A Fine-Tooth Comb for Measurement Reliability: Predicting Variance Components in Bayesian Hierarchical Models

Donald R. Williams, Stephen R. Martin,
Michaela C. DeBolt, Lisa M. Oakes, and Philippe Rast
University of California, Davis

Author Note

Research reported in this publication was supported by funding from the National Science Foundation Graduate Research Fellowship to DRW and the National Institute On Aging of the National Institutes of Health under Award Number R01AG050720 to PR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Abstract

The primary objective of this work is to extend classical test theory (CTT), and in particular, for the case of repeated measurement studies. The guiding idea that motivates this work is that any theory ought to be expanded when it is not compatible with commonly observed phenomena–namely, that homogeneous variance appears to be the exception and not the rule in psychological applications. Additionally, advancements in methodology should also be considered in light of theory expansion, when appropriate. We argue that both of these goals can be accomplished by merging heterogeneous variance modeling with the central tenants of CTT. We introduce novel methodology that is based on the mixed-effects location scale model. This allows for fitting explanatory models to the "true" score (between-group) and error (within-group) variance. Two illustrative examples, spanning from educational research to infant cognition, highlight such possibilities. The results revealed that there can be substantial variability in error variance, which necessarily implies the same for reliability, and that "true" score variance can be a function of covariates. We incorporate this variance heterogeneity into novel reliability indices. These extend traditional formulations that assume the variance components are fixed and non-varying. This powerful approach can be used to identify predictors of "true" score or error variance, which can then be used to refine measurement. The methods are implemented in the user-friendly R package `ICCier`.

Keywords: measurement reliability, Bayesian, classical test theory, heterogeneous variance, hierarchical model

A Fine-Tooth Comb for Measurement Reliability: Predicting Variance Components in Bayesian Hierarchical Models

> It seems likely that in mental testing we shall find ourselves in the future increasingly concerned with measurement of fluctuating mental functions.
>
> —(p. 340, Thouless, 1936)

A fundamental concept underlying psychological science is measurement reliability. From educational researchers to clinical and developmental psychologists to psychophysicists, we all understand that gathering reliable measurements plays an important role in psychological inquiry in particular. That is, while reliability is important for all sciences, we often face the difficult task of investigating latent constructs that cannot be directly measured. For example, using a set of items to measure positive affect, rating mother-child interactions for the purpose of attachment classification, or investigating cognitive function experimentally, all share that common thread. This challenge has resulted in a rich literature relating to psychological assessment, wherein the primary concern is measurement. Here the foundation for thought is organized around classical test theory (CTT, Thorndike, 1904; Traub, 1997).

In CTT, observed measurements are composed of a given individual's "true" score plus *random* error, $X = T + E$ (Lord & Novick, 1968; Novick, 1965). Thus, the variability of observed measurements is defined as the sum of true-score variance, σ^2_T, plus random error variance, σ^2_E . Reliability is by definition the proportion of the total observed variance, $\sigma^2_T + \sigma^2_E$, that is true score variance. This work explicitly focuses on the latter source of variance. A critical assumption underlying the gathered measurements is that the errors are random and uncorrelated. That is, the residuals do not exhibit systematicity such that they are constant over, say, time of measurement, experimental condition, and across individuals. Indeed, to date, "systematic errors are not handled well in CTT" (p.

94, Kline, 2005). In fact, we would argue that they are not handled at all. However, several lines of recent research point towards randomness in the error structure being the *exception* rather than the rule.

These insights into the error structure are not restricted to specific psychological applications. They appear to apply generally. For example, Williams, Zimprich, and Rast (2019) examined the residual variance in a learning task, where it was found that the errors across repeated trials followed a nonlinear trajectory, including individual variation therein. This conflicts with the underpinnings of CTT, in that fluctuations cannot be a function of learning (Kline, 2005). Additionally, in Williams, Liu, Martin, and Rast (2019), it was found that the errors can be explained by an individuals emotional state the previous day, including both positive and negative affect, as well as physical activity. This again hints at systematicity in the error structure. These findings are also not restricted to observational studies, and have recently emerged from experimental settings. In particular, Williams, Rouder, and Rast (2019) and Williams, Martin, and Rast (2019) demonstrated that there was substantial individual variation in the error structure of cognitive inhibition tasks. This is especially important, because, in commonly used reliability indices, each person is assumed to have a common error variance.

We recognize that the variance structure is commonly investigated in the social-behavioral sciences. For example, in the case of significance testing, homogeneity of variance is an important assumption for ensuring nominal error rates (Blanca, Alarcón, Arnau, Bono, & Bendayan, 2018; Delacre, Lakens, & Leys, 2017). Further, in Ruscio and Roche (2012), it was shown that variance heterogeneity is common in psychological research and it is a function of the number of groups. That is, the ratio between the smallest to largest variance increases with more groups (see Table 2, Ruscio & Roche, 2012). The present work is not focused on significance testing. Rather, by incorporating variance heterogeneity into the descriptive nature of reliability indices, this can provide richer information about measurement.

There is also an interesting and storied literature on modeling within-person variance (i.e, the error, Cleveland, Denby, & Liu, 2003; Hedeker, Mermelstein, & Demirtas, 2008, 2012; Leckie, French, Charlton, & Browne, 2014; Rast & Ferrer, 2018; Rast, Hofer, & Sparks, 2012). The aforementioned findings do stand apart from the existing literature in that there was a heavy emphasis on data visualization. By plotting individual variation in the error structure, the degree of systematicity was more than readily apparent–it was striking.[1] For instance, it was recently shown that the vast majority of individuals differed from the average within-person variance in hierarchical models (see Figure 3 in: Williams, Martin, & Rast, 2019). It was further noted that "...heterogeneous within-person variance is a defining feature of these [cognitive] tasks..." (p.1, Williams, Martin, & Rast, 2019). This is notable, as we describe below, because the average variability is used to compute reliability. Together, these findings suggest that a large component of error may be systematic and largely under-explored in psychological research.

