Educational Psychology Review (2025) 37:37
https://doi.org/10.1007/s10648-025-10015-5

COMMENTARY

Why We Might Still be Concerned About Low Cronbach's Alphas in Domain-specific Knowledge Tests

Steffen Zitzmann[1] · **Gabe A. Orona**[2,3]

Accepted: 26 March 2025 / Published online: 11 April 2025

Abstract

Edelsbrunner et al. *Educational Psychology Review, 37*, 1–43, 2025 recently published a systematic review and meta-analysis of Cronbach's alphas in domain-specific knowledge tests. While appreciating their analysis and agreeing with most findings, we disagree with three messages regarding the use of alpha in knowledge tests: (1) alpha measures the strength of interrelations among items, (2) a low alpha indicates validity, and (3) thresholds for alpha be abandoned. We discuss these messages in a constructive manner and present a way to counteract the inflation of seemingly high alphas in educational psychology.

Keywords Domain-specific knowledge test · Measurement · Validity · Reliability · Cronbach's alpha · Confidence interval · Bayesian credible interval

Domain-specific knowledge tests are measurement instruments. They were developed to gain information on individual capacities within specific knowledge areas. To assure high quality, scores from such measuring devices should be reliable (consistent over time, raters, and items), and authors should provide reliability evidence as stated in American Educational Research Association, American Psychological Association, and National Council on Measurement in Education's (2014) published Standards for Educational and Psychological Testing. Cronbach's alpha (Cronbach, 1951) is most prominent among different ways to provide such evidence. Recently, this journal published a systematic review that incorporates a meta-analysis of alphas in knowledge tests. Contrary to expectation, the authors of this study, Edelsbrunner et al. (2025), found a relatively high average alpha

✉ Steffen Zitzmann
steffen.zitzmann@medicalschool-hamburg.de

[1] Department of Psychology, Medical School Hamburg, 20457 Hamburg, Germany

[2] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, 72072 Tübingen, Germany

[3] Lynch School of Education and Human Development, Boston College, Chestnut Hill, MA 02467, USA

despite a low interrelatedness of items and mainly attributed these findings to the publication process of knowledge tests that they think had led to an overrepresentation of high alphas. They then moved on and discussed implications for knowledge tests and how a large threshold for alpha can bias research. We appreciate their excellent analysis and largely agree with the authors' findings. Edelsbrunner et al. (2025) convey, however, three messages regarding the use of alpha in knowledge tests that we consider problematic. First, the authors emphasize alpha as a measure of the strength of interrelations among items rather than a coefficient of reliability. Second and related to the first message, they interpret a low alpha as indicating a high level of content validity and thus as a strength rather than a limitation. Third, they suggest that educational psychologists abandon thresholds for alpha completely to move the field forward. However, we are hesitant to fully agree with them. In this commentary, we will present counterarguments against the authors'views and make a less revolutionary, simpler suggestion for improving current practice, without changing the interpretation of alpha, playing reliability off against validity, or dropping well justified thresholds.

Cronbach's Alpha Should not be Misinterpreted

Contrary to what has been the intention behind the development of alpha and contrary to standard practice, the authors refrain from interpreting alpha as a coefficient of reliability but view it merely as a measure of the strength of interrelations among items. This interpretation is unnecessarily imprecise if not flawed.

Alpha was first published by Cronbach (1951), who developed this coefficient against the background of classical test theory, which is a theory of errors and thus of reliability. As Sijtsma et al. (2024) put it, this theory"provides impressive contributions to psychometrics based on very few assumptions"(p. 84). Alpha can, however, also be justified by domain sampling or generalizability theory. With some limitations, alpha is even in line with a certain reading of factor analysis. And sometimes, it is reported in studies where an item response theory perspective was adopted, although this practice appears to be more ad hoc than theoretically backed up. Since the advent of alpha, there have been cautions about unrestricted use, particularly in factor analytic applications, accompanied by alternative solutions. Some scholars suggested to either use general alternatives such as what has been named greatest lower bound (Sijtsma, 2009), McDonald's omega (e.g., Green & Yang, 2009; McDonald, 1999; Raykov, 1997; Revelle & Zinbarg, 2009; Yang & Green, 2011), or more specific coefficients depending, among other features, on the specified measurement model (e.g., Bentler, 2009; McNeish, 2018; Zinbarg et al., 2005). However, these alternatives have been criticized as well for their misbehavior (e.g., Bell et al., 2024; Edwards et al., 2021; Sijtsma & Pfadt, 2021) or impracticability (e.g., Davenport et al., 2016; Revelle & Zinbarg, 2009; Viladrich et al., 2017). Also, there is an ongoing debate as to how Cronbach's alpha should be interpreted, and scholars have made different suggestions, such as internal consistency, unidimensionality, or homogeneity and other suggestions (e.g., Cortina, 1993; Crano et al., 2024; Green et al., 1977; McNeish, 2018; Revelle & Zinbarg, 2009; Schmitt, 1996; Sijtsma, 2009). We particularly disagree with Edelsbrunner et al.'s (2025) view that alpha would mainly indicate the interrelatedness of

items. This is not only problematic because alpha does not differentiate between the case in which each item is related to only a few other items from the case in which each item is related to nearly all other items (Revelle & Zinbarg, 2009). What is more, we find this perspective on alpha to be even dangerous when educational psychologists jump on this bandwagon and interpret a low alpha as only reflecting low item intercorrelations. Rather than this, alpha has a deeper meaning tied to reliability more broadly. Unlike Edelsbrunner et al. (2025), we will show and argue that alpha is, first and foremost, what it was meant to be at the time when Cronbach suggested this coefficient, namely the reliability of a scale score (see Cho, 2016; Raykov & Marcoulides, 2015, for similar views). Before we begin, we wish to note that as the basis for our discussion, we consider each of K items as one indicator of the domain under study, which we assume is a relatively unidimensional concept. Moreover, while these items differ in wording and their specific content, they are comparable in another important respect: when these items are answered by an individual person, their responses to the items differ mostly as a consequence of measurement error. This property defines the classical notion of parallel items.

