare when it comes to significance testing, we used a sample of $N = 60$ (a typical effect size seen in practice)—to achieve stable estimates of p-values we averaged results over 20,000 repetitions.

2.2 Statistical analysis

To assess test-retest reliability, we used the intraclass correlation coefficient (ICC) (Fleiss, 2011; Koo and Li, 2016; Liljequist et al., 2019):

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}, \tag{2}$$

where for clarity we omit the within-subject variance term in the denominator because throughout our analysis it was kept at 0 between test and retest measurements.

To measure group effects, we used Cohen's d as the main effect size metric for both within- and between-group effects. To account for unequal variances between groups when performing 90/10%

split (for controls vs. patients), we used d^*, a variant of Cohen's d that does not assume equal variances:

$$d = \frac{\mu_2 - \mu_1}{\sigma_p}, \quad \text{where}$$
$$\sigma_p = \sqrt{\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1+n_2-2}} \tag{3}$$
$$d^* = \frac{\mu_2 - \mu_1}{\sigma_{np}}, \quad \text{where} \quad \sigma_{np} = \sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}} \tag{4}$$

Equation 3 is the standard method for calculating Cohen's d using the pooled standard deviation, where in the numerator we have the difference between the means of the two groups (μ_1 and μ_2), with n_1 and n_2 denoting the sample sizes of each group, and σ_1^2 and σ^2 denoting the variances of each group. In contrast, Cohen's d^* (Equation 4) is based on the non-pooled standard deviation. While the use of a non-pooled standard deviation somewhat complicates the interpretation of the resulting effect size metric, empirical investigations have shown that it possesses robust inferential properties and may be a more practical option given that the equal variance requirement is rarely met in practice (Delacre et al., 2021).

To be more comprehensive in our analysis, alongside Cohen's d, we also report a non-parametric alternative, the rank-biserial correlation coefficient (r_{rb}). To perform the associated statistical significance tests, we use the t-test and the Mann-Whitney U test, respectively.

To quantify the impact of reliability on statistical power, we calculated the required sample sizes for each effect size metric to achieve 80% power at $\alpha = 0.05$ (Cohen, 2013). The critical values $z_{\alpha/2}$ and z_β are defined as:

$$z_{\alpha/2} = \Phi^{-1}(1-\alpha/2) \quad \text{and} \quad z_\beta = \Phi^{-1}(1-\beta), \tag{5}$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution, $\alpha = 0.05$ is the two-sided significance level, and $\beta = 0.20$ (corresponding to 80% power) is the Type II error rate. With these parameters, $z_{\alpha/2} = 1.96$ and $z_\beta = 0.84$.

For Pearson correlation, we used the Fisher z-transform approach (Cohen, 2013):

$$N_r = 3 + \left(\frac{z_{\alpha/2}+z_\beta}{\text{atanh}(|r_{obs}|)}\right)^2, \tag{6}$$

where $|r_{obs}|$ is the absolute value of the observed correlation attenuated by measurement error. For Cohen's d from median split analysis, we used the standard two-sample t-test power formula (Cohen, 2013):

$$N_d = 4\left(\frac{z_{\alpha/2}+z_\beta}{d_{obs}}\right)^2, \tag{7}$$

where d_{obs} is the observed effect size attenuated by reliability. For the rank-biserial correlation with equal group sizes, we used (Noether, 1987):

$$N_{rb} \approx \frac{4}{3}\left(\frac{z_{\alpha/2}+z_\beta}{r_{rb,obs}}\right)^2, \tag{8}$$

where $r_{rb,obs}$ is the observed rank-biserial correlation coefficient.

3 Results

3.1 The reliability paradox

First, we performed data simulations to illustrate the reliability paradox—namely, that strong within-group effects are inherently at odds with high test-retest reliability. We generated multiple sets of synthetic data by independently varying between-subject variance (σ_b^2) and measurement error variance (σ_e^2), and explored how this affects test-retest reliability and observed within-group effects (Figure 1). The key takeaway here is that while test-retest reliability is determined by the proportion of between-subject variance relative to total variance $\sigma_b^2/(\sigma_b^2+\sigma_e^2)$, within-group effects depend on the total variance $\sigma_b^2+\sigma_e^2$. In other words, increasing error variance σ_e^2 will reduce both reliability and within-group effect sizes, whereas increasing between-subject variance σ_b^2 will improve reliability but will reduce within-group effects, since a fixed mean difference becomes smaller relative to the larger total variance.

Note that for simplicity here we assumed the between-subject variance in condition 1 and condition 2 to be uncorrelated (Figure 1C). Under this assumption, the variance of the difference scores (i.e., the individual-level condition differences used to compute paired-sample Cohen's d) is equal to the sum of the variances in each condition. Due to this linear relationship, Figure 1F can therefore be interpreted as referring to the between-subject variance of difference scores. In practice, however, task conditions are often positively correlated, which reduces the variance of the difference scores—thereby inflating effect sizes while reducing the reliability of the underlying scores (Cronbach and Furby, 1970; Hedge et al., 2018; Draheim et al., 2019). This would introduce non-linearities in how the between-subject variance of each condition relates to the observed effect size, but the relationship between the between-subject variance of difference scores and the observed effect size, which we aim to convey here, still holds.

3.2 Group differences: data simulations

3.2.1 Data simulations for groups created by dichotomizing continuous measures

Next, using data simulations we investigated how measurement reliability affects group differences when the groups are derived by thresholding a continuous measure (e.g., symptoms or traits). To make this more intuitive we considered two common scenarios: 1) mental disorders, which can be generally thought of as occupying the low end of the wellbeing distribution (Huppert et al., 2005) or any specific symptom dimension, and 2) "low" vs. "high" cognitive traits formed by a median split (Figure 2). Considering these scenarios using simulations helps illustrate one key insight: raw group differences scale together with between-subject variance (Figures 2B, C). Hence, unlike with within-group effects, reducing between-subject variance does not lead to larger group effects. Adding measurement error can further

