



# Understanding the role of subpopulations and reliability in between-group studies

Selena Wang<sup>1</sup> · Paul De Boeck<sup>1</sup>

Accepted: 30 August 2021 / Published online: 7 February 2022  
© The Psychonomic Society, Inc. 2022

## Abstract

The replication crisis has led to a renewed discussion about the impacts of measurement quality on the precision of psychology research. High measurement quality is associated with low measurement error, yet the role of reliability in the quality of experimental research is not always well understood. In this study, we attempt to understand the role of reliability through its relationship with power while focusing on between-group designs for experimental studies. We outline a latent variable framework to investigate this nuanced relationship through equations. An under-evaluated aspect of the relationship is the variance and homogeneity of the subpopulation from which the study sample is drawn. Higher homogeneity implies a lower reliability, but yields higher power. We proceed to demonstrate the impact of this relationship between reliability and power by imitating different scenarios of large-scale replications with between-group designs. We find negative correlations between reliability and power when there are sizable differences in the latent variable variance and negligible differences in the other parameters across studies. Finally, we analyze the data from the replications of the ego depletion effect (Hagger et al., 2016) and the replications of the grammatical aspect effect (Eerland et al., 2016), each time with between-group designs, and the results align with previous findings. The applications show that a negative relationship between reliability and power is a realistic possibility with consequences for applied work. We suggest that more attention be given to the homogeneity of the subpopulation when study-specific reliability coefficients are reported in between-group studies.

**Keywords** Reliability · Power · Effect size · Replication · Homogeneity

The reputation of psychological science is at stake when prominent psychological findings (e.g., Carney et al., 2010; Strack et al., 1988; Sripada et al., 2014; Zhong & Liljenquist, 2006; Meltzoff & Moore, 1977) fail to replicate. The lower-than-expected replication success is most often labeled the replication crisis (Francis, 2013; Ioannidis, 2005; Pashler & Wagenmakers, 2012). Whether the crisis is genuine or only a matter of perception, it has led to extensive reflections and proposed remedies on how to improve psychological research (e.g., Meehl, 1990; Rosenthal, 1979;

De Boeck & Jeon, 2018; Shrout & Rodgers, 2018; Tackett et al., 2017; Nosek & Lakens, 2014).

Improving the reliability of the dependent variable (DV) is a common recommendation to counter the replication crisis (e.g., Flake et al., 2017; Funder et al., 2014; Stanley & Spence, 2014). Yet, it is unclear that high reliability coefficients as such are the solution independent of the type of study. Consistently replicated effects, e.g., produced by robust cognitive tasks, were found with low reliability coefficients (Hedge et al., 2018) suggesting that the relationship between replicability and reliability is not simply positive. The phenomenon that a low reliability can potentially co-occur with high replicability of an effect and high power is referred as a paradox (Overall & Woodward, 1975). The paradox refers to the possible co-occurrence of low reliability and high power in experimental (or quasi-experimental) studies due to the homogeneity of the subpopulation.

The often-assumed positive association between reliability and power of experimental studies has long been debated

---

The authors would like to thank Theodore P. Beauchaine, Robert Cudeck and Jolynn Pek for their comments and discussions on this topic.

---

✉ Selena Wang  
wang.10171@osu.edu

<sup>1</sup> Department of Psychology, The Ohio State University, Columbus, OH 43210, USA

(e.g., Fleiss, 1976; Hedge et al., 2018; Overall & Woodward, 1975; Overall & Woodward, 1976; Zimmerman & Zumbo, 2015). The controversy can be attributed to the fact that reliability is a population-dependent concept (Mellenbergh, 1996, 1999; Zimmerman & Zumbo, 2015); an estimate of the reliability is highly influenced by the subpopulations from which samples are drawn. The reliability coefficient as reported for a study indicates more than the quality of a measure; it is also a feature of the subpopulation, i.e., its homogeneity vs. heterogeneity. The more heterogeneous the subpopulation, the higher the reliability coefficient. At the same time, the power of an experimental study is lower as a result of the larger standard error because variance (i.e., true variance) is a component of the standard error. Semantically speaking, the steady association of a high reliability with a low or even an absence of error variance may seem natural, but it is not consistent with the mathematical definition of reliability in the Classical Test Theory (Zimmerman & Zumbo, 2015). Following the classical definition, true variance and error variance are the two components of test scores. Reliability is the portion of true variance relative to the observed variance, which implies that a high reliability can also stem from a large true variance, and a low reliability can stem from a small true variance, as in a homogeneous subpopulation.

Discussions of the replication crisis often involve descriptions of research findings as “unreliable” (e.g., Button et al., 2013; Tressoldi, 2012; Stanley & Spence, 2014), and the replicability of research findings has also been conceptually associated with the idea of consistency and stability. It is therefore not surprising that the reliability coefficient has been associated with the replicability of research findings. A first reason for our study is to clarify that in between-group studies, a high reliability coefficient can be associated with a low power rate, and therefore with less consistency of research findings, not just in theory but also in practice. Certain recommendations to counter the replication crisis fixated on improving reliability can be too general and should be interpreted in a nuanced way, as will be explained. A second but related reason is that the choice of a subpopulation and its homogeneity vs. heterogeneity is an under-discussed issue with consequences for power in between-group studies.

We focus on between-group studies (experimental or quasi-experimental) in this manuscript. They are also called between-subjects studies and need to be differentiated from within-subjects studies and from individual differences (i.e., correlational) studies. The term “reliability paradox” or simply “paradox” refers to the possible opposition between reliability and power in experimental studies, whether the design is between subjects or within subjects, in line with the use of the term by Overall & Woodward (1975, 1976). The same term can also be interpreted

as the differential role of reliability in experimental studies versus individual differences studies. The reliability coefficient of a variable is an index of how well the variable consistently differentiates between individuals, so that a higher reliability coefficient is unambiguously beneficial for individual differences studies. In this study, we use the term exclusively in the former sense, not for the contrasting role of reliability in experimental studies and individual differences studies. To avoid possible misunderstandings, we will explain the differential role of the reliability coefficient in individual differences studies versus experimental studies to emphasize that our analysis and its conclusion are applicable for between-group studies, but not for individual differences studies. We will also differentiate between between-subjects studies and within-subjects studies because for within-subjects studies, the paradox concerns the reliability of the intra-individual changes in the dependent variable (i.e., difference scores), and not the dependent variable itself. We will explain, in our discussion of the literature, why this is a complication that has led us to focus on between-subjects studies. Another study would be required to relate the reliability coefficient of the dependent variable itself to the power of a within-subjects study.

We explicitly address the role of three important choices associated with planning for a between-group study: (1) the measure for the dependent variable, (2) the sample size, and (3) the subpopulation. To be clear, we use the term “subpopulation” to refer to the subset of the general population from which a sample is drawn. For example, the set of psychology college students is a subpopulation of the U.S. population. The measure and the sample size are commonly agreed to be important considerations for upholding the standards of psychology research. The size of the sample is important for the power of a study and the credibility of the effect. High reliability coefficients of the DV and high power values are sought after to achieve those standards; high reliability is often coveted for the benefit of power (e.g., LeBel & Paunonen, 2011).

In this paper, we draw attention to the role of the subpopulation when planning an experiment. The subpopulation and its homogeneity vs. heterogeneity are much less explicit points of attention than the other two choices (the DV measure and the sample size). The prevalent use of homogeneous samples, such as freshmen psychology students and MTurk recruits, may not be the optimal choice for psychology research, yet subpopulation concerns are rarely addressed in published writings. The homogeneity of the subpopulation is a crucial factor in the controversial relationship between reliability and power (e.g., Fleiss, 1976; Hedge et al., 2018; Overall & Woodward, 1975; Overall & Woodward, 1976; Zimmerman & Zumbo, 2015).

In our study, we will show that homogeneity of the subpopulation has consequences for the relationship between reliability and power. Depending on the subpopulation, researchers would not be able to obtain high reliability coefficients and high values of power in the same study. This phenomenon will be explored in detail in the following sections.

In this paper, we first review and investigate the relationship between reliability and power for between-group designs in the context of the classical test theory. We reiterate the core issue and provide an overview of the literature on the controversy. We then go beyond the classical test theory and discuss the relationship between reliability and power based on the latent variable framework. In the process, we seek theoretical evidence that supports and explains the reliability paradox while formalizing not only the role of the measure but also the role of the subpopulation and its homogeneity. We demonstrate our findings using a visual illustration based on variations of parameters important for the relationship between reliability and power. We also seek empirical evidence via re-analyzing data from two published large-scale between-subjects replication studies. In these direct replications, the procedure and the measure of the dependent variable are the same, and the only difference is the subpopulation. Even though the subpopulations may seem rather equivalent on purpose, the empirical evidence suggests they are not in terms of homogeneity. In this way, we empirically demonstrate the role of the subpopulation. Finally, we reflect on possible reasons behind the persistence of common notions related to reliability, formulate recommendations, and discuss limitations of our study and the role of a latent variable framework.