To facilitate readability, assume without loss of generality that all items are standardized. Then, alpha can be expressed as:

$$\alpha = \frac{K \cdot \overline{\rho}}{1 + (K-1) \cdot \overline{\rho}} \tag{1}$$

where $\overline{\rho}$ is the average correlation among items. This is the usual formula for the standardized alpha. Assuming for a moment that the correlation is the same for all pairs of items, we can drop the bar from the $\overline{\rho}$ symbol, and because of the standardization of items, this correlation is equal to the variance of the true score σ_T^2 so that:

$$= \frac{K \cdot \sigma_T^2}{1 + (K-1) \cdot \sigma_T^2} = \frac{\sigma_T^2}{\frac{1-\sigma_T^2 + K \cdot \sigma_T^2}{K}} \tag{2}$$

Further, due to standardization, $1 - \sigma_T^2$ is equal to the variance of the measurement error σ_E^2. If we insert this error variance and simplify the equation, we obtain:

$$= \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2/K} \tag{3}$$

The denominator is the usual formula for the variance of the scale score. Using σ_X^2 for it, we arrive at:

$$= \sigma_T^2 / \sigma_X^2 \tag{4}$$

This ratio compares the amount of true score variance to the scale score variance, which is exactly what a reliability coefficient is expected to do (Cho, 2016). Indeed, the ratio equals the reliability of the scale score under classical test theory, domain sampling, and the generalizability theory perspective. This completes the proof that alpha is indeed a reliability coefficient and should be interpreted as such.

A similar, yet less stringent argument can be made for the general case, where items may be nonparallel, although this argument shows that when items are considerably nonparallel, alpha provides an underestimate of reliability (see Cho, 2016; Lord & Novick, 1968; Revelle & Zinbarg, 2009). However, knowledge tests are based on a sampling design that is called"blueprint"and involves the establishment of a valid selection of items (design-based approach; e.g., Camilli, 2018; Robitzsch & Lüdtke, 2022; White, 2025; Zitzmann, List, Lechner, Hecht, & Krammer, 2024). Very often, these items have equal weights, implying equal contributions. We therefore doubt whether typical tests are intended to include items that are highly nonparallel. Based on this, we speculate that underestimation of reliability will be at best marginal. Thus, one might ask oneself what one will lose if alpha continues to be the standard in knowledge tests, and conclusions from empirical examples and simulation studies are clear? Not much (e.g., Edwards et al., 2021; Gu et al., 2013; Hussey & Hughes, 2020; Savalei & Reise, 2019). Raykov and Marcoulides (2023; see also Raykov et al., 2024) developed a method for estimating the magnitude of underestimation. Employing this method, they demonstrated that even when items deviated considerably from being parallel, there was effectively no underestimation, which led them to conclude that the usual criticism of alpha (e.g., McNeish, 2018; Revelle & Zinbarg, 2009; Sijtsma, 2009) is exaggerated and misleading. Even.

those who had been critical about alpha have rolled back (e.g., Raykov et al., 2023; Sijtsma & Pfadt, 2021; Sijtsma et al., 2024). This and similar reasons may explain why alpha played a prominent role in the development and application of most knowledge tests and will arguably continue to do so in the future.

To further support our view that interpreting alpha as reflecting item intercorrelations is problematic, consider the above formula of alpha. It evident that the variance of the scale score in the denominator decomposes into true score variance and an error variance σ_E^2/K that depends on the number of items. If the amount of error variance depends on K, so does the reliability. This dependency is an important property of alpha and of reliability in general: it can be easily improved by expanding the test's length. Rather than an obstacle as suggested by Edelsbrunner et al. (2025), this is a great feature over mere interrelatedness of items. You cannot increase the interrelatedness by adding arbitrary items from the domain. Such an increase in interrelatedness is only possible when items are added that correlate highly with most of the other items or, in other words, by adding items that are redundant in the sense that they can be well predicted from the other items. Increasing reliability via the addition of items does not require that an additional item is well predictable by existing items. It only requires that this item is not much less reliable than the other items. From the angles of domain sampling and generalizability theory, this has two great advantages. Adding such items improves the generalizability of the scale score to the true score that had been obtained if all items from the pool had been administered. Moreover, adding items can greatly improve validity of the measurement in the sense that the resulting scale score will cover more content, a fact that Edelsbrunner et al. (2025) also mention.

Taken together, we argued that alpha is more in line with the idea of a reliability coefficient than with the view of Edelsbrunner et al. (2025) and others as reflecting merely the interrelatedness of items. The question arises as to what consequences

would follow if we failed to consider alpha as a reliability coefficient? In the following, we will make the case that such views may make researchers and practitioners believe that a low alpha is acceptable, a position that may worsen educational psychological research and practice rather than improve it.

Challenging Conventional Wisdom

While we generally agree with Edelsbrunner et al. (2025) that in knowledge tests, items may be only weakly correlated, we want to point out that by viewing a low alpha as a feature of knowledge tests rather than a limitation, the authors unintentionally provide an"argument"that other authors could easily exploit to play down or even legitimate a low alpha. As a consequence, these authors might get easily away with low reliability as long as, in their view, a high level of content validity is upheld. We do not dispute the outstanding significance of validity, and we thank Edelsbrunner et al. (2025) for bringing this aspect back to the reader's mind. Before moving on with reliability, we take the opportunity to engage more with their perspective on the interplay between reliability and content validity, which appears to express largely what we call the conventional wisdom. It traces back to Loevinger (1957) and has aptly been summarized by Steger et al. (2023). According to this wisdom, when selecting items that are highly correlated, a test will be highly reliable as indicated by a high alpha, but not very valid. In contrast, choosing items that are less correlated will result in a lower alpha but capture the concept much better. This position is also very much in line with a recent position by White (2025), who argued that test developers would face a peculiarity: whereas the usual guidelines for constructing valid knowledge tests emphasize the necessity to include items that capture different aspects of the concept in order to better represent the concept in its full breadth, the usual validation practice would consider these aspects as measurement error. Our critics could adopt this argument and defend Edelsbrunner et al. (2025) by arguing that even guidelines would imply low alphas. However, what seems like a logical implication is none. The reason is that the guidelines focus on item content, recommending that this content varies in such a way that the concept will be semantically represented by these items. Alpha, on the other hand, expresses the amount of true score variance in scale scores. Given a limited number of items, a low alpha implies a weak interrelatedness among these items. However, the link between content and correlation is less clear than educational psychologists often think. To see that variability in item.