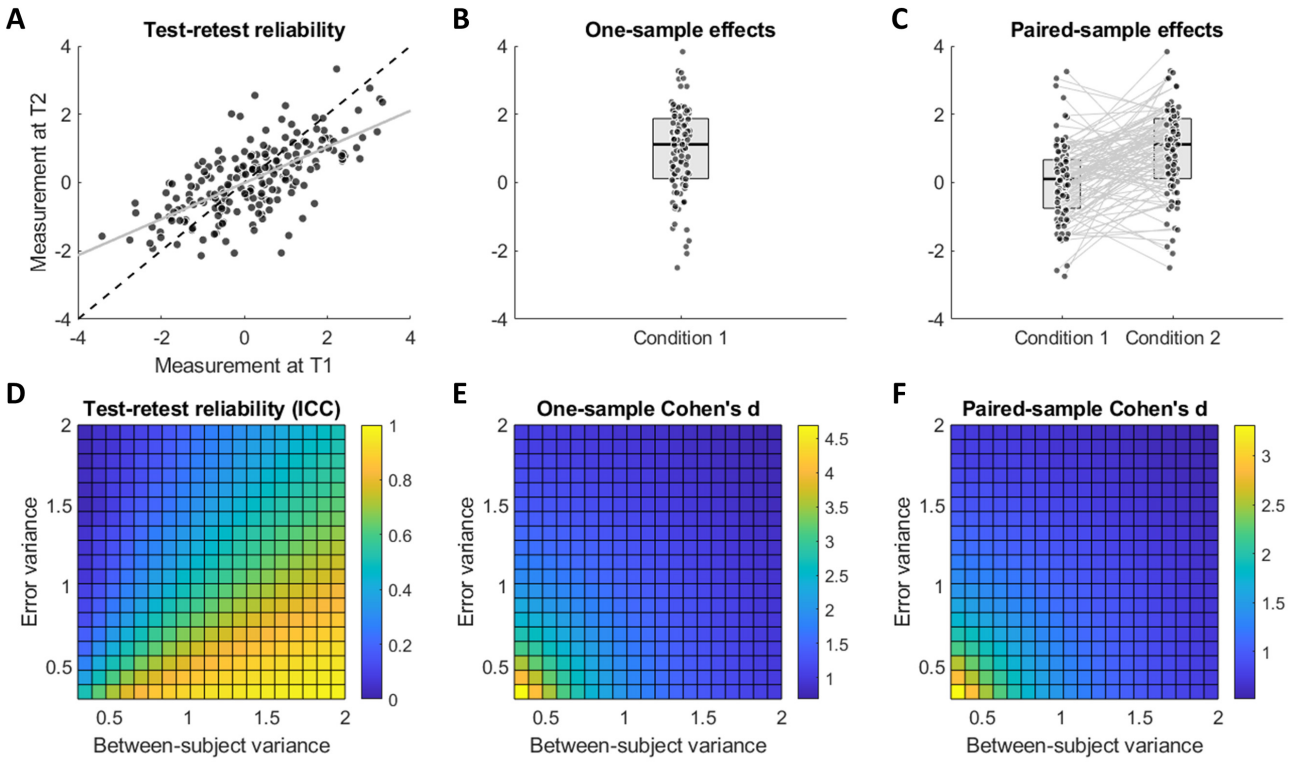

FIGURE 1
The reliability paradox. Top panels **(A–C)** illustrate the statistical tests under consideration: **(A)** test-retest correlation (same measure obtained twice), **(B)** one-sample test (mean of a single condition compared to zero), and **(C)** paired-sample test (mean difference between two conditions). Bottom panels **(D–F)** show how the observed outcomes of these tests depend on the relative contributions of error variance and between-subject variance. Test-retest reliability **(D)** increases when error variance is minimized and between-subject variance is maximized, whereas observed one-sample and paired-sample effect sizes **(E, F)** increase when both error and between-subject variances are minimized.

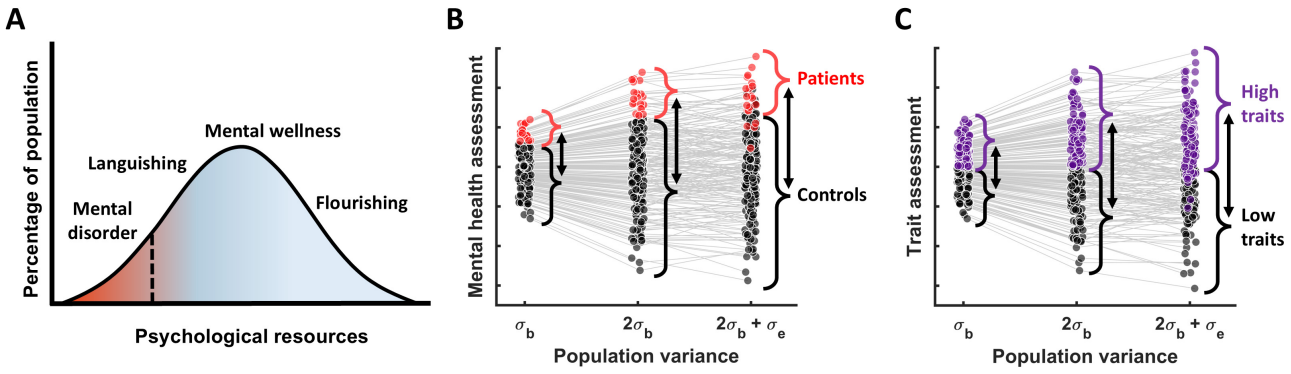

FIGURE 2
Group differences as a function of population variance. **(A)** The dimensional view of mental disorders; adapted from Huppert et al. (2005). **(B)** The relationship between patients vs. controls group differences and population variance, assuming that patients are defined as 10% of the population with the poorest mental health. **(C)** A more general case illustrating the group differences resulting from the median split of the data (based on some cognitive measure) as a function of population variance. In both **(B)** and **(C)** we see that as true population variance increases, raw group difference increase too, while adding measurement error to the true scores results in misclassification of some individuals—which will end up attenuating observed group differences in the measures of interest.

increase raw group differences, but it also leads to misclassification (Figures 2B, C), which ultimately reduces observed group differences in any other measures of interest, as we will see next.