### Why low reliability can be observed with high power

Statistical power is defined as the probability of rejecting the null hypothesis when it is false. A high statistical power indicates a high probability of detecting a true effect. When the reliability of a measure is low in a high-power study, a counter-intuitive situation occurs, where a low reliability is associated with a high power, and thus with a high replicability.

We first explain the core of the issue through equations for the reliability coefficient and for the standard error of an effect. An (unstandardized) effect is, in most cases, the difference between two means. The standard error is required not only for power calculations and null hypothesis significance testing, but also for confidence intervals. Therefore, our investigation is relevant for the null hypothesis testing and beyond (e.g., Cumming, 2013; Cumming, 2014). By directly looking at the standard error instead of power, the role of the subpopulation and its homogeneity is also easier to understand.

The reliability coefficient, in its simplest form, is based on the decomposition of the total variance into the true variance and error variance:

$$\text{Reliability} = \frac{\text{True Variance}}{\text{True Variance} + \text{Error Variance}}. \quad (1)$$

Holding the other variance constant, reliability increases with an increase in the true variance and decreases with an increase in the error variance, which has consequences for the standard error of the effect:

$$\text{SE} = \sqrt{\frac{\text{Variance}}{N}}. \quad (2)$$

As shown in Eq. 2, the standard error of an (unstandardized) effect is the square root of the ratio between the (sampling) variance and the sample size. The variance and the sample size depend on the design. For a within-subjects design, the variance refers to the variance of the pairwise differences, and the sample size refers to the number of pairwise differences. For a between-subjects design, assuming equal sample size and equal variance of the groups, the variance refers to two times the within-group variance (estimated through pooling); and the sample size refers to the sample size per group.

Following the classical test theory, the variance of the test scores, regardless of the experimental design, is the sum of the true variance and the error variance. Thus, in the within-subjects design, the reliability is the reliability of the pairwise differences; whereas in the between-subjects design, the reliability is the reliability of the measure scores. The true variance and the error variance of the pairwise differences are not directly related to those of the measure scores. The relationship between the reliability of the pairwise differences and that of the measure scores is undetermined unless highly constraining assumptions are made (e.g., Nicewander & Price, 1983; Levin, 1986). For example, in an intervention study, the measure scores in the pretest can be highly reliable while the differences between pretest and posttest are unreliable; and it is also possible that the pretest scores are unreliable while the differences between pretest and posttest are highly reliable.

The situation is much clearer for the between-subjects design because the true variance and the error variance that constitute the reliability coefficient of measure scores can be used directly in Eq. 2. More specifically,

$$\text{SE} = \sqrt{\left(\frac{2 \times (\text{True Variance} + \text{Error Variance})}{N}\right)}. \quad (3)$$

The standard error of the estimated effect, the ostensible inverse of power, is a positive function of both the true variance and the error variance while the reliability of the dependent measure is the ratio of the true variance to the total variance. Therefore,

- an increase of the true variance leads to an increase of the reliability, an increase of the standard error and a loss of power.
- an increase of the error variance leads to a decrease of the reliability, an increase of the standard error and a loss of power.

A higher reliability can be associated with a gain of power as well as with a loss of power. Both situations can occur depending on how much the true variance varies in comparison to the error variance. One extreme is when the true variance is reduced to zero. In such a case, a zero reliability yields maximum power given constant error variance. The other extreme is when the error variance is reduced to zero. In this case, perfect reliability yields maximum power given constant true variance.

### Literature on the relationship between reliability and power

The early discussion of the relationship between reliability and power focuses on the reliability in the within-subjects designs (Fleiss, 1976; Overall & Woodward, 1975, 1976), which is confounded with the reliability of change scores. Overall and Woodward (1975) point out the “paradox in the measurement of change” showing that the maximum power of a study is reached when the variance of the difference scores (differences between pre and post scores) consists of only error variance, i.e., the only-error assumption. The reliability of the difference is zero when the only-error assumption applies. This paradox refers to the combination of two seemingly contradictory elements: maximum power and minimum reliability. The only-error assumption also implies that the true effect is the same for all individuals (i.e., the true variance for the difference scores is zero). In a response to Overall and Woodward (1975), Fleiss (1976) argues that the only-error assumption is unrealistic because, in his view, there should always be an interaction between the individuals and the treatment, i.e., there always should be individual differences in the effect of a treatment. Fleiss (1976) further argues that the maximum power is reached when the error variance is zero implying perfect reliability. In their rejoinder, Overall and Woodward (1976) reassert the paradox stating that “other things constant” (p. 776), power has its largest value when the reliability is zero. In a follow-up article, Nicewander and Price (1978) explains that both points of view are correct. Power increases with reliability if the error variance is reduced, and it decreases with reliability if the true variance is increased. Nicewander and Price (1978) refer to the between-subjects designs for this synthesis. The synthesis is restated in several more recent articles by Zimmerman and Williams (1986),

Zimmerman et al. (1993), and Zimmerman and Zumbo (2015). Focusing on the reliability of individual differences, Parsons et al. (2019) state that the robustness of an effect does not imply a high reliability of individual differences in a simple measure and that the reliability of individual differences in an experimental effect (i.e., a difference measure instead of a simple measure) is important when one is interested in the latter type of individual differences. In our paper, we differentiate individual differences research from experimental research, and we argue that, for a between-subjects design experiment, the reliability paradox should be taken into consideration.

Part of the literature on the relationship between reliability and power is inspired by the initial focus on within-subjects designs and the measurement of change. While the core issue for within-subjects studies and between-subjects studies is the same—the issue is rooted in the decomposition of variance into true variance and error variance—there also is an important difference. The difference is that, in between-subjects designs, the decomposition is of the scores from the measures; whereas, in within-subjects designs, the decomposition is of the difference scores. It is an important distinction because reliabilities of the difference scores and reliabilities of the simple measure scores are influenced by different factors of the study. In within-subjects studies, the relationship between reliability of a simple measure and power is indeterminate, which is pointed out, among others, by Levin (1986) and Collins (1996). In within-subjects studies, selecting a measure based on its reliability does not improve power, except with the special circumstance outlined in Nicewander and Price (1983) and Levin (1986). These authors show with derivations that, in within-subjects studies, the relationship between reliability and power is positive across measures that fulfill two conditions: (1) the true scores of the measures are linearly related, and (2) the true scores in the control condition and the experimental condition are linearly related. Condition 1 is fulfilled for measures of the same construct. Condition 2 is fulfilled if the true score changes are equal for all individuals or are perfectly correlated with the true scores in the control condition, both of which are rare.

In summary, the relationship between reliability and power is unclear for within-subjects studies. For between-subjects studies, Sutcliffe (1958), Cleary and Linn (1969), and Hopkins and Hopkins (1979) believe that power increases with an increase in the reliability of the measure. The silent assumption is that this relationship is a relationship between power and reliability across different measures given the same subpopulation. The possibility of different subpopulations with varying heterogeneity is not considered. As argued by De Schryver et al. (2016), conditioning on the measure, reliability and power would be negatively related across subpopulations if the subpopulations differ in



heterogeneity. As high power is sought after in experimental studies, this negative relationship would lead to a selection of homogeneous subpopulations at the cost of reliability. Together with Overall and Woodward (1975, 1976), Zimmerman and Williams (1986), Zimmerman et al. (1993), Zimmerman and Zumbo (2015) and Nicewander and Price (1978), De Schryver et al. (2016) pose an opposing voice to the more common belief that an increase in reliability always indicates higher replicability of findings (e.g., Asendorpf et al., 2013; Funder et al., 2014; LeBel & Paunonen, 2011).

Different from past literatures that focus on reliability and power from a classical test theory perspective, we investigate the relationship based on a latent variable model. The latent variable framework is a more recent and more general framework. Latent variables replace the notion of true scores in classical test theory. In psychology, theories are often formulated in terms of constructs; and constructs can be thought as latent variables underlying a set of observed dependent variable measures. The latent variable framework lends itself better to (considering) different measures and subpopulations at the same time. In such a process, these different measures should be measures of the same latent variable. The measures can differ from one another with respect to their loadings and residual variances.