Alternative frameworks have been proposed to overcome limitations of CTT. In particular, generalizability theory (*G*-theory) emerged to explicitly broaden how we think about and evaluate measurement reliability (Brennan, 1992; Cronbach, 1972; Shavelson & Webb, 2010) . This is also a primary goal of this work–that is, "In classical test theory measurement error is undifferentiated random variation; the theory does not distinguish among various possible sources" Shavelson and Webb (p. 599 2012). For example, in *G*-theory, the variance is partitioned into several components. A classic example considers variance at the level of the person, item, and time of measurement, including interactions therein. This flexible framework allows for investigating, say, reliability (e.g., intraclass correlation coefficients) for each subject across time points. However, *G*-theory still assumes that the error variance is "constant for all persons, regardless of true score" (p. 321 Shavelson & Webb, 2012) This assumption in particular has been the target of much

[1] This stands in contrast to exclusively using some form of hypothesis testing to communicate findings. For example, when reporting only a significant *p*-value or large Bayes factor, the degree of systematicity is not readily apparent.

attention that dates back to Lord (1955). As but one example of violating this assumption, there is a well-known relationship between reaction time means and standard deviations. Thus the error variance is partially defined by the location of the "true" score. Importantly, this is an assumption of commonly used statistical methodology (e.g., ANOVA) used to compute reliability. But in general this assumption can be relaxed.

Consider the case of a one-way random effects model (Bartko, 1966), which is the focus of this work. There are two sources of variation, that is,

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}. \tag{1}$$

This is commonly referred to as ICC(1), and it serves as a reliability index for single scores that ranges from $0 - 1$ (Shieh, 2016). In (1), σ_b^2 is the between-group variance or the variance in "true" scores. Thus reliability is the proportion of variance attributed to σ_b^2. Further, σ_w^2 is the within-group variance and it is often referred to as measurement error. In cognitive inhibition tasks, for example, it captures trial-to-trial "noise" in reaction times. Thus, assuming that σ_b^2 is held constant, increasing σ_w^2 will necessarily decrease reliability (Hedge, Powell, & Sumner, 2018). This definition of ICC does not allow for the possibility of individual differences in reliability. This was the primary focus of Williams, Martin, and Rast (2019), where it was demonstrated that there are individual differences in σ_w^2 and thus also in reliability. This key insight is not possible within both CTT or *G*-theory, and it forms the impetus for this current work.

We introduce novel methodology to probe measurement reliability at both the level of the numerator and denominator in (1). This extends Williams, Martin, and Rast (2019), where only a random intercepts model was fitted to σ_w^2, such that each person had their own error variance and person-specific reliability. In this work, we now predict the "true " score (σ_b^2) and error (σ_w^2) variance with sub-models. The latter can also capture individual differences in error variance. This is a powerful approach for characterizing reliability as it allows for investigating whether σ_b^2 can be explained by, say, age groups. That is, perhaps older individuals are *more* homogeneous, which implies they are *less* reliable, assuming σ_w^2 is held constant. This general idea extends beyond people. For example, suppose that σ_w^2 is permitted to vary among schools. Here each school would then have their own error variance that can also be predicted, by, say, gender. This would result in school-specific reliability that is a function of gender. These kinds of insights are possible by seeking to *explain* the variance components in (1).

This work is organized as follows. In the first section we introduce the model. Our intention here is to describe key aspects of the proposed methodology. This serves as the foundation for the remainder of the paper. The rest of the work consists of case studies that span from educational to cognitive psychology. This demonstrates the utility of the methodology. In this section, we also emphasize the connection between the presented methodology and classical test theory in hierarchical models. We end by summarizing our major and novel contributions, as well as discussing implications for measurement in psychological research.

Model Formulation

The presented methodology is based upon a straightforward extension to the traditional mixed-effects approach, which allows for partitioning the unexplained variance, or within-group variance, as well as the between-group variance, or "true" score variance. The technique to do so is termed mixed-effects *location scale* model (MELSM, pronounced mel·zəm, Hedeker et al., 2008, 2012), which combines earlier work on variance heterogeneity Aitkin (1987) and models for random scale effects Cleveland et al. (2003). In this work, we build upon this foundation and further demonstrate that the MELSM has untapped potential as a fine-tooth comb for assessing measurement reliability.

Mixed-Effects Location Scale Model

The starting point is the standard linear mixed effects model for $i = 1, 2, \dots, G$ groups (e.g., people or schools) and j ($j = 1, 2, ..., n_i$) measurements that may be specified as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where $\mathbf{y}_i$ is the $n_i \times 1$ response vector for observations in group i. $\mathbf{X}_i$ is the $n_i \times k$ design matrix for the fixed effects for observations in group i. Note that an intercept only results the unconditional model (Raudenbush & Bryk, 2002), whereas including explanatory variables in $\mathbf{X}_i$ allows for computing conditional ICCs (Rabe-Hesketh & Skrondal, 2008). $\boldsymbol{\beta}$ captures the fixed effects and its dimension is $k \times 1$. The random effects are in the $n_i \times q$ matrix $\mathbf{Z}_i$ for observations in group i where $\mathbf{b}_i$ is the according $q \times 1$ vector with the random effects coefficients. These effects characterize the group means for the response (i.e., the location). $\boldsymbol{\epsilon}_i$ is a vector of errors specific to group i. The general assumption in standard mixed effects models is that random effects are $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Phi})$. Where $\boldsymbol{\Phi}$ is a $q \times q$ covariance matrix for the random effects with the variances σ_b^2 and the covariances $\sigma_{bb'}$ (for $q \neq q'$). The errors $\boldsymbol{\epsilon}_i$ are also assumed to be normally distributed with a mean of $\mathbf{0}$ and covariance of $\sigma_\epsilon^2\boldsymbol{\Psi}_i$ where $\boldsymbol{\Psi}_i$ is a $n_i \times n_i$ matrix which can take different structures. In these models the between-group variance is captured by σ_b^2 and the within-group variance is represented in σ_w^2. In the context of reliability, the former is termed "true" score variance. This formulation leads to computing traditional measures of reliability, for example, ICC(1) given in (1)–i.e., $\sigma_b^2/\sigma_b^2+\sigma_w^2$