We generated correlated "symptoms/traits" and "task performance" datasets such that they had $r_{true} = 0.7$ Pearson's correlation. To derive the groups, we defined patients as occupying the 10% of the population with the poorest mental health

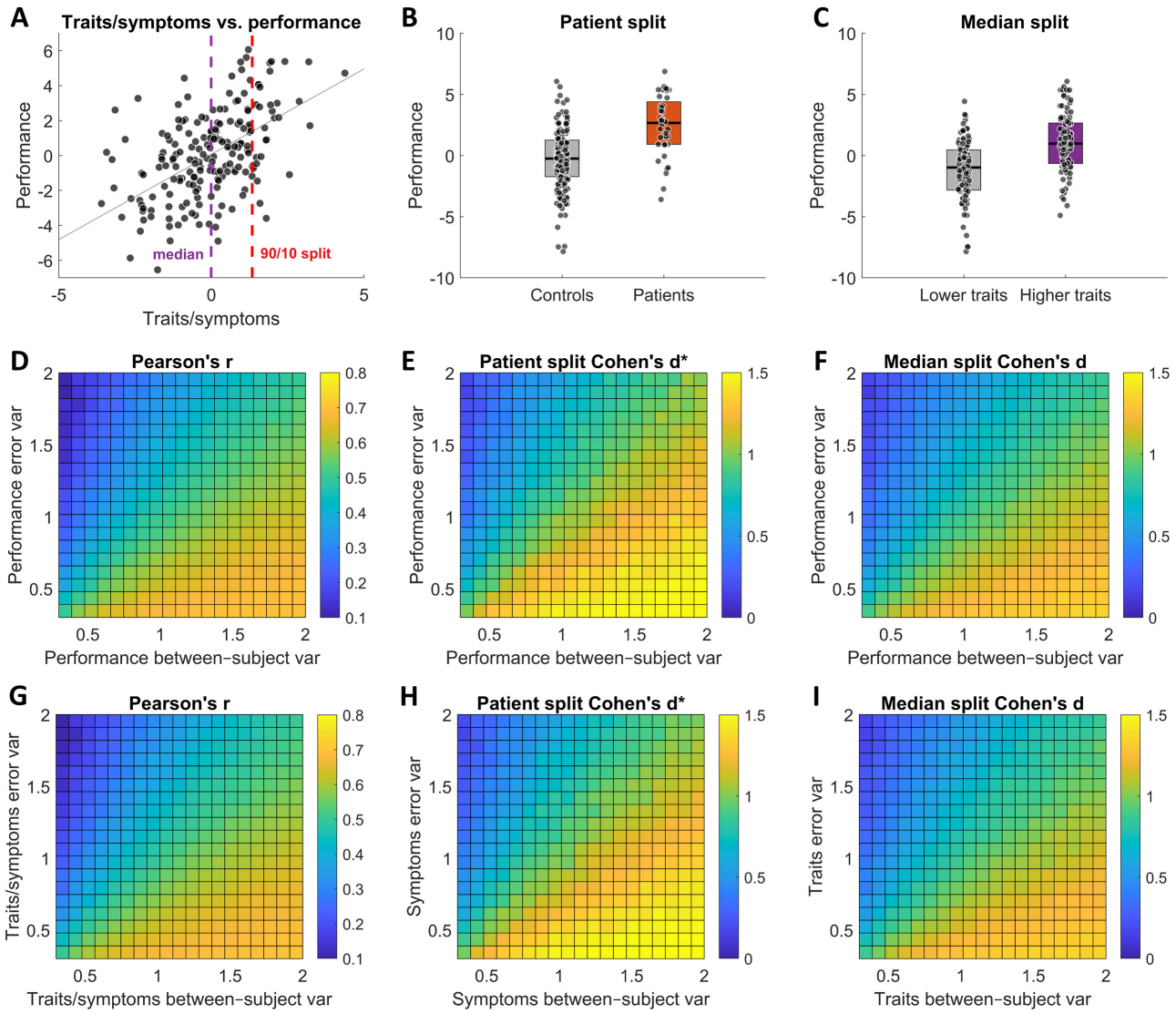

FIGURE 3
Test-retest reliability effects on observed group differences. The top row panels **(A–C)** illustrate the different analysis scenarios, while the 2 row panels **(D–F)** show the corresponding observed effects for different error and between-subject variance values of task performance, and the bottom row panels **(G–I)** show the corresponding observed effects for different error and between-subject variance values of symptoms/traits. **(A)** An illustration of correlation between traits or symptoms and task performance. The vertical dashed lines indicate how the data was split for the two analysis scenarios. **(B)** An illustration of patient and control groups created by assigning 10% of the population with the poorest mental health to patient group and the remaining 90% to control group. **(C)** An illustration of "low" and "high" trait groups by performing a median split. Overall, the test results show that both observed correlation strength and observed group differences increase with increasing test-retest reliability (i.e., with reducing error variance/increasing between-subject variance) both when varying between-subject and error variances of task performance measures **(D–F)** and symptoms/traits measures **(G–I)**.

(Figures 3A, B), with the rest of the population being controls; using the median split along the trait dimension Figure 3A, we grouped individuals into "low" and "high" trait groups (Figure 3C).

We then examined how the observed effect size metrics were affected by independently varying σ_b and σ_e of task performance and then of traits/symptoms (**see Methods for more details**). In both cases, we find the same results: reducing reliability in either task performance measures (Figures 3D–F) or symptoms/traits measures (Figures 3G–I) leads to attenuation of observed effect sizes that mirror those of correlational analyses of individual differences (Equation 9).

3.2.2 Comparing attenuation effects across effect size metrics

In a correlational analysis, the true correlation strength between a measure x and a measure y is attenuated by their respective reliabilities following (Spearman, 1904):

$$r_{observed} = r_{true}\sqrt{ICC_x ICC_y}. \quad (9)$$

We compared the observed effect sizes from our simulations to the predicted attenuation relationship and found that both parametric (Cohen's d) and non-parametric (rank-biserial r_{rb})

estimates closely followed the same reliability-based scaling as Pearson's r, especially for moderate true correlations ($r_{\text{true}} = 0.5$; Figure 4A). However, Cohen's d deviated more substantially when $r_{\text{true}} = 0.9$ (Figure 4A) inset due to increasing deviations from normality caused by dichotomization. Thus, when the assumptions of the effect size metric hold, observed between-group differences can be approximated as:

$$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{ICC_x ICC_y} \tag{10}$$

Although attenuation similarly affects both correlations and group differences, it is important to keep in mind that correlational analyses generally retain greater statistical power. Figure 4B illustrates that p-values for group comparisons of dichotomized data are larger than those for correlation tests ($N = 60$, $r_{\text{true}} = 0.5$) and Figure 4C similarly illustrates that required sample sizes (to have 80% power at $\alpha = 0.05$) for dichotomized data are substantially larger than those for correlational analysis, especially when reliability is low. That is simply because the variance discarded during dichotomization results in information loss. This is well documented in previous work (e.g., MacCallum et al., 2002; Royston et al., 2006; Naggara et al., 2011; Streiner, 2002) therefore we will not go into any further details here. Just, please, avoid dichotomizing your continuous data as much as you can.

3.3 Group differences: mathematical derivations

3.3.1 Attenuation for exogenously defined groups

In previous sections, we used simulations to illustrate how poor reliability attenuates group differences when groups are derived from a noisy continuous measure. These visualizations were meant to provide an intuitive understanding of the attenuation effect. Here, we derive the same effect mathematically, but this time considering exogenously defined groups (e.g., male vs. female or treatment vs. control), which are categorical and not subject to measurement error.