content does not necessarily imply a low alpha, consider the following analogy, which involves the two measures head circumference and body height. We have learned that both are indicative of the physical development of toddlers. While these measures are obviously semantically distinct, they tap into the same underlying concept, and are highly rather than only loosely correlated. Our point here is that reliability and validity can both be high, calling the logic of what has been termed reliability-validity trade-off into question. Needless to say, this trade-off can occur empirically. By this, we mean that choosing items of diverse content may reduce alpha. Such findings may, however, not be due to content diversity per se, but may

result from violations of unidimensionality. Similarly, one cannot infer from a low alpha that content validity is high, as the low value may be due to other causes (e.g., unreliability of items), highlighting that validity evidence too can suffer from problems. Extending our example, consider vocabulary size as a third measure. While correlating only moderately with our two measures of physical development, possibly decreasing alpha when added to these two, vocabulary size can hardly be considered a valid indicator of physical but cognitive development. As vocabulary size would not increase validity, a low alpha may not always be an indication of validity, showcasing that establishing validity is a difficult task. Yet, we agree Edelsbrunner et al. (2025) that validity is important, doubting however whether it is a good move to play reliability off against validity. Developers should opt for a more balanced strategy. We do not discuss specific procedures here but refer instead to Clifton (2020) and the great potential of natural language processing in combination with AI-based optimization in this matter.

While acknowledging that a low alpha might reflect content validity in some cases, Edelsbrunner et al.'s (2025) perspective is incomplete or potentially risky if misapplied, particularly when this comes at the expense of reliability. A high reliability of the scale score as indicated by a high level of alpha or a comparable, in some applications more suitable measure, such as omega, is a prerequisite not only for doing research well but also for determining a student's state of knowledge in a precise way. In the next section, we will proceed as follows. We will begin with explanations why low alphas can bias results regarding relationships among.

variables in many different ways, followed by a reminder that only tests with high values for alpha can distinguish different states of knowledge. Our argument is based on the view that alpha is not only the most prominent coefficient of reliability but also often the only viable option. Although alternatives such as the test–retest coefficient exist, obtaining them with knowledge tests is very challenging due to the malleability of individual knowledge and to various methodological problems (e.g., retest effects). These challenges have been debated for a long time and need not to be reiterated.

Low Alphas can Have Adverse Effects

Readers might be familiar with Spearman's (1904) well-known correction for attenuation method (see also Spearman, 1910), which corrects a correlation coefficient for the attenuation caused by a low alpha of at least one of the variables. While most educational psychologists are aware of this bias, other consequences of unreliability are largely unknown, because these consequences are seldom discussed in their courses. We now move to list some of the more important consequences. Readers interested in the full picture are referred to Fuller's (1987) excellent, yet technical monograph entitled Measurement Error Models or, alternatively, to Buonaccorsi's (2010) Measurement Error: Models, Methods, and Applications.

Similar to the correlation coefficient, there will be a downward bias of the regression coefficient in a simple regression of the dependent variable Y on a predictor variable X when X exhibits a low alpha. Assume for a moment that both variables X

and Y are perfect measures, meaning they are equal to their true scores T_X and T_Y, respectively. Using ordinary least squares (OLS), we can express the regression coefficient β, which describes the relationship between X and Y, as the covariance between the variables'true scores $\sigma_{T_X T_Y}$ divided by the variance of the predictor's true score $\sigma^2_{T_X}$. But what coefficient can be expected when the predictor is error-prone? To address this question, we apply OLS first, yielding:

$$\beta^* = \frac{\sigma_{XY}}{\sigma^2_X} = \frac{\sigma_{T_X T_Y}}{\sigma^2_X} \tag{5}$$

Replacing $\sigma_{T_X T_Y}$ by $\sigma^2_{T_X} \cdot \beta$, which follows from solving the OLS form of β for $\sigma_{T_X T_Y}$, we get:

$$= \sigma^2_{T_X} / \sigma^2_X \cdot \beta = \alpha_X \cdot \beta \leq \beta \tag{6}$$

From this equation, it is evident that the coefficient will be attenuated by alpha, with the bias being large if alpha is low. To illustrate this bias, consider the following calculation, which is inspired by results from real data and can thus be considered empirically defensible. Edelsbrunner et al. (2025) provides insights regarding which values for alpha researchers can expect. With the lower bound of the prediction interval being 0.35, many of these values will indeed lie well below common thresholds. Take $\alpha_X = .60$ as an example. Assuming that the actual effect of prior knowledge on learning in a study is$\beta = 0.30$, which is positive and thus typical for such effects (Simonsmeier et al., 2022), we expect to yield$\beta^* = .60 \cdot 0.30 = 0.18$, which is only half this effect and thus demonstrates underestimation. It is important to note our argument does not involve the sample size, meaning that the shown bias is not a finite-sample bias (i.e., it does not specifically occur in small samples). Rather, the bias exists in large samples, where standard errors are negligibly small and results tend to be interpreted as facts, so that a low alpha may lead to wrong conclusions.