## Theoretical latent variable model

Our investigation is based on a between-subjects design because, as explained earlier, the relationship between reliability of different scores and power is indeterminate without further specific and perhaps unrealistic assumptions about the nature of change within individuals. The nature of intra-individual change is an important but different topic. We also investigate the standard error and the  $d$ -measure of effect size (Cohen, 1962). The standardized effect size (unstandardized effect divided by the standard deviation of the measure scores) is related to the variance of the measure scores.

## Model and equations

We study the experimental effect, e.g., the effect of an intervention, in a between-subjects design using a theoretical<sup>1</sup> latent variable model. Consider a between-subjects study with  $N$  subjects per condition and  $J$  possible

<sup>1</sup>We are in the theoretical discussion of this framework to illustrate the relationship between reliability and power. This means that we consider the parameters defined and discussed in this model to be true parameters and not estimates. Additional constraints and considerations must be taken into account if this model is to be estimated.

DVs. A DV can be a single measure (e.g., response times) or a measure with one or multiple items (e.g., sum scores). Using an observable DV, researchers intend to measure a construct of interest. For example, an attitude construct can be measured through responses to a set of questions regarding attitudes (e.g., Albarracín & Shavitt, 2018). Often, only one measure or DV (possibly with multiple items) is used in an experimental study. Using the latent variable model as a framework, we consider a study with  $J$  possible DVs measuring the same latent construct. The experimental effect is exerted on the latent variable (i.e., on the construct) and is shown through each of the DVs. A path diagram for such a design is shown in Fig. 1.

The group effect (on the observed scores) is manifested through the difference between the means of the observed scores in the two groups. In theory, the effect on the latent variable (a latent DV) can be estimated if a latent variable model<sup>2</sup> was used with multiple DVs. In this paper, we investigate the effect on the observed scores as commonly seen in psychological studies with an experimental design.

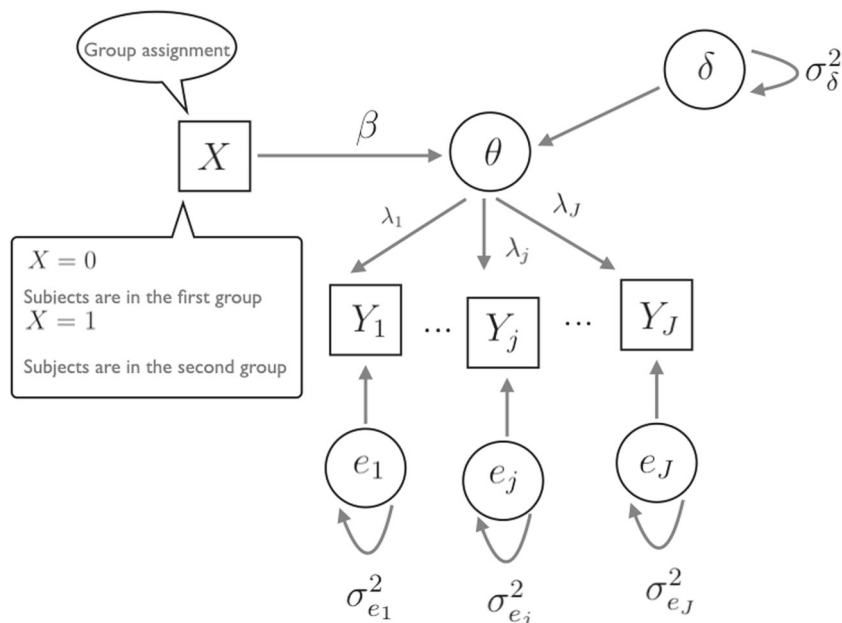
The model is presented in Eq. 4. Here and in what follows, we consider the reliability and power associated with  $DV_j$ , and therefore for simplicity, we omit the subscript  $j$  in the following model equation:

$$y_i = \lambda * \beta * x_i + \lambda * \delta_i + e_i \quad i = 1, 2, \dots, 2N, \quad (4)$$

where the latent variable,  $\delta$ , and the error term,  $e$ , are assumed to be independently and normally distributed. The factor loading  $\lambda$  indicates the degree of association between the latent variable and the DV. Coefficient  $\beta$  is the regression coefficient of  $x_i$  (denoting the independent variable with values 0 and 1) and represents the experimental effect on the latent variable. The product,  $\lambda\beta$ , reflects the experimental effect on the observed scores. The power of a two-sample  $z$  test (the within-group variance is assumed to be known in this theoretical discussion), the reliability for  $DV_j$ , the standard error (SE), the Cohen's  $d$  (i.e., the standardized or relative effect size) and their relationships are derived in the Appendix. A summary of the relationship between reliability and power through shared parameters appears in Table 1.

In the table, + signals positive association and – signals negative association. Keeping the other parameters constant, we find that

<sup>2</sup>The latent variable model is not a common approach for experimental studies. The latent variable model requires large sample sizes and multiple measures. However, these restrictions are not a problem when the latent variable model is used as a framework to understand the relationships between reliability and power. To be clear, we do not suggest using latent variable models for the analysis of experimental studies; and it would be incorrect to think that our conceptual analysis does not apply when latent variable models are not used in a study.



**Fig. 1** The path diagram of the latent variable model. The scores of  $2N$  subjects of DV  $j$  are represented by  $Y_j$ ,  $j = 1, 2, \dots, J$ . The parameters of the model are:  $\lambda_j$ , the factor loading of DV  $j$  on  $\theta$ ;  $\sigma_{e_j}^2$ , the residual variance of DV  $j$ ;  $\sigma_{\delta}^2$ , the within-group latent variable variance and  $\beta$ , the unstandardized regression coefficient for regressing  $\theta$  on  $X$ . We use  $X$  to indicate the group assignment. Subject  $i$  is in the first group when  $x_i = 0$ ; subject  $i$  is in the second group when  $x_i = 1$ ,  $i = 1, 2, \dots, 2N$ . We assume equal variance and strict measurement invariance between the two conditions

1. a larger regression coefficient yields a larger effect size and larger power, without consequences for the standard error and reliability;
2. a larger latent variable variance leads to a smaller effect size, less power, a larger standard error and higher reliability;
3. a larger error variance yields a smaller effect size, less power, a lower reliability, and a larger standard error;
4. a larger factor loading leads to a larger effect size, more power, a larger standard error and higher reliability.

Note that points 1 and 4 also apply to the absolute effect size, and not just to Cohen's  $d$ .

As far as between-group studies are concerned, previous recommendations to increase power by increasing reliability are based on the 3rd inference (above), which is that minimizing measurement error variance raises both reliability and power (Zimmerman & Williams, 1986; Zimmerman et al., 1993; Humphreys, 1993). These recommendations are also reasonable when the increase in reliability stems from an increase in the factor loading<sup>3</sup>, as shown in the 4th inference in Table 1. The relationships follow directly from the

<sup>3</sup>A higher factor loading does not imply a smaller error variance in this latent variable model. Loadings and error variances are different parameters.

latent variable model. It is clear that a higher reliability does not always indicate a higher power. Following Table 1, one can find conditions for which the correlation between reliability and power is positive and other conditions for which the correlation is negative. To use reliability as a tool against the replication crisis (when the experimental design is a between-group design), researchers should consider conditions under which a higher reliability is beneficial to power. From Table 1, we also find that Cohen's  $d$  changes in the same direction as power and that the standard error increases with increases in all parameters, except for  $\beta$ .

The above relationships between reliability and power apply to different types of reliability coefficients. When estimating reliability, different assumptions are applied to estimate the true variance and the observed variance in Eq. 1. As long as they are good approximate estimates, the same conclusions can be drawn. For example, the equal loadings assumption is made for estimating Cronbach's alpha (Cronbach, 1951), which is often used for sum scores (e.g., De Boeck & Elosua, 2016; Sijtsma, 2009). Without the equal loadings assumption the same parameters still influence the relationship between reliability and power. Therefore, our conclusions would still be true. Similarly, our conclusions can be applied to other estimates of reliability: the parallel-test reliability (Guttman, 1945), split-half reliability (Spearman, 1910), glb (Sijtsma, 2009), and McDonald's omega (McDonald, 1999).

**Table 1** Relationships of the latent variable model parameters with  $d$ , SE, power, and reliability conditional on other parameters

Parameters	$d$	SE	Power	Reliability
$\beta$ (regression coefficient)	+		+	
$\sigma_{\delta}^2$ (latent variable variance)	–	+	–	+
$\sigma_{\epsilon}^2$ (error variance)	–	+	–	–
$\lambda$ (factor loading)	+	+	+	+

## Illustration

The equation-based relationship between reliability and power, as summarized in Table 1, has direct consequences for the correlations between reliability and power across studies of the same effect. To illustrate this, we imitate replication studies in four hypothetical conditions, where different variations in the shared parameters are observed. The variations in the parameters are not meant to represent the ideal of replications, but rather to illustrate the implications of measures and subpopulations for reliability and power. For illustrative reasons, we work with uniform distributions of the parameters shown in Table 2.