Error Variance. In this standard form, the error variance σ_ϵ^2 is a fixed entity. In order to allow it to differ at the group level, we add the subscript i to the within-group variance term (cf. Hoffman, 2007; Myles, Price, Hunter, Day, & Duffy, 2003) and we also allow it to differ among j-measurements to obtain $\sigma_{\epsilon_{ij}}^2$. Changes in the within-group variance $\sigma_{\epsilon_{ij}}^2$ are explained by group varying covariates in the $n_i \times m$ matrix $\mathbf{W}_i$ for the fixed effects and $\mathbf{V}_i$, with dimension $n_i \times p$ (and $m \geq p$) for the random effects (Rast et al., 2012). Hence, with the inclusion of level one covariates the within-group variance not only varies across groups but also across measurements given the model:

$$\boldsymbol{\varphi}_i = \exp(\mathbf{W}_i\boldsymbol{\eta} + \mathbf{V}_i\mathbf{t}_i). \tag{3}$$

$\boldsymbol{\varphi}_i$ then is the $n_i \times 1$ vector that contains all error variances $\sigma^2_{\epsilon_{ij}}$ for group i and for each measurement j within that group. $\boldsymbol{\eta}$ is comparable to the regression weights $\boldsymbol{\beta}$ in (2). That is, for an intercept and slope term, η_0 defines the average within-group variance and η_1 weights the influence of the predictor on the variance. The individual departures from the fixed effects that are captured in the random effects $\mathbf{t}_i$ are normally distributed with $\boldsymbol{t}_i \sim N(\mathbf{0}, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is a covariance matrix of dimension $p \times p$ that contains the random effects of the scale. Note that $\mathbf{W}_i$ and $\mathbf{V}_i$ may, or may not, be the same as $\mathbf{X}_i$ and $\mathbf{Z}_i$. The exponent is used to ensure that the variance is restricted to positive values, and thus, is lognormally distributed (Hedeker et al., 2008).

"True" Score Variance. We also introduce a sub-model for the between-group variance. In other words, we can *predict* the "true" score variance in (1). It is important to note here that we now have random effects $\mathbf{b}_i$ from the *location* of the model (the means structure) and random effects $\mathbf{t}_i$ from the *scale* of the model (the within-group variance structure). All these random effects are assumed to come from a normal distribution with mean zero. Hence, we can stack both $\mathbf{b}_i$ and $\mathbf{t}_i$ vectors, resulting in $\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$. This also means that $\boldsymbol{\Sigma}_i$ contains the variances and covariances of both, the location and scale. In order to define a variance model for $\boldsymbol{\Sigma}_i$, we can decompose $\boldsymbol{\Sigma}_i = \boldsymbol{\tau}_i\boldsymbol{\Omega}\boldsymbol{\tau}_i'$, where $\boldsymbol{\tau}_i$ is a diagonal matrix for group i in which the diagonal elements are the random-effect standard deviations and $\boldsymbol{\Omega}$ is the correlation matrix that contains the correlations among *all* random effects. That is, $\boldsymbol{\Omega}$ is of dimension $(q+p) \times (q+p)$ and contains the correlations among the random effects of the location, the correlations among the random effects of the scale, and the correlations among the random effects of the location and the scale. Given this definition, $\mathbf{\Omega}$ remains constant across conditions. We can now define a model for the random effects SDs, which can be defined as

$$\mathrm{diag}(\boldsymbol{\tau}_i)' = \exp(\boldsymbol{g}_i\boldsymbol{\iota}), \tag{4}$$

where $\boldsymbol{g}_i$ is the design matrix that contains between-group predictors (e.g., age group or gender), and $\boldsymbol{\iota}$ is a matrix with $(q+p)$ columns of coefficients. ι_{rc} is the effect of the rth column in $\boldsymbol{g}_i$ on the cth random effect SD (i.e., $\tau_{c,c}$). For example, ι_{12} is the intercept for $\tau_{2,2}$, and ι_{23} is a slope parameter for $\tau_{3,3}$. Consequently, the random-effects variance is not constant but may change due to group-specific characteristics (e.g. Leckie et al., 2014). For example, if there is a positive effect of age, this would mean that older people are relatively more heterogeneous in their "true" scores and thus also more reliable.

Hence, having specified all elements, we can define the full MELSM as

$$\begin{aligned}
y_i &\sim N(\mu_i, \boldsymbol{\varphi}_i)\\
\mu_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i\\
\boldsymbol{\varphi}_i &= \exp(\mathbf{W}_i\boldsymbol{\eta} + \mathbf{V}_i\mathbf{t}_i)
\end{aligned}$$

with the random effects for both the location and the scale coming from the same multivariate distribution

$$\begin{aligned}
\begin{bmatrix}\mathbf{b}_i\\ \mathbf{t}_i\end{bmatrix} &\sim N(\mathbf{0}, \mathbf{\Sigma}_i)\\
\mathbf{\Sigma}_i &= \boldsymbol{\tau}_i\mathbf{\Omega}\boldsymbol{\tau}_i'\\
\mathrm{diag}(\boldsymbol{\tau}_i) &= \exp(\boldsymbol{g}_i\boldsymbol{\iota}).
\end{aligned}$$

Group-Specific Reliability

Modeling the variance structure leads to cluster-specific reliability. First consider the case of a random intercept models fitted to both the location and scale, which estimates a variance for each group. Computing reliability can be accomplished with a straightforward extension to ICC(1) given in (1)–i.e.,

$$\rho_i = \frac{\tau_{1,1^2}}{\tau_{1,1^2} + \exp[\eta_0 + t_i]^2} \tag{5}$$

Note that the subscript i denotes the ith group, $\tau^2_{1,1}$ is the "true" score variance (i.e., σ^2_b), and $\exp[\eta_0 + t_i]^2$ is the group-specific variance. More specifically, with $i = 1$, this formulation would provide the group-specific estimate of reliability for the first group. Further, in (5), the covariance between two observations from the same group remains unchanged from the customary definition of ICC(1). In other words, the only modification is that the correlation is now expressed as a function of the within-group variances.