For such externally defined groups, random measurement error does not systematically bias the raw mean difference (Lord and Novick, 1968; Nunnally and Bernstein, 1994) but inflates total variance ($\sigma_b^2 + \sigma_e^2$). Consequently, the raw difference scales with between-subject variance rather than total variance (Cohen, 1988; Ellis, 2010):

$$\mu_2 - \mu_1 = \delta_{\text{true}}\, \sigma_b. \tag{11}$$

The expression for $\delta_{observed}$ will then depend on total variance, such that:

$$\delta_{\text{observed}} = \frac{\mu_2 - \mu_1}{\sqrt{\sigma_b^2 + \sigma_e^2}} \tag{12}$$

$$= \delta_{\text{true}} \frac{\sigma_b}{\sqrt{\sigma_b^2 + \sigma_e^2}} \tag{13}$$

$$= \delta_{\text{true}} \sqrt{ICC}\,, \tag{14}$$

where used Equations 11, 2 to get to the final expression.

Notably, this same attenuation mechanism applies to both externally defined groups and groups formed via dichotomization, although the latter additionally suffers from misclassification effects.

3.3.2 Attenuation when the groups have different reliabilities

Equation 14 assumes both groups have the same measurement reliability, but this is not always the case. Here, we derive a more general formula that accounts for differing reliabilities. If the groups have reliabilities ICC_1 and ICC_2, their observed standard deviations are:

$$\sigma_1 = \frac{\sigma_b}{\sqrt{ICC_1}}, \quad \sigma_2 = \frac{\sigma_b}{\sqrt{ICC_2}}, \tag{15}$$

where we assume both groups share the same true between-subject standard deviation. Using the non-pooled variance estimation (see Methods for more details), the observed standard deviation is given by:

$$\sigma_{np} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}} = \sqrt{\frac{\sigma_b^2}{2}\left(\frac{1}{ICC_1} + \frac{1}{ICC_2}\right)}$$

$$= \sigma_b \sqrt{\frac{1}{2}\left(\frac{1}{ICC_1} + \frac{1}{ICC_2}\right)}. \tag{16}$$

Thus, the observed standardized difference is:

$$\delta_{\text{observed}} = \frac{\mu_2 - \mu_1}{\sigma_{np}} \tag{17}$$

$$= \frac{\delta_{\text{true}}\, \sigma_b}{\sigma_b \sqrt{\frac{1}{2}\left(\frac{1}{ICC_1} + \frac{1}{ICC_2}\right)}} \tag{18}$$

$$= \delta_{\text{true}} \sqrt{\frac{2}{\frac{1}{ICC_1} + \frac{1}{ICC_2}}} \tag{19}$$

$$= \delta_{\text{true}} \sqrt{\frac{2\, ICC_1\, ICC_2}{ICC_1 + ICC_2}}. \tag{20}$$

In the special case where $ICC_1 = ICC_2 = ICC$, this expression simplifies to Equation 14, consistent with the earlier result.

3.3.3 Attenuation when true variances are also unequal

Thus far, we assumed that both groups share the same underlying between-subject standard deviation. Here, we relax

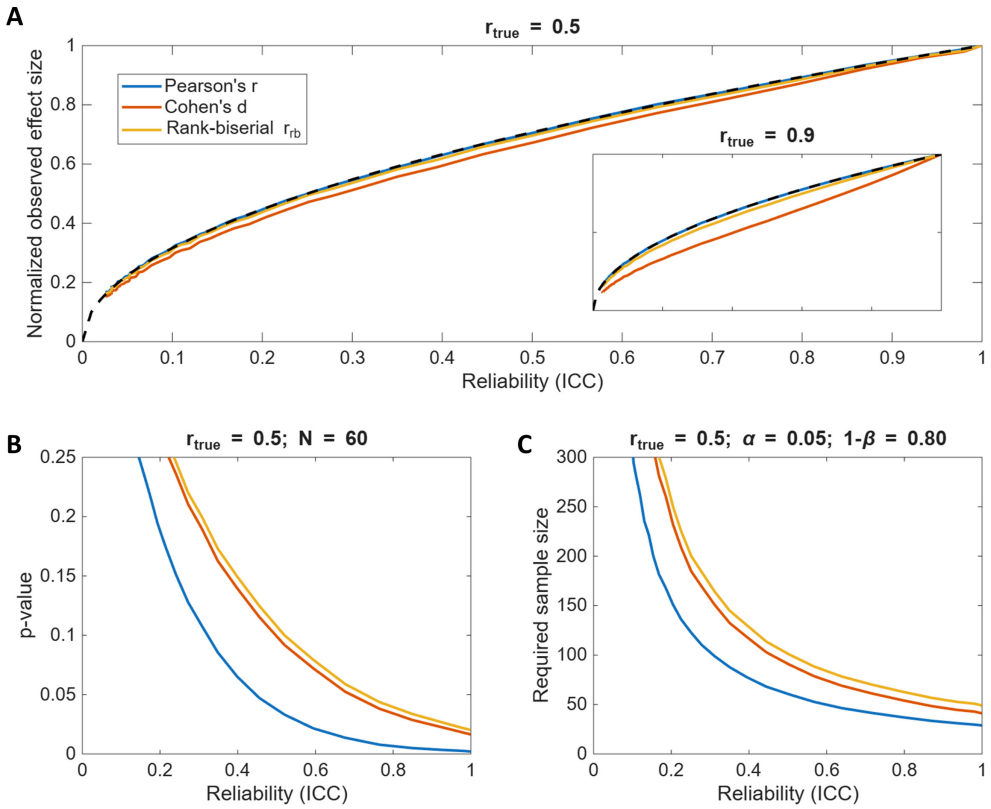

FIGURE 4
Test-retest reliability effects across different effect size metrics and statistical tests. **(A)** The observed effect sizes as a function of reliability for $r_{true} = 0.5$, comparing group differences to correlational strength. Note, because the effect sizes among the tests are not directly comparable, each effect size is normalized by its own maximum value at ICC = 1. The inset shows the results for $r_{true} = 0.9$. The dashed line denotes $r_{observed} = r_{true}\sqrt{ICC_x ICC_y}$. **(B)** The p-value as a function of reliability for $r_{true} = 0.5$ and the total sample size of $N = 60$. Dichotomizing data substantially increases p-values, especially when reliability is low. **(C)** The required sample size to achieve 80% statistical power at $\alpha = 0.05$ as a function of reliability for the three effect size metrics. Dichotomizing data substantially increases the required sample sizes to detect the same true effect, especially when reliability is low.