On the basis of the attenuation of correlation coefficients and regression coefficients in simple regressions, one might think that low alphas will always cause a downward bias, which is indeed a common misconception. In settings that involve more than two variables, almost any bias is possible. To give one example, consider the following scenario, which is often encountered in educational psychology: a multiple regression model in which the dependent variable Y is regressed on a predictor of focal interest X and a covariate Z. Assume that X, Y, and Z would be perfectly reliable for a moment and thus equal to their true scores T_X, T_Y, and T_Z, and denote the regression coefficients for X and Z as β and γ, respectively. We ask what coefficients can be expected when Z is error-prone.

To answer this question, we use a modified version of alpha, which we will refer to as the"adjusted alpha."This adjusted alpha can be calculated in a manner similar to alpha and is thus.

linked to alpha. It is a coefficient of reliability too, albeit of a conditional reliability, which we define in accordance with Mislevy (1991) as the reliability that results for a measure when the influence of another variable is controlled for. We do

not attempt to interpret the adjusted alpha in a substantive manner here but use it as a means to simplify expressions. According to the definition, the adjusted alpha of Z reads:

$$\alpha_Z^* = \frac{\sigma_{T_Z}^2 - \delta^2 \sigma_{T_X}^2}{\sigma_Z^2 - \delta^2 \sigma_{T_X}^2} \tag{7}$$

where δ is the coefficient in the regression of Z on X, which describes the relationship between X and Z. Due to squaring of this coefficient, we can deduce that:

$$\alpha_Z^* \leq \sigma_{T_Z}^2 / \sigma_Z^2 = \alpha_Z \tag{8}$$

suggesting that the adjusted alpha will usually be lower than alpha. It cannot be larger. This means in particular that the adjusted alpha will be even lower than an already low alpha.

The formulas for the requested regression coefficients can be easily obtained. However, because these deductions may be perceived as tedious and are not necessary to follow the argument, we skip them and refer interested readers to the Appendix. Equipped with the adjusted alpha of Z, we can state the coefficient for the error-prone covariate as:

$$\gamma^* = \alpha_Z^* \cdot \gamma \leq \gamma \tag{9}$$

Again, it is evident that the coefficient will be attenuated, with the attenuation being large if the adjusted alpha is low. This finding is not very surprising given the observed attenuation in simple regression, except perhaps that the magnitude of the attenuation is usually even larger. However, what is more interesting but counterintuitive at first glance is the impact that a low alpha of the covariate has on the coefficient for the predictor. This can be seen when this coefficient is written as a weighted sum of two coefficients:

$$\beta^* = \beta + \delta \cdot \left(1 - \alpha_Z^*\right) \cdot \gamma \tag{10}$$

As the resulting coefficient is not equal to the coefficient β but involves also γ, it will usually be biased. However, we cannot conclude by a simple statement such as that this coefficient will be.

attenuated, because things are more complex here. The sign and magnitude of the bias is determined by the specific combination of the relationship between the predictor and the covariate and, more importantly, the magnitude of attenuation in the coefficient for the covariate. As we have seen, this magnitude critically depends on the adjusted alpha of the covariate, with the magnitude becoming large if the adjusted alpha is low (all other conditions held constant). Because we know that the adjusted alpha will be lower than alpha, the statement holds true for alpha as well: the bias will be large if alpha is low. Taken together, this means for an existing effect of the predictor that a low alpha can lead to its attenuation. Also, it can inflate the effect or even artificially produce a spurious effect (also referred to as phantom effect; e.g., Pokropek, 2015; Televantou et al., 2015), calling into question

the practice of ignoring error in the covariates, which can be found in many published articles. Considering that we assumed that the predictor is perfectly reliable, we consider this to be an important finding and a critical case for when low alphas jeopardize research quality. To demonstrate the occurrence of a phantom effect, suppose our study follows a quasiexperimental design, where students are assigned to different intervention groups (standard vs. newly proposed intervention to promote learning) due to their prior knowledge. Now, assume that in a given study, there is actually no effect of the newly proposed intervention when controlling for prior knowledge, $\beta = 0.00$. This means that both interventions are equally effective. Further, assume that the relationship between prior knowledge and intervention is $\delta = 1.00$, which appears to be a reasonable assumption given the way students were assigned to groups, so that an adjusted reliability $\alpha_Z^* = .50$ may also be reasonable. Given the assumed effect of prior knowledge on learning from the previous example, we expect to yield $\beta^* = 0.00 + 1.00 \cdot (1 - .50) \cdot 0.30 = 0.15$ and thus a positive effect where there is actually none.

Whereas an attenuation raises the risk that the practical significance of an effect is underrated, inflation or a phantom effect can lead to an overstatement. Readers might ask which scenario is most problematic, which may remind us of the discussion about which statistical error is more severe, type I error or type II error. We think that this question can only be addressed from an applied perspective by asking whether it is more"expensive"for researchers and society to.

underestimate an effect or to find and report an effect that is spurious. Unfortunately, there is no general answer to this question other than that it depends on the context. For example, attenuation may lead to the false conclusion that a large-scale intervention (e.g., an education reform) will not be very effective, preventing the implementation of an effective protocol. On the other hand, spurious findings may lead to the implementation of an intervention that is ineffective and ties up resources that cannot be spent on other, potentially effective measures.