Predicting Reliability

The above formulation is restricted to estimating group-specific reliability, for, say, a person or school. The central idea behind this work is to also fit explanatory models to both the "true" score (σ^2_b) and error (σ^2_w). This allows for predicting reliability and it extends customary theories of measurement, where the variance components are taken to be fixed and non-varying constants. To this end, we follow Hedeker et al. (2008) and define the covariance for any two measurement from the same group as

$$\begin{aligned} \mathrm{Cov}(y_{ij}, y_{ij'}) &= \sigma^2_{bi} & (6) \\ &= \tau^2_{i1,1} = \exp(\boldsymbol{g_i \iota}_1)^2 \text{ for } j \neq j'. & (7) \end{aligned}$$

Note that subscript i is used to emphasize that the variance is a function of the design matrix $\boldsymbol{g}_i$ and the regression weight vector $\boldsymbol{\iota}_1$. For each group, the predicted reliability then takes on the following form

$$\boldsymbol{\rho_i} = \frac{\exp(\boldsymbol{g}_i\boldsymbol{\iota}_1)^2}{\exp(\boldsymbol{g}_i\boldsymbol{\iota}_1)^2 + \exp\left(\mathbf{W}_i\boldsymbol{\eta} + \mathbf{V}_i\mathbf{t}_i\right)^2} \tag{8}$$

Hence, reliability can be probed at both the level of the numerator of denominator. Of course, if there is not much individual variability in the variance structure, and if the covariates have a minimal effect on the variances, this would result in (1) and (8) producing similar estimates. This is because a mixed-effects model is a special case of the MELSM, but with an implicit fixed intercept only model fitted to the variance σ^2_w. In the case studies, we provide specific models that will further clarify this formulation.

A Note of Acknowledgement

Before proceeding to the illustrative examples, we want to emphasize our acknowledgement of the relevant literature. While varying ICCs naturally arise from heterogeneous variance modeling, these ideas are most prominent in research areas that gather intensive longitudinal data (Hamaker, Asparouhov, Brose, Schmiedek, & Muthén, 2018; Hedeker et al., 2012; Rast & Ferrer, 2018; Watts, Walters, Hoffman, & Templin, 2016; Williams, Liu, et al., 2019). Indeed, to our knowledge, the notion was first described in the context of ecological momentary assessment. In particular, Hedeker et al. (2008) described how the variances (e.g., σ^2_b and σ^2_w) could be a function of covariates. This provided the foundation for Brunton-Smith, Sturgis, and Leckie (2017) and Williams, Martin, and Rast (2019). These works in particular estimated group-specific ICCs in hierarchical models.

There are several novel aspects of the present work. First, we *fully* merge the ideas stemming from the variance modeling literature with psychological measurement. In particular, we provide the key insight that the MELSM is ideal for assessing and *predicting* reliability. The implications of this are far reaching–e.g., the long-standing issue of assuming a constant within-group variance is addressed. Second, we adopt a Bayesian framework for extending CTT. The advantages of Bayesian methods for estimating "true" score and error variance were described in Lindley (1969) and Novick, Jackson, and Thayer (1971). Our work builds upon those ideas. But with the full power of modern Markov chain Monte Carlo algorithms (Betancourt, 2017), which were not computationally feasible until the 1990's (e.g., see, Gelfand & Smith, 1990; Robert & Casella, 2011). Third, we demonstrate the utility of this framework in several psychological applications. This highlights the generality of our methodology. We have also implemented it in the user-friendly R package `ICCier`, which serves as a high-level interface to the programming language Stan (Stan Development Team, 2016).

Illustrative Examples

In this section, we employ the methodology in a wide range of psychological applications. Our intention here is to demonstrate the utility of heterogeneous modeling for assessing reliability. Further, recall that the motivation for this work was partly based on recent findings demonstrating that homogeneity in the variance structure appeared to be rare in psychological applications. Thus, in these examples, we start by characterizing the within-group variance and then proceed to *predicting* reliability. This necessarily requires formulating a variety of models. However, it is important to note that the central ideas from each case study apply generally. This includes the relation between measurement reliability and hierarchically modeling.

Case 1: Educational Research

Educational researchers often encounter hierarchically structured data. That is, while students serve as the unit of measurement, they are commonly nested within schools that can in turn be nested within a higher-level unit, say, the respective school district, county, or even at the state level. It is for this reason educational data is often used to demonstrate key aspects of hierarchically modeling. To our knowledge, however, the MELSM has only been used once to investigate within-school variance heterogeneity (Leckie et al., 2014).

An important question in education is to identify factors that are related to student success and academic achievement. Such factors include school location (Logan & Burdick-Will, 2017), teacher expectations (de Boer, Timmermans, & van der Werf, 2018), and the socioeconomic status composition among attending students (Sirin, 2005). However, reducing education inequality between, for example, suburban and urban schools (Sandy & Duncan, 2010), has proven to be far from trivial in that " ...income-related gaps both in access to and in success in higher education are large and growing" (p. 125, Haveman & Smeeding, 2006). Consequently, an important question is to not only study differences between developed environments, but to also investigate whether certain schools differ from each other in test scores. This can provide important information at the school-level (e.g., rankings) and it is the focus of the following example.