that assumption and allow the two groups to have different true variances, such that the total true variance is

$$\sigma_{b,\mathrm{np}} = \sqrt{\frac{\sigma_{b,1}^2 + \sigma_{b,2}^2}{2}}, \tag{21}$$

and the observed variances are

$$\sigma_{\mathrm{obs},1} = \frac{\sigma_{b,1}}{\sqrt{ICC_1}}, \quad \sigma_{\mathrm{obs},2} = \frac{\sigma_{b,2}}{\sqrt{ICC_2}}, \tag{22}$$

and so the total observed unpooled variance is

$$\sigma_{\mathrm{np}} = \sqrt{\frac{\sigma_{\mathrm{obs},1}^2 + \sigma_{\mathrm{obs},2}^2}{2}} = \sqrt{\frac{\sigma_{b,1}^2/ICC_1 + \sigma_{b,2}^2/ICC_2}{2}}. \tag{23}$$

Thus, the observed standardized difference is

$$\delta_{\mathrm{observed}} = \frac{\mu_2 - \mu_1}{\sigma_{\mathrm{np}}} = \frac{\delta_{\mathrm{true}}\,\sigma_{b,\mathrm{np}}}{\sigma_{\mathrm{np}}}. \tag{24}$$

Substituting Equations 21, 23 into Equation 24 yields

$$\delta_{\mathrm{observed}} = \delta_{\mathrm{true}} \sqrt{\frac{\sigma_{b,1}^2 + \sigma_{b,2}^2}{\sigma_{b,1}^2/ICC_1 + \sigma_{b,2}^2/ICC_2}}. \tag{25}$$

In the special case where $\sigma_{b,1} = \sigma_{b,2} = \sigma_b$, Equation 25 simplifies to Equation 20.

3.3.4 Attenuation by classification reliability

We can further extend these equations to take into account the reliability of group labels (e.g., patients vs. controls). Note, that we have already demonstrated via simulations that when group classification is error prone, the observed group differences scales as $\sqrt{ICC}$ for the underlying continuous measure. However, when comparing two groups, a more likely measure of classification reliability that would be used is Cohen's Kappa (κ) (Cohen, 1960), which measures the reliability of categorical labels (and is often used to quantify the inter-rater reliability of clinical diagnoses). The relationship between κ and the underlying reliability of continuous measures ICC can be shown to be Kraemer (1979):

$$\kappa = \frac{2}{\pi} \arcsin\left(\sqrt{ICC}\right). \quad (26)$$

Rearranging this for ICC gives

$$ICC = \sin^2\left(\frac{\pi}{2}\kappa\right). \quad (27)$$

Now, the expression Equation 10 can be re-expressed in terms of classification reliability, while Equations 20, 25 can be further extended to account for classification reliability:

$$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{ICC \cdot \sin\left(\frac{\pi}{2}\kappa\right)}, \quad (28)$$

$$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{\frac{2\,ICC_1\,ICC_2}{ICC_1 + ICC_2} \sin\left(\frac{\pi}{2}\kappa\right)}, \quad (29)$$

$$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{\frac{\sigma^2_{b,1} + \sigma^2_{b,2}}{\sigma^2_{b,1}/ICC_1 + \sigma^2_{b,2}/ICC_2} \sin\left(\frac{\pi}{2}\kappa\right)}. \quad (30)$$

We summarize all the attenuation equations in Box 1.

4 Discussion

This report extends the implications of the reliability paradox beyond its original focus on individual differences (Hedge et al., 2018), demonstrating that it presents the same problems when studying group differences. When groups are formed by thresholding continuous measures (e.g., patients vs. controls), the resulting loss of statistical power makes detecting group differences (vs. individual differences) even harder when reliability is low. We hope that this work will help raise awareness of measurement reliability implications in group differences research and that the provided mathematical expressions will help researchers better account for the magnitude of the effect size attenuation in their studies.

4.1 Implications for clinical translation

Poor reliability leads to small observed effects, which severely impedes clinical translation (Karvelis et al., 2023; Nikolaidis et al., 2022; Gell et al., 2023; Tiego et al., 2023; Moriarity and Alloy, 2021; Paulus and Thompson, 2019; Hajcak et al., 2017). For example, for a measure to have diagnostic utility—defined as $\geq$ 80% sensitivity and $\geq$ 80% specificity—it must show a group difference of $d \geq 1.66$ (Loth et al., 2021). Note that $d \geq 0.8$ is considered "large" and it is rarely seen in practice. This may also explain why treatment response prediction research, where it is common to dichotomize symptom change into responders vs. non-responders, has so far shown limited success (Karvelis et al., 2022). Improving the reliability of measures to uncover the landscape of large effects is therefore of paramount importance (DeYoung et al., 2025; Nikolaidis et al., 2022; Zorowitz and Niv, 2023). This applies not only to cognitive performance measures—where the reliability paradox discussion originates—but equally to other instruments including clinical rating scales and diagnostic criteria (Regier et al., 2013; Shrout, 1998), self-report questionnaires (Enkavi et al., 2019; Vrizzi et al., 2025), and experience sampling methods (ESM) (Dejonckheere et al., 2022; Csikszentmihalyi and Larson, 1987). To begin uncovering large effect sizes, however, reliability analysis and reporting must first become a routine research practice (Karvelis et al., 2023; Parsons et al., 2019; LeBel and Paunonen, 2011). While some guidelines such as APA's JARS for psychological research (Appelbaum et al., 2018) and COSMIN for health measurement instruments (Mokkink et al., 2010) do encourage routine reporting of reliability, others, such as PECANS for cognitive and neuropsychological studies (Costa et al., 2025), do not mention reliability or psychometric quality at all, underscoring the need to continue raising awareness of measurement reliability issues.

4.2 Double bias: reliability attenuation and small-sample inflation

Correct interpretation of observed effects requires considering not only the attenuation effects we describe here, but also

BOX 1 Attenuation of observed group differences in different scenarios.