So far, we have considered regression scenarios in which a low alpha occurred for the predictor or the covariate. To change perspective, suppose the case of a simple regression model where the dependent variable Y, not the predictor X, shows a low alpha-a situation in which there will be no bias. Assuming once more that both variables would instead be equal to their true scores T_X and T_Y, we can define the coefficient β as $\sigma_{T_X T_Y}$ divided by $\sigma^2_{T_X}$. Addressing the question how the coefficient looks like when the dependent variable is error-prone requires the applications of OLS and the definition of β, resulting in:

$$\beta^* = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{T_X T_Y}}{\sigma^2_{T_X}} = \beta \tag{11}$$

From this identity, it becomes clear that the coefficient is unbiased and thus not affected by a low alpha, a statistical truth typically cited by educational psychologists in their responses to reviewers that criticized the low alphas of dependent variables. However, it is a mistake to think that the result applies equally well to the standardized coefficient that is often calculated and interpreted and that will be biased downward. Assume that $X = T_X$ and $Y = T_Y$ for a moment so that the

standardized coefficient B can be expressed as $\sqrt{\sigma^2_{T_X}/\sigma^2_{T_Y}} \cdot \beta$. To clarify what happens when Y is error-prone, we apply OLS along with our previous result and obtain:

$$B^* = \sqrt{\frac{\sigma^2_X}{\sigma^2_Y}} \cdot \beta^* = \sqrt{\frac{\sigma^2_{T_X}}{\sigma^2_Y}} \cdot \beta^* = \sqrt{\frac{\sigma^2_{T_X}}{\sigma^2_Y}} \cdot \beta \tag{12}$$

Multiplying by"1"yields:

$$= \sqrt{\frac{\sigma^2_{T_Y}}{\sigma^2_{T_Y}} \cdot \frac{\sigma^2_{T_X}}{\sigma^2_Y}} \cdot \beta = \sqrt{\frac{\sigma^2_{T_Y}}{\sigma^2_Y}} \cdot \sqrt{\frac{\sigma^2_{T_X}}{\sigma^2_{T_Y}}} \cdot \beta = \sqrt{\frac{\sigma^2_{T_Y}}{\sigma^2_Y}} \cdot B = \sqrt{\alpha_Y} \cdot B \leq B \tag{13}$$

showing that unlike the unstandardized coefficient, the standardized coefficient will be attenuated, particularly when alpha is low. Admittedly, this could have been found more easily by noticing that in simple regression, the standardized coefficient is equal to the correlation, which-as we know from Spearman-will be attenuated in the presence of unreliability.

Besides this biasing effect, a low alpha of the dependent variable has yet another less known consequence, which can be problematic as well: a drop in statistical power. This can be easily seen by looking at the standard error for the unstandardized coefficient, which is roughly:

$$se_\beta \approx \left(\frac{\left(1 - \rho^2_{XY}\right) \cdot \sigma^2_Y}{\sigma^2_X} / n \right)^{1/2} \tag{14}$$

where ρ_{XY} is correlation between X and Y. If we make use of the proportionality principle and perform some simplifying manipulations, we get:

$$se_\beta \propto \sqrt{\left(1 - \rho^2_{XY}\right) \cdot \sigma^2_Y} = \sqrt{\left(1 - \alpha_Y \cdot B^{*2}\right) \cdot \sigma^2_Y} = \sqrt{\left(1 - \alpha_Y \cdot B^{*2}\right) \cdot \sigma^2_{T_Y}/\sigma^2_{T_Y} \cdot \sigma^2_Y} \tag{15}$$

$$\propto \sqrt{\left(1 - \alpha_Y \cdot B^{*2}\right)/\alpha_Y} = \sqrt{1/\alpha_Y - B^{*2}} \tag{16}$$

Note that if α_Y is low, $1/\alpha_Y$ will be large, while B^{*2} will be small, together implying that the square root of the difference between these two terms will be large. Due to proportionality, the standard error will be large as well, with the typical consequence that statistically testing the coefficient will run a high risk to fail to detect an existing effect.

A critic might remark that our examples of adverse effects due to low alphas would merely be"formal games"of no relevance for educational psychologists, who routinely engage in modeling relationships among latent variables rather than among scale scores. Indeed, the use of latent variables inherently adjusts correlations and regression coefficients for unreliability when these variables capture what is common among their indicators. While it is true that latent variable modeling is capable of performing these adjustments, we would like to point out that to this day, not

all educational psychologists are sufficiently familiar with this methodology, even those who are do not apply it by default. There is still much research in educational psychology that does not involve latent variables, particularly from areas in which experimentally oriented.

approaches dominate. Moreover, latent variable modeling is not without problems, and these problems can affect results as well, suggesting that it may not always be the optimal choice (e.g., Ledgerwood & Shrout, 2011; Li & Beretvas, 2013; Lüdtke et al., 2011; Savalei, 2019; Ulitzsch et al., 2023; Zitzmann, 2018; Zitzmann et al., 2016). However, in light of these limitations, we believe that it is all the more important for applied researchers constructing and/or analyzing data from knowledge tests to secure high levels of alpha.

The final consequence of a low alpha that we wish to highlight before we move on should be intuitively clear to most educational psychologists, and this is why we only briefly mention it here. A large area of application of alpha is the assessment of individual students in order, for example, to detect learning delays for tailoring instruction to these students (Zitzmann et al., 2024a, 2024b, 2024c, 2024 d, 2024e, 2024f, 2024 g) or to report learning gains back with the potential effect to further motivate them (Zitzmann et al., 2024a, 2024b, 2024c, 2024 d, 2024e, 2024f, 2024 g). To ensure that measurement is trustworthy (i.e., not to a great extent determined by chance), it is essential that the used knowledge test is reliable. As argued, alpha is primarily a reliability coefficient. It can be used to express the amount of uncertainty that remains about a student's true level of knowledge or, in other word, to express to what extent one should remain skeptical about their scale score. Thus, a low alpha may cause a practitioner not to trust the test, or even worse, if they ignores measurement error, they may run a high risk of making the wrong decision, with potentially unfavorable consequences for the student-a point that cannot be repeated often enough (for recent reminders, see Zitzmann & Lindner, 2024; Zitzmann et al., 2023a, 2024).

To conclude so far, it should have become clear that deviating from the traditional interpretation of alpha, thereby inviting authors to legitimate a low alpha, can have numerous negative consequences. Finally, we will take a constructive stance and suggest that the usual reporting practice be complemented by confidence or Bayesian intervals for alpha.