Average Score Reliability. Assessing school-level differences in average test scores is inextricably linked to average score reliability. As a point of reference, in relation to people, this is analogous to the study of individual differences in the location or mean structure. This is based upon CTT, in that, with infinite measurements, the idea is that we will converge on the "true" score. In hierarchical models, this relates to the precision with which the cluster (e.g., schools or individuals) specific averages or the mean test scores are measured. What is less appreciated is the link to shrinkage, which is a defining feature of hierarchical modeling. In the presence of error, group estimates are smoothed towards the grand mean (the groups become more similar to one another). The degree of smoothing for each group i is defined as

$$\lambda_i = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2/n_i}, \tag{9}$$

which is known as the so-called "shrinkage factor." And it is *equivalent* to average score reliability or ICC(2) (McGraw & Wong, 1996; Shieh, 2016). In (9), n_i is the number of measurements for cluster i. Note that this is the same as ICC(1) given in (1), but in this case, σ^2_w is divided by the respective number of measurements gathered from each cluster i. This establishes the foundational link to CTT, in that reliability approaches one as $n_i \rightarrow \infty$, with the assumption that σ^2_b is greater than zero. The predicted estimate $\hat{\mu}_i$ for school i is then computed with

$$\hat{\mu}_i = (1 - \lambda_i)\bar{\text{Y}}_G + \lambda_i \bar{\text{Y}}_i, \tag{10}$$

where $\bar{\text{Y}}_G$ is the grand mean and $\bar{\text{Y}}_i$ is the school-specific mean (McCulloch, 2003). This is commonly referred to as the Best Linear Unbiased Predictor (BLUP) and it is related to the so-called "Stein's paradox" (Efron & Morris, 1977; Morris & Lysy, 2012; Stein, 1956). However, it is important to note that the original derivation is grounded in psychometrics and CTT. That is, (10) was given as an individuals best estimate of their "true" score in Kelley (p. 178, see eq. 22, 1927).

Herein lies how our methodology can be viewed as an extension to CTT. Namely, in (9), the calculation of average score reliability assumes a common within-group variance σ^2_w. This is not so for the MELSM, and in reference to group-specific reliability given in (5), we can further establish a connection to the Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910), that is,

$$\Psi(\rho)_i = \frac{J \cdot \rho_i}{1 + (J - 1)\rho_i}. \tag{11}$$

Here the Spearman-Brown equation is applied to the cluster-specific reliability, ρ_i, based on ICC(1), or $\text{ICC}(2)_i = \Psi\{\text{ICC}(1)_i\}$, such that each cluster (e.g., school) has their own predicted reliability for J measurements. Note that (11) approximates ICC(2), that was given in (9), asymptotically as J increases (Bliese, 2000). Additionally, because the "shrinkage factor" is equivalent to ICC(2), the Spearman-Brown equation in (11) can also be understood as predicting the expected degree of shrinkage in a one-way random effects model. The generalization to predicting group-specific reliability and the connection to shrinkage are major contributions of this work.

Model Specification. We now apply the MELSM and the Spearman-Brown equation for predicting reliability based on J measurements in an educational setting. The basic idea is to predict reliability, in an effort to reach an acceptable level while also accounting for (possible) heterogeneity in the within-school variance structure. Further, assuming the goal is to detect school-level differences, (11) can also be used to predict the expected shrinkage. This is critical to detect differences, in that, with increasing shrinkage, the school-level means or "true" scores become more homogeneous (Gelman, Hill, & Yajima, 2012).

We use data from the General Certificate of Secondary Education (GCSE) exam, which is an academic qualification for the United Kingdom. There are 65 schools and 4,059 students from six Inner London Education Authorities. These data have been used in several examples demonstrating the utility of hierarchical models. In particular, the school-level residuals have also been examined in an attempt to make inference about the variance structure (Goldstein et al., 1993, see Figure 2,), which is a naive MELSM. In the present analysis, the outcome is normalized test scores at age 16. We also predict "true" score variance, σ_b^2, with school-level averages for the London Reading Test (LRT) at age 11. The schools have been categorized into three groups (based on the averages), consisting of the bottom 25%, middle 50%, and top 25%.

For the ith school and jth measurement, the one-way random effects model is defined as

$$y_{ij} = \beta_0 + u_{0i} + \epsilon_{ij}, \tag{12}$$

where β_0 is the fixed effect and u_{0i} the individual deviation. More specifically, β_0 is the average of the school means or observed scores and for, say, the first school ($i = 1$), their respective observed score is defined as $\beta_0 + u_{01}$.

"True" Score Variance. The random effects are then assumed to be drawn from a common distribution, that is,

$$u_{0i} \sim \mathcal{N}(0, \sigma^2_{bi}).$$

Here the between-school variance σ^2_b captures the variability in the random effects $var(u_{0i})$, or the "true" score variance, that are assumed to be normally distributed with a mean of zero. We then predict the "true" score variance with the school-level rankings of test scores. This log-linear model is defined as

$$\sigma^2_{bi} = \exp\Big(\kappa_0^{(-25\%)} + \kappa_1^{(50\%)} X_i + \kappa_2^{(+25\%)} X_i\Big).$$

Note $\kappa_0^{(-25\%)}$ is the between-school variance for the schools in the bottom 25th percentile in scores at age 11. This serves as the reference category. Consequently, for, say, the schools in the top 25th percentile, the variance in their "true" scores is $\kappa_0^{(-25\%)} + \kappa_2^{(+25\%)}$.

Error Variance. Further, given the MELSM approach, the residuals or errors are also assumed to be normally distributed with a mean of zero and variance $\sigma^2_{\epsilon_{ij}}$, given as a function of a linear model, such that

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\epsilon_{ij}}) \text{ with}$$
$$\sigma^2_{\epsilon_{ij}} = \exp(\eta_0 + u_{1i})^2.$$

As indicated by the subscripts i and j, the error variance $\sigma^2_{\epsilon_{ij}}$, is now allowed to vary across i schools and j measurements given a log-linear model. The parameters in the scale model (the model for the error variance) are analogous to those in (14). That is, η_0 represents the intercept and defines the average of the within-school variances and u_{1i} represent the random effect, that is, the school departures from η_0. Again for the first school ($i = 1$), $\eta_0 + u_{11}$ is the variability of their exam scores. The random school effects are also normally distributed, $u_{1i} \sim \mathcal{N}(0, \sigma_1^2)$, and like the mean structure, they can be shrunken towards the average within-school variance η_0.

In the `ICCier` package, the maximal random effects covariance structure is always estimated (Barr, Levy, Scheepers, & Tily, 2013). Consequently, we do not write out the full model specification for each case study. Further details are provided in the appendix.