$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{ICC}$	Continuous outcome measured with reliability ICC; classification is error-free.
$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{ICC \cdot \sin\left(\frac{\pi}{2}\kappa\right)}$	Continuous outcome measured with reliability ICC while classification reliability is κ.
$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{\frac{2\,ICC_1\,ICC_2}{ICC_1+ICC_2}}$	Continuous outcome with group-specific reliabilities ICC_1 and ICC_2; classification is error-free.
$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{\frac{2\,ICC_1\,ICC_2}{ICC_1+ICC_2} \sin\left(\frac{\pi}{2}\kappa\right)}$	Continuous outcome measured with group-specific reliabilities ICC_1 and ICC_2 while classificaiton reliability is κ.
$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{\frac{\sigma^2_{b,1}+\sigma^2_{b,2}}{\sigma^2_{b,1}/ICC_1+\sigma^2_{b,2}/ICC_2}}$	Continuous outcome with group-specific variances $\sigma^2_{b,1}$, $\sigma^2_{b,2}$ and reliabilities ICC_1, ICC_2; classification is error-free.
$\delta_{\text{observed}} = \delta_{\text{true}} \sqrt{\frac{\sigma^2_{b,1}+\sigma^2_{b,2}}{\sigma^2_{b,1}/ICC_1+\sigma^2_{b,2}/ICC_2} \sin\left(\frac{\pi}{2}\kappa\right)}$	Continuous outcome measures with with group-specific variances $\sigma^2_{b,1}$, $\sigma^2_{b,2}$ and reliabilities ICC_1, ICC_2, while classification reliability is κ.

sampling error. While low measurement reliability attenuates observed effect sizes, small samples produce unstable estimates that are often selectively reported, leading to systematic inflation of reported effects—known as the winner's curse (Sidebotham and Barlow, 2024; Button et al., 2013; Ioannidis, 2008). Currently, research in cognitive neuroscience and psychology is dominated by small samples, with an estimated 50% of research reporting false positive results (Szucs and Ioannidis, 2017); also see Schäfer and Schwarz (2019). While the attenuation of effect sizes can be addressed by the equations we provide, inflation due to the winner's curse can be mitigated by collecting larger samples, preregistering analyses, applying bias-aware estimation or meta-analytic techniques (Button et al., 2013; Nosek et al., 2018; Zöllner and Pritchard, 2007; Vevea and Hedges, 1995).

4.3 Broader implications for real-world impact

Although we presented our statistical investigation with psychiatry and cognitive sciences in mind, the implications of our results are quite general and could inform any area of research that relies on group comparisons, including education, sex, gender, age, race, and ethnicity (e.g., Hyde, 2016; Ones and Anderson, 2002; Roth et al., 2001; Rea-Sandin et al., 2021; Perna, 2005; Vedel, 2016). The reliability of measures is rarely considered in such studies, but the observed effect sizes are often treated as proxies for practical importance (Cook et al., 2018; Funder and Ozer, 2019; Kirk, 2001, 1996; Olejnik and Algina, 2000) and are used to inform clinical practice (e.g., Ferguson, 2009; McGough and Faraone, 2009) and policy (e.g., Lipsey et al., 2012; Pianta et al., 2009; McCartney and Rosenthal, 2000). Not accounting for the reliability of measures can therefore create a very misleading scientific picture and lead to damaging real-world consequences.

4.4 Limitations and caveats

Our derivations of effect size attenuation are based on parametric assumptions and may not give precise estimates when the data is highly non-normal or is contaminated with outliers. By extension, they may not give precise estimates for non-parametric group differences metrics, although it should still provide a good approximation. Furthermore, we should highlight once again that our derivations rely on using non-pooled variance for calculating standardized mean differences, which allows dropping the assumption of equal variance. Thus, when the true variances are indeed not equal between the groups, it is important to use the non-pooled variance version of Cohen's d* (see Delacre et al., 2021, for further details) when using the attenuation equations. However, if the true variances are roughly equal, the attenuation relationships derived here will hold just as well for the standard Cohen's d, which uses pooled variance.

Data availability statement

The datasets presented in this study can be found in online repositories. The code for producing the data simulations and figures is available at: https://github.com/povilaskarvelis/clarifying_the_reliability_paradox.

Author contributions

PK: Conceptualization, Formal analysis, Funding acquisition, Investigation, Visualization, Writing – original draft, Writing – review & editing. AD: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. PK is supported by CIHR Fellowship (472369). AD was supported by NSERC Discovery Fund (214566) and the Krembil Foundation (1000824).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., and Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. *Am. Psychol.* 73:3. doi: 10.1037/amp0000191

Blair, R. J. R., Mathur, A., Haines, N., and Bajaj, S. (2022). Future directions for cognitive neuroscience in psychiatry: recommendations for biomarker design based on recent test re-test reliability work. *Curr. Opin. Behav. Sci.* 44:101102. doi: 10.1016/j.cobeha.2022.101102

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475

Clayson, P. E. (2024). The psychometric upgrade psychophysiology needs. *Psychophysiology* 61:e14522. doi: 10.1111/psyp.14522

Cohen, J. (1960). A cofficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge: New York. doi: 10.4324/9780203771587

Cook, B. G., Cook, L., and Therrien, W. J. (2018). Group-difference effect sizes: Gauging the practical importance of findings from group-experimental research. *Learn. Disabil. Res. Pract.* 33, 56–63. doi: 10.1111/ldrp.12167

Costa, C., Pezzetta, R., Toffalini, E., Grassi, M., Cona, G., Miniussi, C., et al. (2025). Enhancing the quality and reproducibility of research: preferred evaluation of cognitive and neuropsychological studies-the pecans statement for human studies. *Behav. Res. Methods* 57:182. doi: 10.3758/s13428-025-02705-3

Cronbach, L. J., and Furby, L. (1970). How we should measure "change": or should we? *Psychol. Bull.* 74:68. doi: 10.1037/h0029382

Csikszentmihalyi, M., and Larson, R. (1987). Validity and reliability of the experience-sampling method. *J. Nerv. Ment. Dis.* 175, 526–536. doi: 10.1097/00005053-198709000-00004

De Schryver, M., Hughes, S., Rosseel, Y., and De Houwer, J. (2016). Unreliable yet still replicable: a comment on lebel and paunonen (2011). *Front. Psychol.* 6:2039. doi: 10.3389/fpsyg.2015.02039

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., et al. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychol. Assess.* 34:1138. doi: 10.1037/pas0001178

Delacre, M., Lakens, D., Ley, C., Liu, L., and Leys, C. (2021). Why hedges'g* s based on the non-pooled standard deviation should be reported with welch's *t*-test. doi: 10.31234/osf.io/tu6mp. [Epub ahead of print].