Suppose an effect is replicated 20 times, an approximate number of studies in Hagger et al. (2016), and each time, a between-subjects study is implemented. In Table 2, we define the different uniform distributions of the parameters in the four conditions. As can be seen, in conditions 1, 2, and 3, the variation is large for only one parameter ( $\sigma_{\epsilon}^2$ ,  $\lambda^2$ , or  $\sigma_{\delta}^2$ ), while the variations of the other two parameters are small. In condition 4, the variations of all three parameters are large. The ratios between upper and lower bound of the parameter intervals are 1.1 to 1.25 when the variations of parameters are small. The ratios are 2.5 to 3 when the variations of parameters are large—two to three times larger than the small intervals. These variations are created through repeated random sampling from uniform distributions.

Cohen (1992) categorizes standardized effect size  $d = 0.2, 0.5$  and  $0.8$  as small, medium, and large, respectively. We will consider each of these. When  $\beta = 1$ , we use the assigned parameter values to obtain a distribution of

standardized  $d$  values centering around 0.8. Similarly, when  $\beta = 0.6$  and  $0.25$ , the means of the  $d$  distributions become approximately 0.5 and 0.2. Using each set of these  $\beta$  values and the sampled parameter values, we can calculate reliability, power and their correlation across the 20 replication studies. This process is repeated 100 times for each of the four conditions and for each of the three different  $\beta$  values.

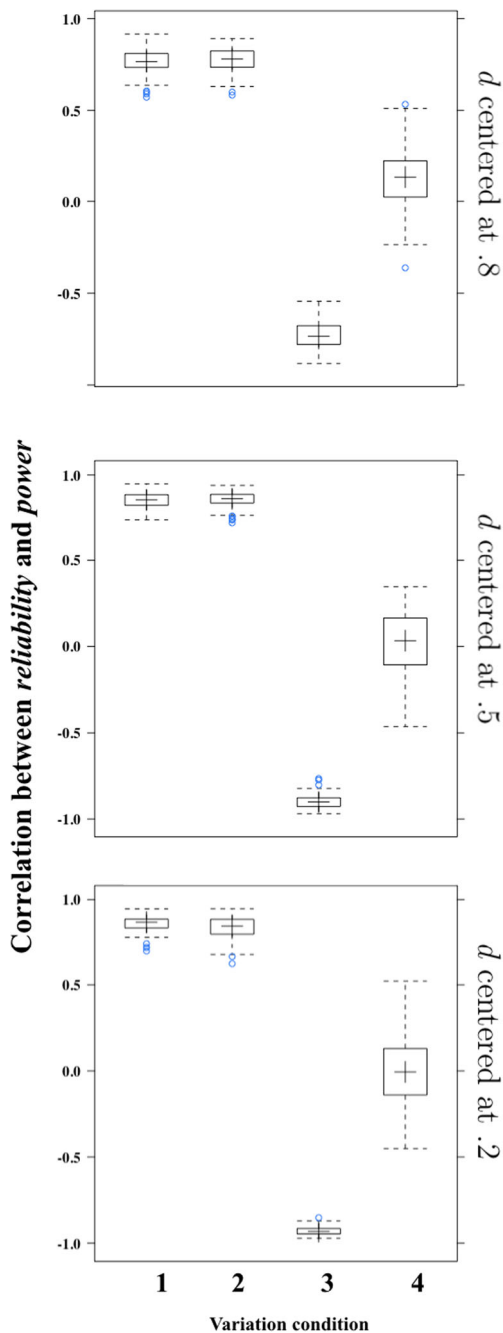
In Fig. 2, we present the correlations between reliability and power in the first three conditions with small, medium, and large effect sizes. As can be seen, the correlations between reliability and power are positive and close to 1 when the variation of either  $\lambda^2$  or  $\sigma_{\epsilon}^2$  is large, and the variations of the other two parameters (including  $\sigma_{\delta}^2$ ) are small. The correlations become less positive and more negative as the variation of  $\sigma_{\delta}^2$  becomes larger. This change in the signs of the correlations is consistent across different effect sizes. The negative signs of the correlations between reliability and power depend strongly on the variation of  $\sigma_{\delta}^2$ . In condition 3, where the variation of only  $\sigma_{\delta}^2$  is large, and the variations of the other two parameters are small, the correlations are highly negative and close to  $-1$ ; see Table 3.

In condition 3, only the variation of  $\sigma_{\delta}^2$  is large, and  $\beta$  is assigned values 1, 0.6 and 0.25. The magnitude of power decreases with decreases in the effect sizes (i.e., the means of  $d$  decrease roughly from 0.8, 0.5 to 0.2). Despite this change in the magnitudes of power and  $d$ , the correlations between reliability and power remain strongly negative. This illustrates that the strong negative correlation between reliability and power is driven by changes in the true variance,  $\sigma_{\delta}^2$ , regardless of the size of the effect. In

**Table 2** The uniform distributions of  $\lambda^2$ ,  $\sigma_{\epsilon}^2$  and  $\sigma_{\delta}^2$  across 20 replications of an effect in four different conditions

Condition	$\lambda^2$ (squared loading)	$\sigma_{\epsilon}^2$ (error variance)	$\sigma_{\delta}^2$ (latent variable variance)
1	(0.60, 0.68)	(0.18, 0.54)	(0.9, 1.1)
2	(0.36, 0.92)	(0.32, 0.40)	(0.9, 1.1)
3	(0.60, 0.68)	(0.32, 0.40)	(0.5, 1.5)
4	(0.36, 0.92)	(0.18, 0.54)	(0.5, 1.5)

Note: Each interval represents a uniform distribution of a parameter



**Fig. 2** The correlations between reliability and power in the four conditions with large, medium, and small effect sizes

condition 4, where the variation of all parameters is large, the correlation between reliability and power is close to zero. Together, the illustration shows that the correlation between reliability and power can range from extremely high and positive to extremely high and negative. The sign and magnitude of the correlations depend on features of the

studies including the dependent variable measure and the subpopulation.

## Empirical examples

For the empirical study of the relationship between reliability and power, we re-analyze data from large-scale replication studies with a between-group design. As mentioned before, these replication studies use the same procedures and the same measures, but they differ in subpopulations as the replications are conducted by investigators from different parts of the world. However, there are no manifest indications that the homogeneity of the subpopulations is different based on their demographics. It is of course possible that the homogeneity of the subpopulations differs empirically. If a negative relationship between reliability and power is observed under these conditions, then the difference in subpopulations is likely the culprit. If this is empirically confirmed, then it is a fortiori true for other studies in which less attention is paid to the selection of subpopulations.

**Replication studies of the grammatical aspect effect**  
Eerland et al. (2016) organized a large-scale replication study with 12 studies on whether an individual's perceptions of others' actions are influenced by the grammatical aspect (i.e., imperfective versus perfective) of the language used to describe the events. The effect of an action description using imperfective vs. perfective aspect was expected to show in three measurable ways: (1) stronger perceived intentionality of the actor (Intentionality), (2) more imagery in the mind of the perceiver (Imagery), i.e., more detail in the perceiver's imagination of the action, (3) stronger perceived responsibility of the actor for the action (Intention Attribution). The three DVs were respectively measured with 3, 4 and 3 self-report items, as reported by Eerland et al. (2016). The meta-analytic effect size for Intention Attribution was reported as 0, and thus the Intention Attribution variable was omitted for power calculations. By using the meta-analytic effect sizes for power calculations, the calculated power is conceptually closer to post hoc power than to "true" power, and yet we argue that this imprecision does not interfere with our illustration of the relationship between reliability and power. An underestimation or overestimation of power due to a different estimate of the effect size has no effect on its correlation with reliability, as has been demonstrated in the previous section. It is the variability of the shared parameters across studies that influences the signs of the correlations. The



**Table 3** The means of the distributions for reliability, power, Cohen's  $d$ , and the correlation between reliability and power in condition 3

	Reliability	Power	Cohen's $d$	Correlation
$\beta = 1$	0.6286	0.9997	0.8092	-0.7352
$\beta = .6$	0.6297	0.9188	0.4852	-0.9009
$\beta = .25$	0.6269	0.3015	0.2028	-0.9296

unstandardized meta-analytic effect sizes for Intentionality and Imagery were reported as  $-0.24$  and  $-0.08$ . Using these reported values, we calculated power for independent  $t$  tests with an alpha level of 0.05. We calculated Cronbach alpha<sup>4</sup> as estimates for the reliability of the measure scores. For Intentionality and Imagery, we present the relationship between reliability and power in the left and middle panel of Fig. 3, respectively. The corresponding correlations across the 12 studies are  $r = -0.75$  and  $r = -0.17$ .