Reliability Indices. Together, this model allows for predicting cluster-specific reliability, given the within-school (error) and between-school ("true" score) variance. That is, for school i their respective reliability, or ICC(1), is

$$\rho_i = \begin{cases} \dfrac{\exp(\kappa_0)^2}{\exp(\kappa_0)^2 + \exp(\eta_0 + u_{1i})^2} & \text{if } \ -25\% \\ \dfrac{\exp(\kappa_0 + \kappa_1)^2}{\exp(\kappa_0 + \kappa_1)^2 + \exp(\eta_0 + u_{1i})^2} & \text{if } \quad 50\% \\ \dfrac{\exp(\kappa_0 + \kappa_2)^2}{\exp(\kappa_0 + \kappa_2)^2 + \exp(\eta_0 + u_{1i})^2} & \text{if } \ +25\% \end{cases}. \tag{13}$$

This can then be used to predict the average score reliability with (11), which as we explained above, also predicts the expected shrinkage in a one-way random effects model.

Results. We first compared the MELSM to a traditional, location only, random intercepts model. This is the traditional "CTT model." We used LOO (Vehtari, Gelman, & Gabry, 2017), which is analogous to AIC asymptotically, but accounts for posterior uncertainty. The difference was nearly 6 standard errors away from zero, which indicates that the MELSM has superior predictive accuracy.

Error Variance. This improved model fit with the MELSM can also be inferred from Figure 1. Panel A includes the within-school standard deviations, and the black bars are 90 % credible intervals that excluded zero. In total, 26% of the schools differed from the average within-school variance. This is important. Recall that this violates the implicit assumption of traditional reliability indices, for example ICC(1) defined in (1). Further, panel A also highlights a key aspect of the model. Namely, there is shrinkage towards the average within-school variance which, in some sense of the word, provides a more "reliable" estimate of variability. Consequently, the cluster-specific estimates of reliability are computed from smoothed estimates of the school-specific variances.

"True" Score Variance. Panel B includes the estimated between-school variability. This is another key aspect of the proposed methodology, in that each category of school-level test scores has their own estimate. Here it can be seen that the middle group, that is those in the middle 50% of LRT scores, were more homogeneous than the top 25 % group. For a point of reference, the 95% credible interval (CrI) for $\kappa_1^{(50\%)}$ excluded zero. In turn, this necessarily translates into the middle 50% group being less reliable. The implications of this cannot be understated, in that, when assuming fixed variances, this can mask important information related to achieving a desired level of reliability at the school level.

Reliability. Panel C includes the school-specific reliability estimates, or ICC(1), computed with (13). Recall that ICC(1) captures the correlation between any two observations from the same cluster. Importantly, these estimates are a combination of Panel A, that includes the within-school variance, and Panel B, that includes the between-school variance. They have been separated according to their test score ranking and the dotted line denotes the fixed effect or average reliability. The results are striking. For all schools in the middle 50% group, their respective 90% CrIs excluded the average reliability in the other categories. Said another way, based on the percentile rank of test scores, it appears that it is possible to differentiate schools based on reliability. This insight was made possible by fitting sub-models to both the between ("true" score) and within-school (error) variance structures.

We next investigated average score reliability, predicted with (11), while accounting for heterogeneity in the variance structures. The idea is that we want to determine the number of measurements needed to achieve adequate reliability. Figure 1 (panel D) displays these results. Note that each line represents a school. As a point of reference, we also included a customary, location only, model that embodies the assumptions of CTT. The critical differences is that the variance components in (1) are fixed and non-varying. For the CTT model, the black line is the predicted reliability and the grey ribbon is the corresponding 90% CrI. The results are again striking. Namely, for the middle 50% category, most of the schools were far below the predicted reliability of the CTT model. For example, on the one hand, the CTT model would have us believe that perhaps less than 25 measurements are needed to obtain, say, 80% reliability. However, on the other hand, by fitting sub-models to the variance components, it is clear that we might actually need 100 measurements! This is attributed to the smaller "true" score variance for schools that placed into the middle 50% of test scores (Figure 1, panel B).

Moreover, the results in panel D also translate into minimizing shrinkage towards the grand mean (Equation 10). This is important to consider when investigating between school mean differences. This is a result of the equivalence between average score reliability, or ICC(2), and the "shrinkage factor" given in (9). Consequently, these results can also be used to inform study design: Based on these results, it is clear that more measurements will be needed from the middle 25% percentile group.[2]

Case 2: Developmental Psychology

In this next section, we shift our focus from education research to infant cognitive development. Developmental researchers have long been interested in measuring individual differences with the desire to predict future behavior or developmental outcomes (Colombo

[2] In these data, there was a large correlation ($r \approx 0.70$) between ICC(2) and the difference between the empirical and shrunken mean estimates. Hence, the less reliable schools tended to also be shrunken more towards the grand mean.

& Fagen, 2014). For example Rose, Feldman, and Wallace (Rose, Feldman, & Wallace, 1988) found that performance on visual recognition tasks in a sample of 6-, 7-, and 8-month-old infants predicted later Stanford-Binet IQ scores at 3 years of age.

A prerequisite of psychological tasks used to measure individual differences is their ability to produce reliable measurements from participants. That is, repeated measurements obtained from a single individual must be correlated (to some extent) with other measurements obtained from the *same* individual on the same task. Recent work has drawn attention to the observation that several commonly used cognitive tasks fail to produce reliable measurements in adults (Hedge et al., 2018; Rouder, Kumar, & Haaf, 2019). As a consequence, the ability for these tasks to measure individual differences has been called into question.

The finding that commonly used cognitive tasks are not well suited to study individual differences is surprising and has generated interest in exploring whether these findings also apply to tasks used in developmental psychology (Arnon, 2018). However, it should be noted that these conclusions are drawn from models that assume a common variance between individuals. As previously demonstrated, the assumption of a common variance does not hold in many observed adult data sets (Williams, Martin, & Rast, 2019). Hence, a core assumption of measurement reliability is routinely violated (i.e., homogeneous variance). And to date a solution to address the so-called "reliability paradox" has not been proposed. Our methodology not only accommodates heterogeneous variance, but as we show below, opens the door for explaining aspects of reliability–that is, "true" score and error variance. This can provide novel insights into cognitive tasks that could then be used to improve measurement.