DeYoung, C. G., Hilger, K., Hanson, J. L., Abend, R., Allen, T. A., Beaty, R. E., et al. (2025). Beyond increasing sample sizes: optimizing effect sizes in neuroimaging research on individual differences. *J. Cogn. Neurosci.* 37, 1–12. doi: 10.1162/jocn_a_02297

Draheim, C., Mashburn, C. A., Martin, J. D., and Engle, R. W. (2019). Reaction time in differential and developmental research: a review and commentary on the problems and alternatives. *Psychol. Bull.* 145:508. doi: 10.1037/bul0000192

Elliott, M. L., Knodt, A. R., and Hariri, A. R. (2021). Striving toward translation: strategies for reliable fmri measurement. *Trends Cogn. Sci.* 25, 776–787. doi: 10.1016/j.tics.2021.05.008

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., et al. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychol. Sci.* 31, 792–806. doi: 10.1177/0956797620916786

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, Cambridge, UK. doi: 10.1017/CBO9780511761676

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., et al. (2019). Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proc. Nat. Acad. Sci.* 116, 5472–5477. doi: 10.1073/pnas.1818430116

Enkavi, A. Z., and Poldrack, R. A. (2021). Implications of the lacking relationship between cognitive task and self-report measures for psychiatry. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 6, 670–672. doi: 10.1016/j.bpsc.2020.06.010

Feng, C., Thompson, W. K., and Paulus, M. P. (2022). Effect sizes of associations between neuroimaging measures and affective symptoms: a meta-analysis. *Depress. Anxiety* 39, 19–25. doi: 10.1002/da.23215

Ferguson, C. J. (2009). An effect size primer: a guide for clinicians and researchers. *Profess. Psychol. Res. Pract.* 40, 532–538. doi: 10.1037/a0015808

Fleiss, J. L. (2011). *Design and Analysis of Clinical Experiments*. New York, NY: John Wiley & Sons.

Foley, É. M., Slaney, C., Donnelly, N. A., Kaser, M., Ziegler, L., and Khandaker, G. M. (2024). A novel biomarker of interleukin 6 activity and clinical and cognitive outcomes in depression. *Psychoneuroendocrinology* 164:107008. doi: 10.1016/j.psyneuen.2024.107008

Fröhner, J. H., Teckentrup, V., Smolka, M. N., and Kroemer, N. B. (2019). Addressing the reliability fallacy in FMRI: similar group effects may arise from unreliable individual effects. *Neuroimage* 195, 174–189. doi: 10.1016/j.neuroimage.2019.03.053

Funder, D. C., and Ozer, D. J. (2019). Evaluating effect size in psychological research: sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* 2, 156–168. doi: 10.1177/2515245919847202

Gell, M., Eickhoff, S. B., Omidvarnia, A., Kueppers, V., Patil, K. R., Satterthwaite, T. D., et al. (2023). The burden of reliability: How measurement noise limits brain-behaviour predictions. *bioRxiv*. doi: 10.1101/2023.02.09.527898

Haines, N., Kvam, P. D., Irving, L., Smith, C., Beauchaine, T. P., Pitt, M. A., et al. (2020). Learning from the reliability paradox: how theoretically informed generative models can advance the social, behavioral, and brain sciences. *PsyArXiv*. doi: 10.31234/osf.io/xr7y3

Haines, N., Sullivan-Toole, H., and Olino, T. (2023). From classical methods to generative models: tackling the unreliability of neuroscientific measures in mental health research. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging.* 8, 822–831. doi: 10.31234/osf.io/ax34v

Hajcak, G., Meyer, A., and Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: internal consistency limits between-subjects effects. *J. Abnorm. Psychol.* 126:823. doi: 10.1037/abn0000274

Hedge, C., Bompas, A., and Sumner, P. (2020). Task reliability considerations in computational psychiatry. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging.* 5, 837–839. doi: 10.1016/j.bpsc.2020.05.004

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186. doi: 10.3758/s13428-017-0935-1

Huppert, F. A., Baylis, N., and Keverne, B. (2005). *The Science of Well-Being*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198567523.001.0001

Hyde, J. S. (2016). Sex and cognition: gender and cognitive functions. *Curr. Opin. Neurobiol.* 38, 53–56. doi: 10.1016/j.conb.2016.02.007

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology* 19, 640–648. doi: 10.1097/EDE.0b013e31818131e7

Karvelis, P., Charlton, C. E., Allohverdi, S. G., Bedford, P., Hauke, D. J., and Diaconescu, A. O. (2022). Computational approaches to treatment response prediction in major depression using brain activity and behavioral data: a systematic review. *Netw. Neurosci.* 6, 1–52. doi: 10.1162/netn_a_00233

Karvelis, P., Hauke, D. J., Wobmann, M., Andreou, C., Mackintosh, A., de Bock, R., et al. (2024). Test-retest reliability of behavioral and computational measures of advice taking under volatility. *PLoS ONE* 19:e0312255. doi: 10.1371/journal.pone.0312255

Karvelis, P., Paulus, M. P., and Diaconescu, A. O. (2023). Individual differences in computational psychiatry: a review of current challenges. *Neurosci. Biobehav. Rev.* 148:105137. doi: 10.1016/j.neubiorev.2023.105137

Kennedy, J. T., Harms, M. P., Korucuoglu, O., Astafiev, S. V., Barch, D. M., Thompson, W. K., et al. (2022). Reliability and stability challenges in abcd task FMRI data. *Neuroimage* 252:119046. doi: 10.1016/j.neuroimage.2022.119046

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educ. Psychol. Meas.* 56, 746–759. doi: 10.1177/0013164496056005002

Kirk, R. E. (2001). Promoting good statistical practices: some suggestions. *Educ. Psychol. Meas.* 61, 213–218. doi: 10.1177/00131640121971185

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 44, 461–472. doi: 10.1007/BF02296208

LeBel, E. P., and Paunonen, S. V. (2011). Sexy but often unreliable: the impact of unreliability on the replicability of experimental findings with implicit measures. *Pers. Soc. Psychol. Bull.* 37, 570–583. doi: 10.1177/0146167211400619

Liljequist, D., Elfving, B., and Skavberg Roaldsen, K. (2019). Intraclass correlation-a discussion and demonstration of basic features. *PLoS ONE* 14:e0219854. doi: 10.1371/journal.pone.0219854

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., et al. (2012). *Translating the Statistical Representation of the Effects of Education Interventions*

into More Readily Interpretable Forms. Washington, DC: National Center for Special Education Research.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

Loth, E., Ahmad, J., Chatham, C., López, B., Carter, B., Crawley, D., et al. (2021). The meaning of significant mean group differences for biomarker discovery. *PLoS Comput. Biol.* 17:e1009477. doi: 10.1371/journal.pcbi.1009477

MacCallum, R. C., Zhang, S., Preacher, K. J., and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychol. Methods* 7:19. doi: 10.1037//1082-989X.7.1.19

McCartney, K., and Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Dev*. 71, 173–180. doi: 10.1111/1467-8624.00131

McGough, J. J., and Faraone, S. V. (2009). Estimating the size of treatment effects: moving beyond *p* values. *Psychiatry* 6:21.