**Replication studies of the ego-depletion effect** (Hagger et al., 2016) organized a large-scale replication of the ego-depletion effect with 23 studies (1 study was excluded from our analysis for lack of information). The study shows that there is not much of an effect, as later confirmed by other large-scale replication studies (Vohs et al., 2021; Dang et al., 2021). We did not directly analyze the ego-depletion effect itself because its DVs (reaction time variability, RTV and the mean reaction time, RT) are single measures, whose reliability coefficients could not be calculated. Instead, we analyzed the manipulation check variable (referred as the Arduousness variable in this paper): the mean of four self-reported items (i.e., effort, difficulty, fatigue, and frustration) that describe how demanding the tasks are perceived by participants. Power was calculated for an independent  $t$  test with an alpha level of 0.05 and an estimated unstandardized meta-analytic effect size of 1.21. The Cronbach's alpha was also calculated. In the right panel of Fig. 3, we present the relationship between reliability and power across the 22 replication studies. The correlation is  $-0.31$ , but we observe a ceiling effect in the power values for Arduousness, which we contribute to the large effect size estimate. For a moderate effect size, i.e., an unstandardized effect size of 0.4 and  $d \approx 0.5$ , the relationship between reliability and power is shown in Fig. 4, and the correlation is  $-0.49$ .

In Figs. 3 and 4, the relationships between reliability and power are consistently negative for the three different DVs. This result may seem surprising, but it is surprising only if one believes that higher reliability is always associated with smaller measurement error variance, which

is correct only when the true score variance is kept constant. In other words, the results from these two large-scale replication studies show that the constant true score variance assumption likely does not hold in practice and that there must be substantial variations of the true variances across studies in these two large-scale replications.

In Fig. 5, we show that the estimated reliability is positively associated with the pooled sample variances. In Fig. 6, we show that estimated power is negatively associated with the pooled sample variances. Together with the observed negative correlations between reliability and power, Figs. 5 and 6 provide real-world evidence for our theoretical model. They show that the results from our theoretical analyses are not unrealistic representations of the relationship between reliability and power in psychological research and that the traditional belief can be misleading.

## Discussion

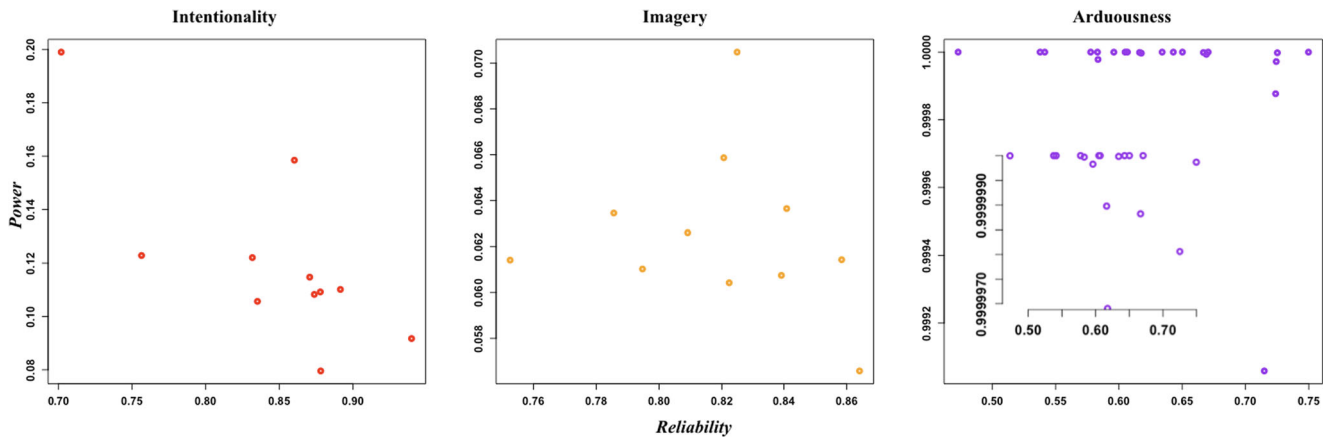
In this section, we first reflect on possible misconceptions about the reliability coefficient and how reliability relates to the power of studies. Next, we formulate three recommendations regarding the reliability coefficient for studies with a between-group design. Finally, limitations of our contribution and the implications of a latent variable framework are discussed.

### Reflections on reliability and power

Several explanations are possible for the steady misconception that a higher reliability coefficient is always associated with higher power, and thus more replicable research. The first is that authors do not always differentiate between semantically related terms such as reliability, consistency, precision, and dependability, the more precise meaning of which depends on its referent. The referent can be individual differences, or differences between experimental conditions, etc. We know that reliability refers to the consistency of individual differences and not the consistency of estimated effects of experimental conditions (Zimmerman & Zumbo, 2015).

Two other explanations, although related to the previous, refer to implicit beliefs. The first implicit belief is that the true variance does not vary, i.e., the invariance of the

<sup>4</sup>Cronbach alpha is a lower bound estimate of the reliability, and it is sometimes referred as an estimate of the internal consistency. It is the most popular reliability coefficient.



**Fig. 3** Power vs. reliability plots for dependent variables Intentionality (*left*), Imagery (*middle*), and Arduousness (*right*). The plots for Intentionality and Imagery use data from Eerland et al. (2016); the plot for Arduousness uses data from Hagger et al. (2016)

true variance. Variations of the reliability coefficients are commonly assumed to be stemming from variations of the measurement error variance, which implies that the true variance does not vary. Only rarely is the assumption made explicit as by Zimmerman and Williams (1986):

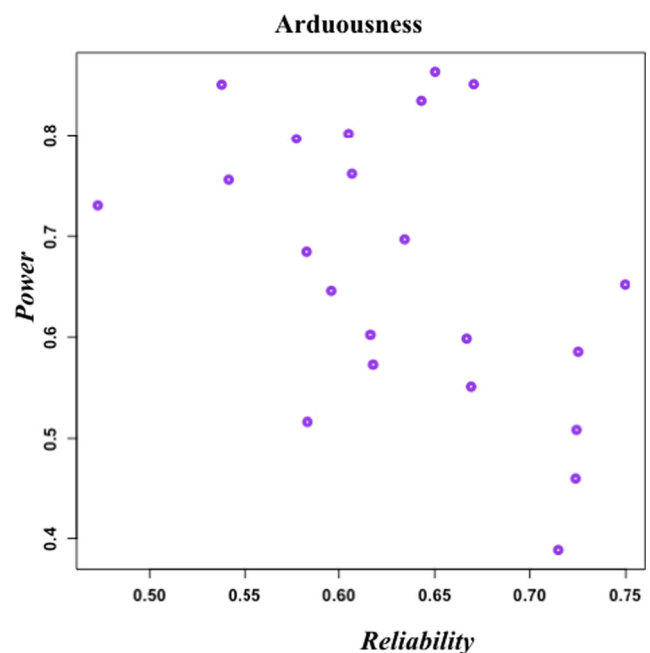
In experimental contexts, improvement in the reliability of a measure is usually interpreted as a reduction in error variance attributable to increased precision of an instrument or elimination of extraneous variables. Improvement in reliability in these contexts is not usually conceptualized as an increase in the heterogeneity of the group of subjects measured. (p. 124)

This usually silent assumption would be correct when we compare measures for the same subpopulation—this is the case when we compare measures within the same study—but there is no evidence supporting this assumption when we compare measures across different subpopulations. In fact, our re-analyses of the data from large-scale replication studies provide counterevidence. The negative correlations between reliability and power suggest that the true variance changes across different replications of the same effect.