Model Specification. We now apply the MELSM to characterize performance on the IOWA task (described in more detail below). Our focus here is two-fold. First, there is emerging evidence that heterogeneous within-person variance is a defining feature of cognitive tasks. However, to date, this has only been investigated in college-aged adults.

Hence, in this case study, we first characterize within-infant variability and also infant-specific reliability in an experimental task that has four conditions. We then dive deeper into the reliability of those conditions. Here we first look at how the number of trials completed influences reliability and then we assess reliability differences between conditions.

We assessed infants using an adaptation of the Infant Orienting With Attention (IOWA) task developed by Ross-Sheehy and colleagues (Ross-Sheehy, Schneegans, & Spencer, 2015). We use eyetracking data collected from 98 full term typically developing infants ranging from 5- to 12-months of age. In this task, each trial begins with a central fixation stimulus (a looming smiley face paired with classical music). As infants fixate that stimulus, a 100 ms spatial attention cue (a small black dot) is presented left or right of midline, followed by a target (a realistic photograph of an object, e.g., rattle or banana). Infants received four types of trials: valid cue trials (target appeared in the location of the cue), invalid cue trials (target appeared on the opposite side of the cue), double cue trials (two cues appeared followed by a single target), and no cue baseline trials (no cue before the target appeared; this condition served to measure the infants' *baseline* responding when no cue was present).

Spatial attention and orienting speed were assessed using reaction time, or latency to fixate the target on each trial. We fitted separate models for each of the four conditions. For the ith infant and jth measurement, the one-way random effects model is defined as

$$y_{ij}^{(c)} = \beta_0^{(c)} + u_{0i}^{(c)} + \epsilon_{ij}^{(c)}. \tag{14}$$

Note that $(c) = \{1, 2, 3, 4\}$ is used to denote the respective experimental condition. Hence $\beta_0^{(c)}$ is the fixed effect and $u_{0i}^{(c)}$ the individual deviation for a given condition c. More specifically, $\beta_0^{(c)}$ is the average of the infants means or observed scores and for, say, the first infant ($i = 1$), their respective observed score is defined as $\beta_0^{(c)} + u_{01}^{(c)}$.

"True" Score Variance. The random effects are then assumed to be drawn from a common distribution, that is,

$$u_{0i}^{(c)} \sim \mathcal{N}(0, \sigma_{bi}^{2(c)}).$$

Here the between-infant variance $\sigma_{bi}^{2(c)}$ captures the variability in the random effects $var(u_{0i})^{(c)}$, or the "true" score variance, that are assumed to be normally distributed with a mean of zero. We the predict the "true" score variance with the number of trials completed by each infant. This explores the possibility that perhaps reliability is function of trials completed. This log-linear model is defined as

$$\sigma_{bi}^{2(c)} = \exp(\kappa_0^{(c)} + \kappa_1^{(c)} X_i) \tag{15}$$

Here we employed grand mean centering. Hence $\kappa_0^{(c)}$ is the "true" score variance at the average number of trials. We also scaled the predictor such that $\kappa_1^{(c)}$ corresponds to the regression weight for a 10 trial increase. In other words, rather than assuming σ_b^2 is fixed and non-varying, as in CTT, this formulation allows for explaining "true" score variance in a cognitive task.

Error Variance. Further, with the MELSM approach, the residuals or errors are also assumed to be normally distributed with a mean of zero and variance $\sigma_{\epsilon_{ij}}^2$, given as a function of a linear model, such that

$$\epsilon_{ij}^{(c)} \sim \mathcal{N}(0, \sigma_{\epsilon_{ij}}^{2(c)}) \text{ with} \tag{16}$$

$$\sigma_{\epsilon_{ij}}^{2(c)} = \exp(\eta_0^{(c)} + u_{1i}^{(c)})^2. \tag{17}$$

As indicated by the subscripts i and j, the error variance $\sigma_{\epsilon_{ij}}^{2(c)}$, is now allowed to vary across i infants and j measurements given a log-linear model. The parameters in the scale model (the model for the error variance) are analogous to those in (14). That is, $\eta_0^{(c)}$ represents the intercept and defines the average of the within-infant variances and $u_{1i}^{(c)}$ represent the random effect, that is, the individual departures from $\eta_0^{(c)}$. Again for the first infant ($i = 1$), $\eta_0^{(c)} + u_{11}^{(c)}$ is the variability of their latencies for condition c. The random infant effects are also normally distributed, $u_{1i}^{(c)} \sim \mathcal{N}(0, \sigma_1^{2(c)})$, and like the mean structure, they can be shrunken towards the average within-infant variance $\eta_0^{(c)}$. This can be seen in Figure 1 (panel A).

Reliability Indices. With the model specification in hand, we can now define reliability indices for each condition c and infant i, that is,

$$\rho_i^{(c)} = \frac{\exp(\kappa_0^{(c)} + \kappa_1^{(c)} X_i)^2}{\exp(\kappa_0^{(c)} + \kappa_1^{(c)} X_i)^2 + \exp(\eta_0^{(c)} + u_{1i}^{(c)})^2}. \tag{18}$$

This index is infant-specific ICC(1) and X_i corresponds to a given number of trials completed for infant i. The "true" score variance, or the numerator in (18), is then a function of trials completed. Consequently $\rho_i^{(c)}$ is computed with respect to the number of trials completed for each infant and their respective error variance, $\exp(\eta_0^{(c)} + u_{1i}^{(c)})^2$. Note that trials completed is not an individually varying predictor, but is instead a so-called level two variable. This allows for predicting reliability, given the fixed effect averages. This predicted reliability index is defined as

$$\hat{\rho}^{(c)} = \frac{\exp(\kappa_0^{(c)} + \kappa_1^{(c)} \cdot \text{Trials})^2}{\exp(\kappa_0^{(c)} + \kappa_1^{(c)} \cdot \text{Trials})^2 + \exp(\eta_0^{(c)})^2}. \tag{19}$$

Recall from (15) that $\kappa_0^{(c)}$ is the between-infant variance for the average number of trials completed (i.e., the grand mean) and $\kappa_1^{(c)}$ the corresponding regression weight. This allows for predicting reliability across a range of trials completed. Hence, this allows for investigating the effect of completing more trials on reliability at the level of "true" score variance. Further, because the package `ICCier` employs Bayesian estimation, $\hat{\rho}^{(c)}$ has a full distribution which readily allows for assessing reliability differences, say, between conditions and as a function of a level two predictor. This is demonstrated below. This powerful approach for probing reliability is a major and novel contribution to the measurement literature.