Milham, M. P., Vogelstein, J., and Xu, T. (2021). Removing the reliability bottleneck in functional magnetic resonance imaging research to achieve clinical utility. *JAMA Psychiatry* 78, 587–588. doi: 10.1001/jamapsychiatry.2020.4272

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The cosmin checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international delphi study. *Qual. Life Res.* 19, 539–549. doi: 10.1007/s11136-010-9606-8

Moriarity, D. P., and Alloy, L. B. (2021). Back to basics: the importance of measurement properties in biological psychiatry. *Neurosci. Biobehav. Rev.* 123, 72–82. doi: 10.1016/j.neubiorev.2021.01.008

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., and Altman, D. G. (2011). Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *Am. J. Neuroradiol.* 32, 437–440. doi: 10.3174/ajnr.A2425

Nikolaidis, A., Chen, A. A., He, X., Shinohara, R., Vogelstein, J., Milham, M., et al. (2022). Suboptimal phenotypic reliability impedes reproducible human neuroscience. *bioRxiv*. doi: 10.1101/2022.07.22.501193

Nitsch, F. J., Lüpken, L. M., Lüschow, N., and Kalenscher, T. (2022). On the reliability of individual economic rationality measurements. *Proc. Nat. Acad. Sci.* 119:e2202070119. doi: 10.1073/pnas.2202070119

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *J. Am. Stat. Assoc*. 82, 645–647. doi: 10.1080/01621459.1987.10478478

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proc. Nat. Acad. Sci.* 115, 2600–2606. doi: 10.1073/pnas.1708274114

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory*. McGraw-Hill, New York, NY, 3rd edition.

Olejnik, S., and Algina, J. (2000). Measures of effect size for comparative studies: applications, interpretations, and limitations. *Contemp. Educ. Psychol.* 25, 241–286. doi: 10.1006/ceps.2000.1040

Ones, D. S., and Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: some british data. *J. Occup. Organ. Psychol.* 75, 255–276. doi: 10.1348/096317902320369703

Parsons, S., Kruijt, A.-W., and Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Adv. Methods Pract. Psychol. Sci*. 2, 378–395. doi: 10.1177/2515245919879695

Paulus, M. P., and Thompson, W. K. (2019). The challenges and opportunities of small effects: the new normal in academic psychiatry. *JAMA Psychiatry* 76, 353–354. doi: 10.1001/jamapsychiatry.2018.4540

Perna, L. W. (2005). The benefits of higher education: sex, racial/ethnic, and socioeconomic group differences. *Rev. High. Educ*. 29, 23–52. doi: 10.1353/rhe.2005.0073

Pianta, R. C., Barnett, W. S., Burchinal, M., and Thornburg, K. R. (2009). The effects of preschool education: what we know, how public policy is or is not aligned with the evidence base, and what we need to know. *Psychol. Sci. Public Interest* 10, 49–88. doi: 10.1177/1529100610381908

Rea-Sandin, G., Korous, K. M., and Causadias, J. M. (2021). A systematic review and meta-analysis of racial/ethnic differences and similarities in executive function performance in the united states. *Neuropsychology* 35:141. doi: 10.1037/neu0000715

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., et al. (2013). DSM-5 field trials in the united states and canada, part II: test-retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* 170, 59–70. doi: 10.1176/appi.ajp.2012.12070999

Roth, P. L., Bevier, C. A., Bobko, P., SWITZER III, F. S., and Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: a meta-analysis. *Pers. Psychol.* 54, 297–330. doi: 10.1111/j.1744-6570.2001.tb00094.x

Rouder, J. N., Kumar, A., and Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychon. Bull. Rev.* 30, 2049–2066. doi: 10.3758/s13423-023-02293-3

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat. Med.* 25, 127–141. doi: 10.1002/sim.2331

Schaaf, J. V., Weidinger, L., Molleman, L., and van den Bos, W. (2023). Test-retest reliability of reinforcement learning parameters. *Behav. Res. Methods* 56, 1–18. doi: 10.3758/s13428-023-02203-4

Schäfer, T., and Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. *Front. Psychol.* 10:442717. doi: 10.3389/fpsyg.2019.00813

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Stat. Methods Med. Res.* 7, 301–317. doi: 10.1191/096228098672090967

Sidebotham, D., and Barlow, C. (2024). The winner's curse: why large effect sizes in discovery trials always get smaller and often disappear completely. *Anaesthesia* 79,86–90. doi: 10.1111/anae.16161

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159

Streiner, D. L. (2002). Breaking up is hard to do: the heartbreak of dichotomizing continuous data. *Can. J. Psychiatry* 47, 262–266. doi: 10.1177/070674370204700307

Szucs, D., and Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* 15:e2000797. doi: 10.1371/journal.pbio.2000797

Tiego, J., Martin, E. A., DeYoung, C. G., Hagan, K., Cooper, S. E., Pasion, R., et al. (2023). Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. *Nat. Ment. Health* 1, 304–315. doi: 10.1038/s44220-023-00057-5

Vedel, A. (2016). Big five personality group differences across academic majors: a systematic review. *Pers. Individ. Dif.* 92, 1–10. doi: 10.1016/j.paid.2015.12.011

Vevea, J. L., and Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* 60, 419–435. doi: 10.1007/BF02294384

Vrizzi, S., Najar, A., Lemogne, C., Palminteri, S., and Lebreton, M. (2025). Behavioral, computational and self-reported measures of reward and punishment sensitivity as predictors of mental health characteristics. *Nat. Ment. Health* 3, 1–13. doi: 10.1038/s44220-025-00427-1

Zhang, W., and Kappenman, E. S. (2024). Maximizing signal-to-noise ratio and statistical power in ERP measurement: single sites versus multi-site average clusters. *Psychophysiology* 61:e14440. doi: 10.1111/psyp.14440

Zöllner, S., and Pritchard, J. K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* 80, 605–615. doi: 10.1086/512821

Zorowitz, S., and Niv, Y. (2023). Improving the reliability of cognitive task measures: a narrative review. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging.* 8, 789–797. doi: 10.31234/osf.io/phzrb

Zuo, X.-N., Xu, T., and Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nat. Hum. Behav.* 3, 768–771. doi: 10.1038/s41562-019-0655-x