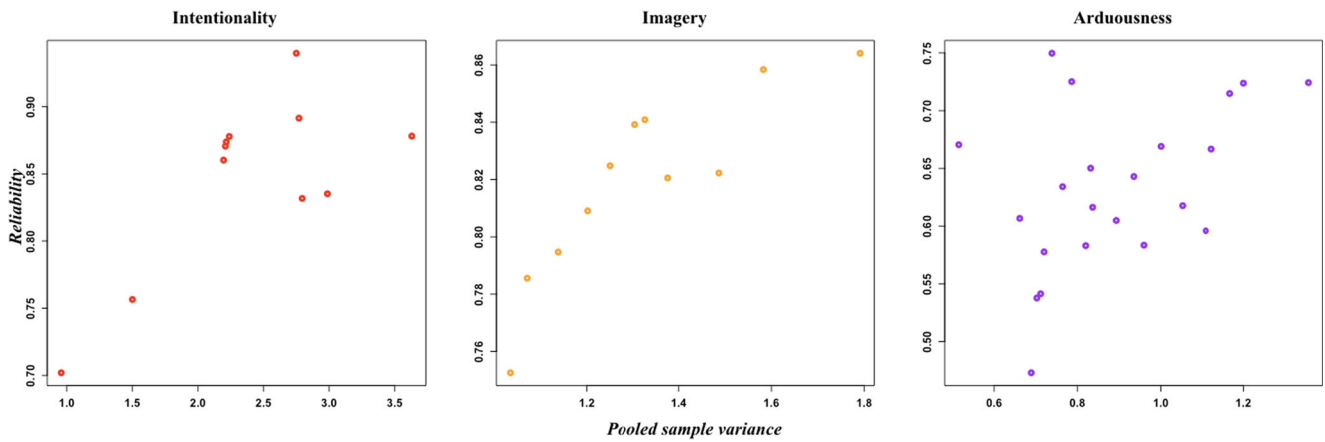
As unfortunate as it may be, equating a reduction in reliability coefficient with a reduction in measurement error variance is how reliability is interpreted and applied. This equivalence and the maybe unintentional ignorance of the role of true variance imply that reliability is a feature of the measure and not of the measure and the subpopulation. The practice that the reliability coefficient of a measure needs to be reported per study and needs to meet a certain threshold for a study to have credibility—even when there are indications in the test manual that the measure is reliable—stems from the belief that only irrelevant sources of variation in the study (i.e., error variance) make the measure unreliable. If the influence of the true variance and the homogeneity of the subpopulation were taken into

account, then using cut-off values to judge the quality of a measure in a study would not have been adopted as a standard practice for publication, whether the study is an experimental study or an individual-differences study. There is very little a researcher can do when the low reliability coefficient comes from low true variance. As Williams et al. (1995) repeats Gulliksen’s (1950) remark:

In general, when we give a test to two different groups and find that the standard deviation of one group is larger than that of the other group, we are dealing with a case where the true variance of one group is greater than that of the other group” (p. 109).



**Fig. 4** Power vs. reliability plots for Arduousness when the effect is of moderate size, i.e., the unstandardized effect size is 0.4 and  $d \approx 0.5$



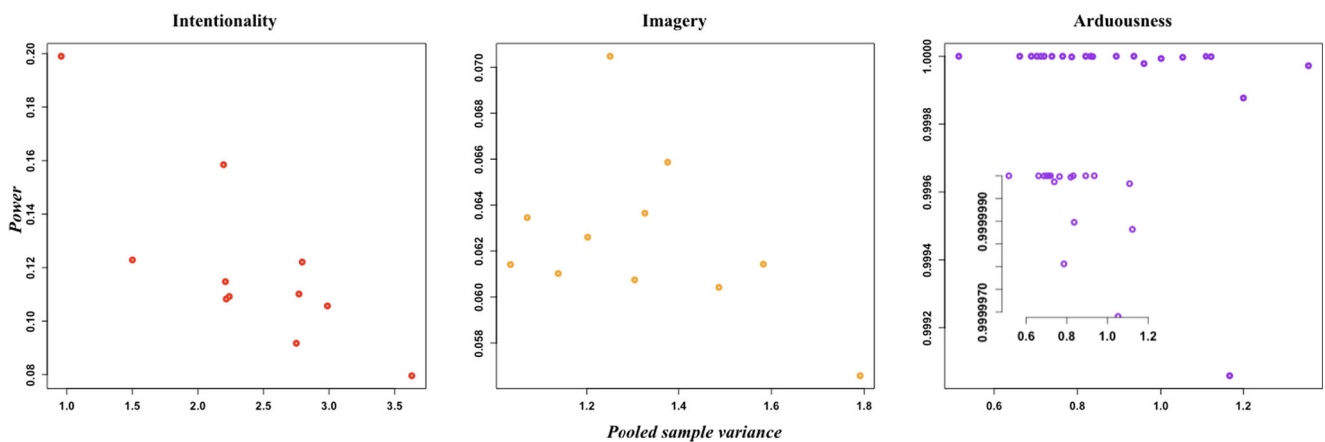
**Fig. 5** Reliability vs. pooled sample variance plots for dependent variables Intentionality (*left*), Imagery (*middle*) and Arduousness (*right*). The plots for Intentionality and Imagery use data from Eerland et al. (2016); the plot for Arduousness uses data from Hagger et al. (2016)

Vague population concepts may contribute to the invariance belief. Via statistical inference, information about the population is inferred based on the sample given that the sample is randomly drawn from “the population”. However, in practice, a sample is rarely randomly drawn from the general population. Instead, for reasons of convenience, it is drawn from subpopulations with possibly distinct characteristics and particularities defined by the chosen sample; and therefore, the samples are not representative of the ideal large population in the common sense. The writings about the study populations can be rather vague or even absent in published studies leading readers to believe that it does not make a difference whether the sample is from a subpopulation or from the idealized general population. This distinction between the general population and the actual study population (more likely the subpopulation) is rarely properly addressed in the research conclusions. The prevalent use of convenience sampling methods as well as the over-generalization of

study conclusions are reflections of the confusion and ambiguity about the concept of population. This ambiguity contributes to the preservation of the invariance belief.

The second implicit belief is that the error variances as part of the observed individual differences also apply to differences between conditions, e.g., between the experimental and control conditions. The reliability coefficient reflects the consistency of individual differences. The consistency of differences between conditions (i.e., the experimental effects) across studies is a different notion. The reliability coefficient quantifies the uncertainty about the former and cannot be generalized to the latter. High reliability of a measure does not imply that the experimental effects can be reliably observed.

Suppose we are interested in measuring the effect of a new diet on reducing weight, and a study is conducted with an experimental group and a control group. The reliability of the weight scale for the sample before the participants are randomly assigned to the two conditions describes how



**Fig. 6** Power vs. pooled sample variance plots for dependent variables Intentionality (*left*), Imagery (*middle*), and Arduousness (*right*)

consistent the weight scale distinguishes people with higher weight from those with lower weight in the sample. The reliability coefficient is a feature of both the measure and the subpopulation (from which the participants are sampled). The coefficient is higher both when the weight scale has more distinguishing power and when people's weights differ more from each other in the subpopulation. The reliability of the weight scale does not determine how reliable the effects of the new diet are or how consistent across studies the weight differences are between the control and the experimental groups. Rather, how invariable the diet effects are across studies and how successful the people are at following the diet determine how reliable the effects of diets are. It is possible that the effects of the diet are inconsistent across studies, and yet the scale is reliable. It is also possible that the diet effect is consistent across studies, and yet there are not enough weight variations in the subpopulation to sustain a high reliability.

Prior calls for greater attention to reliability have generally taken the form of warnings about the impact of poor-quality measures on power and replicability. These warnings about low reliability serve to remind researchers that they can improve power by reducing the random error component as much as possible. This is true for experimental studies as well as for correlational studies. The important distinction made here is that the reliability coefficient is not purely a reflection of measurement quality; it also reflects the heterogeneity of the subpopulation. For correlational studies, a high reliability is advantageous independent of the reason behind the high reliability, whether it is from a small error variance or a large true variance, or both. An increase in reliability implies higher discriminative power for individual differences. However, if the increase in reliability comes from a higher true variance, the discriminative power of between-group studies is reduced as far as the difference between the groups is concerned. The power of a between-group study depends on the manifest variance of the sample, which is the sum of the true variance and the error variance.

To better understand the role of these two components of the manifest variance, we offer the following example and reflections. Suppose that the true variance of the DV in study 1 is 1.00, and its error variance is 0.50, for a total of 1.50. Imagine now two different scenarios for study 2, both with a total variance of 2.0, compared to the 1.5 in study 1. In scenario A, the error variance is increased to 1.0 (0.5 larger than in study 1), and the true variance remains at 1.0 (as in study 1), while in scenario B, the true variance is increased to 1.5 (0.5 larger than in study 1), and the error variance remains at 0.5 (as in study 1). As a result, in study 1, the reliability is 0.67, while in study 2, it is 0.50 for scenario A and 0.75 for scenario B. Let us further assume that the independent variable  $X$  has an effect of 0.4 on the

construct variable (i.e., the latent variable) and that the effect of the construct variable on the observed DV is 1.0 (i.e., the unstandardized loading). The effect of  $X (= 0, 1)$  on the DV follows from the equation for the expected value of the DV:  $E(Y) = 0.4 * \lambda * X + E(\delta) * \lambda + E(e)$ . The expected error,  $E(e)$  is 0; the factor loading,  $\lambda$  is 1; and the expected latent variable,  $E(\delta)$  is the same in both conditions. The expected effect is the difference in  $E(Y)$  for the case  $X$  is 1 instead of 0. For study 1 and for both scenarios of study 2, the expected effect is 0.4 because  $\lambda = 1$ , and  $E(\delta)$  is the same in both conditions. The effect is independent of the sample variance<sup>5</sup>. However, it is common practice in psychology to express the effect size in relative terms (i.e., Cohen's  $d$ ; see Pek and Flora (2018) for a different view), and the sample variance does make a difference for the relative effect. It is 0.327 in study 1, and 0.283 in study 2 independent of the scenario (i.e., 0.4 divided by the within-group SD, which is  $\sqrt{1.5}$  and  $\sqrt{2}$  for study 1 and study 2, respectively).