Do not Drop Guidelines, Use Confidence/Bayesian Intervals

Edelsbrunner et al. (2025) argue that rather than establishing new thresholds for alpha for educational psychological research on the basis of the authors'findings, it would be more advantageous to drop thresholds altogether. We agree with the authors that a threshold can be misused in such a way that the file drawer problem will be increased or that content validity of the tests will be reduced. However, in our view, a well justified threshold can be a valuable means if it is understood as a rough guideline to achieve a level of reliability that is high enough to prevent most of the discussed adverse effects on results. The problem is not so much the existence of the threshold per se or a journals'tendency to accept only manuscripts with

high alphas. The real problem lies in the practice of comparing the value obtained from computing alpha from the data at hand (i.e., the estimated alpha) with the threshold, because this practice does not properly account for sampling error. This error exists because the estimated alpha is obtained from a sample of individuals rather than the usually much larger population of test takers. It can be thought of as the tendency to get a different value for alpha if the analysis would be repeated with another sample from the same population. We believe that one effective remedy is placing the confidence interval or, even better, a Bayesian variant thereof (e.g., Zitzmann et al., 2024a, 2024b, 2024c, 2024 d, 2024e, 2024f, 2024 g) around the estimate in order to correctly convey the uncertainty associated with this value (see Savalei & Reise, 2019, who also suggested this procedure). Intervals for alpha might not be readily available in software, but they can be obtained by employing factor analytic methodology (e.g., Raykov, 1997; Raykov & Marcoulides, 2015; Raykov & Shrout, 2002) or resampling techniques (e.g., Preacher & Hayes, 2008; Zitzmann, Nagengast, Hübner, & Hecht, 2024; Zitzmann et al., 2023a, 2023b). To showcase our suggested procedure, suppose that the estimated alpha is $\widehat{a} = .61$ and thus lower than the threshold 0.70 so that according to standard practice, reviewers and editors tend to conclude that alpha is too low to guarantee valid results regarding the relationships among variables, with the possible consequence that the manuscript will not get published. However, the calculated interval is [0.48,0.74] and includes the threshold,

suggesting a fair chance that the true alpha might be greater than the threshold and thus acceptably high. Note that in the light of the interval, the conclusion may be less pessimistic, and the risk of getting rejection is lowered. Only when the upper bound does not exceed the threshold can one reasonably infer that alpha is too low. This is just one example showing that when intervals are reported, this not only compliments the report on alpha, it might even change the game toward a better, more nuanced evaluation of research.

Summary

To sum up, along with the used domain-specific knowledge test, its reliability should be reported (e.g., American Educational Research Association et al., 2014), and Cronbach's alpha is the most prominent and viable option. Recently, Edelsbrunner et al. (2025) meta-analyzed the reported alphas in knowledge tests and found a high average alpha despite a low interrelatedness of items, which they attributed to the tendency of the publication process to lead to an inflation of high alphas. They discussed implications and how a large threshold for alpha can bias research. While we very much appreciate their study, we disagree with some of their messages regarding the use of alpha. Rather than following the authors and viewing alpha mainly as a measure of the strength of interrelations among items, we argued that alpha should continue to be viewed as it was meant to be, namely a reliability coefficient. Moreover, rather than interpreting a low alpha as indicating a high level of validity, a low alpha should be viewed as a threat to findings from studies involving such tests and to individual diagnostics. A high alpha (or a

comparable reliability coefficient) remains an important prerequisite for research and practice. Finally, rather than abandoning well justified thresholds for alpha, confidence or credible intervals for alpha should be reported and correctly interpreted. A manuscript with an estimated alpha that is too low, and that would otherwise get rejected, may be sent out for review when the interval suggests that the true alpha might be acceptable. This practice may counteract the inflation of seemingly high alphas in educational psychology, thereby improving research rather than worsening it.

In closing, we thank Edelsbrunner et al. (2025) for their reflections on the use of alpha in.

knowledge tests, their initiative to reevaluate alpha, and their inspiring ideas, which led us to think again about this coefficient. In line with the authors, we aim at improving educational psychology and do not see it as the effort of single individuals but a truly collaborative endeavor. As educational psychologists with a strong background in quantitative methods, we wanted to remind researchers and practitioners of the undeniable importance to get seriously engaged with reliability and alpha in particular. While researchers should remain open to novel and even unconventional ideas (e.g., Witte, 2024; Zitzmann & Loreth, 2021; Zitzmann et al., 2024a, 2024b, 2024c, 2024 d, 2024e, 2024f, 2024 g), they should not adopt them in the manner of a rash consumer but carefully evaluate them first. Otherwise, such ideas might fuel the field with myths, contributing to the unjustified use or omission of methods.

Appendix

How a Low Alpha of the Covariate Impacts the Coefficient for the Predictor.

Recall our finding that in simple regression, the coefficient that results for an error-prone predictor is equal to the unbiased coefficient times alpha. In the more general case of two independent variables as described in the main body of the article, a similar statement holds for both coefficients:

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \mathrm{A} \cdot \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{A1}$$

where A is a 2×2 matrix that can be seen as a natural extension of alpha to the case of two independent variables (e.g., Zitzmann, 2018). To obtain A for our specific case (i.e., only the covariate Z is error-prone), we insert the variances and covariances of the variables Z and X and simplify the expression:

$$A = \begin{pmatrix} \sigma_Z^2 & \sigma_{ZX} \\ & \sigma_X^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{T_Z}^2 & \sigma_{T_Z T_X} \\ & \sigma_{T_X}^2 \end{pmatrix} = \begin{pmatrix} \sigma_Z^2 & \sigma_{T_Z T_X} \\ & \sigma_{T_X}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{T_Z}^2 & \sigma_{T_Z T_X} \\ & \sigma_{T_X}^2 \end{pmatrix} \tag{A2}$$

$$= \frac{1}{\sigma_Z^2 \cdot \sigma_{T_X}^2 - \sigma_{T_Z T_X}^2} \begin{pmatrix} \sigma_{T_Z}^2 \cdot \sigma_{T_X}^2 - \sigma_{T_Z T_X}^2 & 0 \\ \left(\sigma_Z^2 - \sigma_{T_Z}^2\right) \cdot \sigma_{T_Z T_X} & \sigma_Z^2 \cdot \sigma_{T_X}^2 - \sigma_{T_Z T_X}^2 \end{pmatrix} \tag{A3}$$

$$= \begin{pmatrix} \frac{\sigma^2_{T_Z} \cdot \sigma^2_{T_X} - \sigma^2_{T_Z T_X}}{\sigma^2_Z \cdot \sigma^2_{T_X} - \sigma^2_{T_Z T_X}} & 0 \\ \frac{\left(\sigma^2_Z - \sigma^2_{T_Z}\right) \cdot \sigma_{T_Z T_X}}{\sigma^2_Z \cdot \sigma^2_{T_X} - \sigma^2_{T_Z T_X}} & 1 \end{pmatrix} \tag{A4}$$

Next, we consider the regression of Z on X with coefficient $\delta = \sigma_{T_Z T_X} / \sigma^2_{T_X}$. Solving for $\sigma_{T_Z T_X}$, replacing, and using the definition of the adjusted alpha α^*_Z (see main body of text) yields:

$$= \begin{pmatrix} \frac{\sigma^2_{T_Z} - \delta^2 \cdot \sigma^2_{T_X}}{\sigma^2_Z - \delta^2 \cdot \sigma^2_{T_X}} & 0 \\ \frac{\left(\sigma^2_Z - \sigma^2_{T_Z}\right) \cdot \delta}{\sigma^2_Z - \delta^2 \cdot \sigma^2_{T_X}} & 1 \end{pmatrix} = \begin{pmatrix} \alpha^*_Z & 0 \\ \delta \cdot \left(1 - \alpha^*_Z\right) & 1 \end{pmatrix} \tag{A5}$$

If we substitute this result for A to compute the coefficients, we finally arrive at:

$$\gamma^* = \alpha^*_Z \cdot \gamma \tag{A6}$$

$$\beta^* = \beta + \delta \cdot \left(1 - \alpha^*_Z\right) \cdot \gamma \tag{A7}$$

The resulting biases are discussed in the article.

Author Contribution S.Z.: writing, mathematical derivations. G.A.O.: writing.

Funding Open Access funding enabled and organized by Projekt DEAL. This research received no external funding.

Declarations

Ethical Approval This research did not involve human participants nor animals.

Conflict of Interest The authors declare no conflict of interest.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). The standards for educational and psychological testing. American Educational Research Association.

Bell, S. M., Chalmers, R. P., & Flora, D. B. (2024). The impact of measurement model misspecification on coefficient omega estimates of composite reliability. *Educational and Psychological Measurement, 84*, 5–39. https://doi.org/10.1177/00131644231155804

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137–143. https://doi.org/10.1007/s11336-008-9100-1

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. CRC Press.

Camilli, G. (2018). IRT scoring and test blueprint fidelity. *Applied Psychological Measurement, 42*, 393–400. https://doi.org/10.1177/0146621618754897

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods, 19*, 651–682. https://doi.org/10.1177/1094428116656239

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods, 25*, 259–270. https://doi.org/10.1037/met0000236

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Crano, W. D., Brewer, M. B., & Lac, A. (2024). *Principles and methods of social research* (4th ed.). Rouledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

Davenport, E. C., Davison, M. L., Liou, P.-Y., & Love, Q. U. (2016). Easier said than done: Rejoinder on Sijtsma and on Green and Yang. *Educational Measurement: Issues and Practice, 35*, 6–10. https://doi.org/10.1111/emip.12106

Edelsbrunner, P. A., Simonsmeier, B. A., & Schneider, M. (2025). The Cronbach's alpha of domain-specific knowledge tests before and after learning: A meta-analysis of published studies. *Educational Psychology Review, 37*, 1–43. https://doi.org/10.1007/s10648-024-09982-y

Edwards, A. A., Joyner, K. J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. *Educational and Psychological Measurement, 81*, 1089–1117. https://doi.org/10.1177/0013164421994184

Fuller, W. A. (1987). *Measurement error models*. Wiley.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827–838.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155–167. https://doi.org/10.1007/s11336-008-9099-3

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. *Methodology, 9*, 30–40. https://doi.org/10.1027/1614-2241/a000052

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*, 166–184. https://doi.org/10.1177/2515245919882903

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology, 101*, 1174–1188. https://doi.org/10.1037/a0024776

Li, X., & Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling, 20*, 241–264. https://doi.org/10.1080/10705511.2013.769391

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*, 444–467. https://doi.org/10.1037/a0024376

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412–433. https://doi.org/10.1037/met0000144

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196. https://doi.org/10.1007/BF02294457

Pokropek, A. (2015). Phantom effects in multilevel compositional analysis: Problems and solutions. *Sociological Methods and Research, 44*, 677–705. https://doi.org/10.1177/0049124114553801

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879–891. https://doi.org/10.3758/BRM.40.3.879

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184.