To interpret an effect, one can either scale it relative to the sample or not scale it. Parsons (2018) follows the former line of thinking and rightfully interprets the attenuation as a smaller (relative) effect following a larger variance from increased error or true variance, or both. In this case, the loss of power due to a larger variance may be considered less of a problem since the size of the relative effect is reduced anyway. While this is true, it can also make sense to stay with the absolute effect. For example, a weight loss in terms of pounds is the same whether obtained in a study with a homogeneous sample or a heterogeneous sample. In a similar way, the absolute effect of a depression treatment may be more intuitive for a patient, and it may also be more important scientifically (considering its invariance) for the evaluation of the treatment. The feature of invariance for the absolute effect is an advantage from a robustness point of view. For the same quality of treatment and the same DV, one may expect the effect to be the same independent of the homogeneity of the sample. If one is interested in the size of the absolute effect, the detection of that effect (using the p-value) is less likely not only when the error variance is large, but also in a heterogeneous sample. In a heterogeneous sample, not only is the size of the absolute effect the same, but also the quality of the measure as such, as long as the error variance does not increase. At the same time, the effect is less likely to be detected.

In sum, depending on the effect, relative vs. absolute, the consequences of a larger manifest variance change, and so does the meaning of the consequences. When opting for

<sup>5</sup>Our analysis also implies that the absolute effect does not depend on the reliability coefficient as such. However, a smaller loading of a DV (i.e., its connection with the latent variable) leads to a lower reliability (through the reduction of true variance) and to a smaller absolute effect. Therefore, if a smaller reliability is due to a smaller loading, then it is associated indeed with a smaller absolute effect.

the relative effect, the size of the effect already takes into account the larger variance, therefore it is not surprising that the *p*-value is larger as well. When opting for the absolute effect, which is usually how an effect is tested, the size of the effect does not shrink with a larger variance, and a larger *p*-value follows from the larger variance and the resulting larger standard error.

The three choices associated with planning an experiment—sample size, subpopulation and measure—are shown to have different effects on power and reliability. The effect of the sample size is evident. Larger sample size is beneficial for power, but it does not affect the size of the reliability coefficient except that there is more information in the reliability estimate. The choice of the subpopulation and its influence on power and reliability are perhaps lesser-known points of contention. As has been shown in this study, the choice of the subpopulation affects not only power and reliability in one study, but also the relationship between power and reliability across studies.

## Recommendations

Based on our study, we formulate three different recommendations for between-group experimental studies. The recommendations concern (1) study-specific reliability coefficients as a criterion for the quality of a study, (2) the selection of a measure for the DV, (3) the selection of a subpopulation.

First, for between-group experimental studies, a low reliability coefficient is not necessarily a counterindication for the quality of the study if there are indications from other studies that the measure is reliable and if the low reliability is due to the homogeneity of the subpopulation. The latter can be checked with the observed variance. Experimental studies are not set up to investigate differences between individuals but to investigate differences between conditions instead. As we have shown, the discriminative power for between-group experimental effects is larger in studies with homogeneous samples. The situation is different if individual differences and their correlations are investigated because the discriminative power for individual differences is higher when the reliability coefficient is higher, a point made by Cooper et al. (2017) and Parsons et al. (2019) among others. It also means that if the effect of a treatment from a within-subjects design is used as a measure to be related to other variables (when a correlation is estimated), it is important that the effect measure in question is reliable.

Second, when selecting measures for an experimental study with a between-group design, a researcher is recommended to compare prior reliability information from roughly equally heterogeneous samples. For example, meta-analytic results may help on the condition that information about the heterogeneity of the samples is available in the

study. Furthermore, because small individual differences lead to low reliability coefficients, independent of the quality of the measure, reliabilities of measures should be compared using rather heterogeneous samples. In this way, the measures are given the opportunity to show their reliabilities.

Third, for the selection of a subpopulation during the design of an experimental study, one may consider homogeneity as a desirable feature. Homogeneous subpopulations increase the power rate of the study even though they yield smaller reliability coefficients. However, an important caveat is that the (absolute) effect under consideration is not modified by the subpopulation, which implies that there should not be an interaction effect of the subpopulation and the independent variable. The absence of interaction needs to be argued to avoid that the result of a study is specific for the selected subpopulation. For the sake of generalization, studies with homogeneous samples should be replicated across samples from different levels of the measure.

A possible source of confusion is that sometimes an effect itself can be considered a measure, such as the Stroop effect. If the individual differences in the effect are small, the reliability of the effect as an individual differences measure is small, but the power of the study to detect the mean effect is large because its standard error is small. However, if the effect is used as an individual differences (correlational) measure, it should of course be reliable. The true variance of the effect is disadvantageous for the power to detect the mean effect, but it is advantageous for the power of a study in which the size of the effect (for different individuals) is used as a measure to be correlated with other variables.

## Limitations and the latent variable framework

Our investigation of the relationship has some restrictions. First of all, we have focused on between-group studies. We have outlined a latent variable framework that, although not covering all possibilities, covers what we believe to be a common scenario (i.e., the between-subjects design). The issue is far more complicated for within-subjects designs, which deserves a separate study. Second, we have assumed measurement invariance across the experimental and control conditions. Measurement invariance cannot be investigated with classical test theory. The advantage of a latent variable framework is that measurement invariance is made explicit and can be investigated. We have also assumed the measures to be measures of the same construct. This assumption can also be investigated with latent variable models, but not with classical test theory. Transitioning from the classical test theory to a latent variable framework, unavoidable restrictions of the traditional theory become optional assumptions in the more flexible latent variable model. Without the assumptions, the relationship between



reliability and power would be even more ambiguous. Future studies should look into the effect of model violations on the issue of subpopulation homogeneity.

Although the latent variable framework is more complex and comes with several assumptions, the framework also has several advantages, such as the use of factor models for the development of scales and for comparing estimates of true variances in different subpopulations. Another possible advantage of the latent variable framework in the context of the replication crisis is the property of invariance for effects. In a latent variable model, effects are first formulated in absolute terms (although after the model is estimated, they can also be re-scaled as relative effects). The formulated absolute effects are invariant, as illustrated with our example that the 0.40 absolute effect does not change as a function of the sample variance. The invariance requires measurement invariance and is not empirically guaranteed. On the other hand, deriving relative effects from a latent variable model requires re-scaling of the absolute effect, which leads to differences in the relative effects across studies even when the corresponding absolute effects are equal. Between absolute effects and relative effects, there may not be a single best choice, but one should be aware of the perspective one is using and its consequences for interpretation.

**Author Contributions** The designs for the theoretical model, the simulation or demonstration and the analyses of empirical data were the result of the collaboration between Selena Wang and Paul De Boeck. Selena Wang proposed different designs, and Paul De Boeck advised, revised and proposed possible improvements and alternatives. Selena Wang derived the mathematical results in the theoretical model, conducted the simulation and the analyses of the empirical data. Paul De Boeck gave comments and suggestions throughout the process. The initial draft was written by Selena Wang. Paul De Boeck added sections to the draft. Both contributed to the rewriting and editing of the manuscript.

The data for all experiments are available at <https://osf.io/d3mw4/> and <https://osf.io/jymhe/>.

## Appendix

### Derivations of statistics in between-subjects studies

For the  $j$ th test (omitting subscript  $j$ ), the null and alternative hypotheses of the test on the true effect are formulated as follows.

$$H_0 : \lambda\beta = 0$$

$$H_a : \lambda\beta \neq 0$$

In the following two equations, we present reliability and power as functions of parameters defined in the above

model:

$$reliability = \frac{\lambda^2 * \sigma_\delta^2}{\lambda^2 * \sigma_\delta^2 + \sigma_\epsilon^2} = \frac{1}{1 + \frac{\sigma_\epsilon^2}{\lambda^2 * \sigma_\delta^2}}. \quad (5)$$

The numerator is the true variance and the denominator is the observed variance (true plus error). Assuming that the within-group variance is known and that  $\alpha = .05$ , we use power for a two-tailed <sup>6</sup> two-sample  $z$  test.

We assume equal and known variances between the two groups due to the theoretical nature of our latent variable model. Assuming that the dependent variable is i.i.d. normally distributed, at  $\alpha = .05$ , power for the two-sample  $z$  test is

$$\begin{aligned} power &= 1 - \Phi\left(z_{\alpha/2} - \frac{\lambda\beta}{\sqrt{Var(Y_{x=0})/N + Var(Y_{x=1})/N}}\right) \\ &\quad + \Phi\left(-z_{\alpha/2} - \frac{\lambda\beta}{\sqrt{Var(Y_{x=0})/N + Var(Y_{x=1})/N}}\right) \\ &= 1 - \Phi\left(1.96 - \frac{\lambda\beta}{\sqrt{(\lambda^2 * \sigma_\delta^2 + \sigma_\epsilon^2)/N + (\lambda^2 * \sigma_\delta^2 + \sigma_\epsilon^2)/N}}\right) \\ &\quad + \Phi\left(-1.96 - \frac{\lambda\beta}{\sqrt{(\lambda^2 * \sigma_\delta^2 + \sigma_\epsilon^2)/N + (\lambda^2 * \sigma_\delta^2 + \sigma_\epsilon^2)/N}}\right) \\ &= 1 - \Phi\left(1.96 - \frac{\lambda\beta}{\sqrt{(\lambda^2 * (\sigma_\delta^2 + \sigma_\epsilon^2) + 2\sigma_\epsilon^2)/N}}\right) \\ &\quad + \Phi\left(-1.96 - \frac{\lambda\beta}{\sqrt{(\lambda^2 * (\sigma_\delta^2 + \sigma_\epsilon^2) + 2\sigma_\epsilon^2)/N}}\right) \\ &= 1 - \Phi\left(1.96 - \frac{\beta * \sqrt{N}}{\sqrt{2\sigma_\delta^2 + 2\frac{\sigma_\epsilon^2}{\lambda^2}}}\right) + \Phi\left(-1.96 - \frac{\beta * \sqrt{N}}{\sqrt{2\sigma_\delta^2 + 2\frac{\sigma_\epsilon^2}{\lambda^2}}}\right) \quad (6) \end{aligned}$$

As  $\beta$  changes, the changes in  $\Phi\left(1.96 - \frac{\beta * \sqrt{N}}{\sqrt{2\sigma_\delta^2 + 2\frac{\sigma_\epsilon^2}{\lambda^2}}}\right)$  are always larger than changes in  $\Phi\left(-1.96 - \frac{\beta * \sqrt{N}}{\sqrt{2\sigma_\delta^2 + 2\frac{\sigma_\epsilon^2}{\lambda^2}}}\right)$ .

Therefore,  $\beta$  and  $\lambda$  have a positive effect on power, and  $\sigma_\delta^2$  and  $\sigma_\epsilon^2$  have a negative effect on power.

The standard error for this between-subjects study is:

$$SE = \frac{\sqrt{Var(Y_{x=0}) + Var(Y_{x=1})}}{\sqrt{N}} = \frac{\sqrt{\lambda^2 * (2\sigma_\delta^2) + 2\sigma_\epsilon^2}}{\sqrt{N}} \quad (7)$$

The true standardized mean effect size (true standardized Cohen's  $d$ ) equals to raw effect sizes divided by the (common) standard deviations of dependent measures,

<sup>6</sup>If the direction of the effect is hypothesized, a one-tailed test should be conducted. The conclusions regarding the relationship between reliability and power does not change for the one-tailed test.

which is the square root of the mean of the two variances when the population variances are different (Cohen, 1988).

$$\begin{aligned}
 d &= \frac{E[Y_{x=1}] - E[Y_{x=0}]}{\sqrt{\frac{\text{Var}(Y_{x=1}) + \text{Var}(Y_{x=0})}{2}}} \\
 &= \frac{\lambda * \beta}{\sqrt{\frac{\lambda^2 * (2\sigma_\delta^2) + 2\sigma_\epsilon^2}{2}}} \\
 &= \frac{\lambda * \beta}{\sqrt{\lambda^2 * (2\sigma_\delta^2)/2 + \sigma_\epsilon^2}} \\
 &= \frac{\beta}{\sqrt{\sigma_\delta^2 + \frac{\sigma_\epsilon^2}{\lambda^2}}} \quad (8)
 \end{aligned}$$

## References

- Albarracín, D., & Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology*, *69*, 299–327.
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . , Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365.
- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*, 1363–1368.
- Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology*, *22*, 49–55.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*, 145.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences 2nd edn*. Hillsdale, NJ: Erlbaum. Cambridge: Academic press.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155.
- Collins, L. M. (1996). Is reliability obsolete? A commentary on “Are simple gain scores obsolete?”. *Applied Psychological Measurement*, *20*, 289–292.
- Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*, 1482.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
- Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, England: Routledge.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.
- Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., & Zinkernagel, A. (2021). A multilab replication of the ego depletion effect. *Social Psychological and Personality Science*, *12*, 14–24.
- De Boeck, P., & Elosua, P. (2016). Reliability and validity: History, notions, methods, and discussion. In Leong, F. T. L., Bartram, D., Cheung, F. M., Geisinger, K. F., & Iliescu, D. (Eds.) *The ITC international handbook of testing and assessment*, (pp. 408–421). New York, NY: Oxford University Press.
- De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin*, *144*, 757.
- De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2016). Unreliable yet still replicable: A comment on LeBel and Paunonen (2011). *Frontiers in Psychology*, *6*, 2039.
- Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., & Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*, 158–171.
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378.
- Fleiss, J. L. (1976). Comment on Overall and Woodward’s asserted paradox concerning the measurement of change. *Psychological Bulletin*, *83*, 774–775.
- Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.
- Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3–12.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282.
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., & Zwienerberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573.
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186.
- Hopkins, K. D., & Hopkins, B. R. (1979). The effect of the reliability of the dependent variable on power. *The Journal of Special Education*, *13*, 463–466.
- Humphreys, L. G. (1993). Further comments on reliability and power of significance tests. *Applied Psychological Measurement*, *17*, 11–14.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.
- LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*, 570–583.
- Levin, J. (1986). Note on the relation between the power of a significance test and the reliability of the measuring instrument. *Multivariate Behavioral Research*, *21*, 255–261.
- McDonald, R. P. (1999). *Test theory: A unified approach*. Hillsdale, NJ: Erlbaum.
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108–141.
- Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299.
- Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, *23*, 87–89.
- Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*, 75–78.
- Nicewander, W. A., & Price, J. M. (1978). Dependent variable reliability and the power of significance tests. *Psychological Bulletin*, *85*, 405.

- Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. *Psychological Bulletin*, *94*, 524–533.
- Nosek, B. A., & Lakens, D. (2014). *Registered reports*. Göttingen, Germany: Hogrefe.
- Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, *82*, 85.
- Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, *83*, 776–777.
- Parsons, S. (2018). Visualising two approaches to explore reliability-power relationships. Center for Open Science. <https://doi.org/10.31234/osf.io/qh5mf>.
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*, 378–395.
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*, 528–530.
- Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638.
- Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *1904-1920*, *3*, 271–295.
- Sripada, C., Kessler, D., & Jonides, J. (2014). Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological Science*, *25*, 1227–1234.
- Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic?. *Perspectives on Psychological Science*, *9*, 305–318.
- Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768.
- Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika*, *23*, 9–17.
- Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . , Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*, 742–756.
- Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or elusive statistical power?. *Frontiers in Psychology*, *3*, 218.
- Vohs, K. D., Schmeichel, B., Fennis, B. M., Gineikiene, J., Hidding, J., Moieni-Jazani, M., . . . , Wagemakers, E. J. (2021). A multi site preregistered paradigmatic test of the ego depletion effect. *Psychological Science*. <https://doi.org/10.1177/0956797621989733>.
- Williams, R. H., Zimmerman, D. W., & Zumbo, B. D. (1995). Impact of measurement error on statistical power: Review of an old paradox. *The Journal of Experimental Education*, *63*, 363–370.
- Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451–1452.
- Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin*, *100*, 123.
- Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement*, *17*, 1–9.
- Zimmerman, D. W., & Zumbo, B. D. (2015). Resolving the issue of how reliability is related to statistical power: adhering to mathematical definitions. *Journal of Modern Applied Statistical Methods*, *14*, 5.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.