Raykov, T., Anthony, J. C., & Menold, N. (2023). On the importance of coefficient alpha for measurement research: Loading equality is not necessary for alpha's utility as a scale reliability index. *Educational and Psychological Measurement, 83*, 766–781. https://doi.org/10.1177/00131644221104972

Raykov, T., & Marcoulides, G. A. (2015). A direct latent variable modeling based method for point and interval estimation of coefficient alpha. *Educational and Psychological Measurement, 75*, 146–156. https://doi.org/10.1177/0013164414526039

Raykov, T., & Marcoulides, G. A. (2023). Evaluating the discrepancy between scale reliability and Cronbach's coefficient alpha using latent variable modeling. *Measurement, 21*, 29–37. https://doi.org/10.1080/15366367.2022.2031485

Raykov, T., Marcoulides, G. A., & Schumacker, R. (2024). Scale reliability evaluation using Bayesian analysis: A latent variable modeling procedure. *Measurement, 22*, 51–60. https://doi.org/10.1080/15366367.2023.2183799

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*, 195–212. https://doi.org/10.1207/S15328007SEM0902_3

Revelle, W., & Zinbarg, R. E. (2009). Coefficient alpha, beta, omega, and the GLB: Comment on Sitjsma. *Psychometrika, 74*, 145–154. https://doi.org/10.1007/S11336-008-9102-Z

Robitzsch, A., & Lüdtke, O. (2022). Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Measurement Instruments for the Social Sciences, 4*, 1–21. https://doi.org/10.1186/s42409-022-00039-w

Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). Collabra: Psychology, 5, 1–8. https://doi.org/10.1525/collabra.247

Savalei, V. (2019). A comparison of several approaches for controlling measurement error in small samples. *Psychological Methods, 24*, 352–370. https://doi.org/10.1037/met0000181

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika, 89*, 84–117. https://doi.org/10.1007/s11336-024-09964-7

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika, 86*, 843–860. https://doi.org/10.1007/s11336-021-09789-8

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist, 57*, 31–54. https://doi.org/10.1080/00461520.2021.1939700

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72–101. https://doi.org/10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2023). The road to hell is paved with good intentions: How common practices in scale construction hurt validity. *Assessment, 30*, 1811–1824. https://doi.org/10.1177/10731911221124846

Televantou, I., Marsh, H., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L.-E. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement, 26*, 75–101. https://doi.org/10.1080/09243453.2013.871302

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling - A comparison of constrained maximum likelihood, bayesian estimation, and fixed reliability approaches. *Psychological Methods, 28*, 527–557. https://doi.org/10.1037/met0000435

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Annals of Psychology, 33*, 755–782. https://doi.org/10.6018/analesps.33.3.268401

White, M. (2025). A peculiarity in psychological measurement practices. *Psychological Methods, Advance Online Publication.* https://doi.org/10.1037/met0000731

Witte, E. (2024). Comment on: Responsible research assessment I and responsible research assessment II. *Meta-Psychology, 8*, 1–3. https://doi.org/10.15626/MP.2023.3685

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*, 377–392. https://doi.org/10.1177/0734282911406668

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133. https://doi.org/10.1007/s11336-003-0974-7

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivariate Behavioral Research, 53*, 612–632. https://doi.org/10.1080/00273171.2018.1469086

Zitzmann, S., & Loreth, L. (2021). Regarding an "almost anything goes" attitude toward methods in psychology. *Frontiers in Psychology, 12*, 1–4. https://doi.org/10.3389/fpsyg.2021.612570

Zitzmann, S., & Lindner, C. (2024). How to assess response. European Archives of Psychiatry and Clinical Neuroscience, Advance online publication. https://doi.org/10.1007/s00406-024-01834-8

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling, 23*, 661–679. https://doi.org/10.1080/10705511.2016.1207179

Zitzmann, S., Lindner, C., Leucht, C., & Leucht, S. (2023a). A potential issue with PANSS responder analysis. *Schizophrenia Research, 261*, 287–290. https://doi.org/10.1016/j.schres.2023.10.009

Zitzmann, S., Weirich, S., & Hecht, M. (2023b). Accurate standard errors in multilevel modeling with heteroscedasticity: A computationally more efficient jackknife technique. *Psych, 5*, 757–769. https://doi.org/10.3390/psych5030049

Zitzmann, S., Bardach, L., Horstmann, K., Ziegler, M., & Hecht, M. (2024a). Quantifying individual personality change more accurately by regression-based change scores. *Structural Equation Modeling, 31*, 909–922. https://doi.org/10.1080/10705511.2023.2274800

Zitzmann, S., Lindner, C., & Hecht, M. (2024b). A straightforward and valid correction to Nathoo et al'.s Bayesian within-subject credible interval. *Journal of Mathematical Psychology, 122*, 1–6. https://doi.org/10.1016/j.jmp.2024.102873

Zitzmann, S., Lindner, C., Leucht, C., & Leucht, S. (2024c). Taking uncertainty in the assessment of response into account: An advanced guideline for computing responder rates in clinical trials. *European Neuropsychopharmacology, 85*, 3–4. https://doi.org/10.1016/j.euroneuro.2024.03.008

Zitzmann, S., List, M., Lechner, C., Hecht, M., & Krammer, G. (2024d). Reporting factor score estimates of teaching quality based on student ratings back to teachers: Recommendations from psychometrics. Manuscript submitted for publication.

Zitzmann, S., Nagengast, B., Hübner, N., & Hecht, M. (2024e). A simple solution to heteroscedasticity in multilevel nonlinear structural equation modeling. Manuscript submitted for publication.

Zitzmann, S., Orona, G. A., König, C., Lohmann, J. F., Bardach, L., & Hecht, M. (2024f). Novick meets Bayes: Improving the assessment of individual students in educational practice and research by capitalizing on assessors' prior beliefs. *Educational and Psychological Measurement, Advance Online Publication.* https://doi.org/10.1177/00131644241296139

Zitzmann, S., Wagner, W., Lavelle-Hill, R., Jung, A., Jach, H., Loreth, L., … & Hecht, M. (2024g). On the role of variation in measures, the worth of underpowered studies, and the need for tolerance among researchers: Some more reflections on Leising et al. from a methodological, statistical, and social-psychological perspective. *Personality Science*, *5*, 1-13. https://doi.org/10.1177/27000710241257413