Results. We first compared the MELSM to a traditional, location only, random intercepts model. This was again done with LOO (Vehtari et al., 2017). The differences ranged from at most 9.4 (control condition) to 6.1 standard errors (double condition) away from zero, which indicates that the MELSMs had superior model fit compared to the customary 'CTT model."

Error Variance. The improved fit of the MELSM can be inferred from Figure 2 (panel A), which includes the within-infant standard deviations SD. Note that for aesthetic reasons they are plotted on the logarithmic scale. The results are striking. That is, across each experimental condition, there is considerable variance heterogeneity at the infant level. Indeed, the proportion of infants that differed from the average within-infant SD (the dotted line) ranged from 0.48 (valid-cue) to 0.63 (baseline). As described in Williams, Martin, and Rast (2019), the average within-infant variance is used to compute traditional reliability indices (Arnon, 2018). Hence, for these data, it is clear that computing reliability with the average would mask individual differences. And it is not entirely clear, from our perspective, what one reliability metric would mean when such variation is present. For example, the maximum–minimum ratio between within-infant variances ranged from 23 (baseline) to 108 (double cue). These ratios were obtained from the shrunken estimates and thus the empirical ratios were much larger. This is further described below (Section *Reliability Indices*).

"True" Score Variance. We also predicted "true" score variance with the number of trials completed for each infant. These results are included in the Appendix (Table A1). Notably, in all four conditions there was a negative effect, such that, given an *increase* in the number of trials completed, "true" score variance is expected to *decrease*. In other words, those infants that completed the most trials tended to also be the most homogeneous clusters in the sample. Importantly, the posterior probability of a negative effect exceeded 95% for three conditions. The exception was the invalid cue condition, where the posterior probability was 87%. The double-cue condition in particular had a large effect. An additional 25 trials completed resulted in “true” score variance reducing from 0.32 to 0.21. A decrease of 34%. We emphasize that this result does *not* imply, in general, that increasing the number of trials completed reduces “true” score variance. And we encourage readers to *not* extend this finding beyond what is an explicitly exploratory context. However, this example does make clear that questions related to improving reliability can be investigated at the level of “true” score variance. This has far reaching implications–“true” score variance can be a function of covariates, which opens the door for explaining a key aspect of reliability.

Reliability Indices. Figure 2 (panel B) includes the infant-specific estimates of reliability. These were computed with (18), which considers the number of trials completed for each infant and their within-infant variance estimate, respectively. The utility of computing cluster-specific reliability is clear. That is, it is readily apparent that there is substantial variation in reliability at the infant level. Importantly, this reliability index is ICC(1), which is the expected correlation between any two observations from the same infant. Hence, for many infants, their latencies were highly correlated. On the other hand, for other infants, their latencies were nearly independent.

It is informative to consider ICC(2), for average scores, which can be computed with (11). Given ICC(1) values of 0.1 and 0.7, and wanting to know the average score reliability of, say, 5 observations, this would translate into ICC(2) values of 0.35 and 0.92. Hence, for some infants, few observations would be needed to reach adequate reliability, whereas for other infants, we would need to gather more measurements. This insight is made possible by extending CTT to accommodate individual differences in reliability.

Panel C includes estimates of reliability that are expressed a function of the number of trials completed. This index was given in (19). In this example, we predicted ICC(1) with a sequence of trials completed that ranged from 1 to 100. A consistent pattern emerged, in that, for all conditions, reliability is expected to decrease with more trials completed. This suggests that observations from the same infant approached independence with more trials completed. Said another way, *in these data*, individual differences tend to be smallest when completing more trials. This should not be interpreted to mean that gathering many trials is necessarily disadvantageous for detecting individual differences. This notion would be naive. For example, with few measurements, the individual effects will also have wider credible intervals. Hence we encourage researchers to interpret this result with care and consider that “true” score variance *could* be related to the number of trials completed in this particular task and age group. The presented methodology not only raised this thought, but it can also be used to look further into this intriguing result.

Panel D also includes predicted reliability as a function of trials completed, but in this case, it is expressed as a difference between conditions. That is, we predicted reliability given a number of trial completed, as in panel C, and then computed pairwise differences in ICC(1). The shaded regions corresponds to values of the covariate (trials completed) that excluded zero. Consider the bottom-right panel, which includes the contrast between the invalid and valid-cue. Here the invalid-cue had notably higher reliability when fewer trials were completed and gradually became similar to the valid-cue condition with more trial completed. This difference was not small, in that, for 25 trials completed, the invalid-cue ICC(1) was 0.60 and the valid-cue ICC(1) was 0.30. On the other hand, while this method does not provide evidence for invariant reliability (i.e., the null hypothesis),[3] it is still informative to note conditions that did not differ from one another. For example, at no value of the covariate (trials completed) did reliability differ between the baseline compared to both the valid and invalid-cue. We emphasize the novelty and utility of this approach. Because we modeled the variance components, this opened the door for

[3] This could be directly tested, but would require introducing an extension to allow for Bayesian hypothesis testing. This is beyond the scope of the current paper.

investigating reliability with an unprecedented level of detail.

Discussion

In this work, we proposed a novel approach for investigating measurement reliability in hierarchical models. The primary motivation for developing this methodology was that classical test theory appears to be incompatible with commonly observed phenomena in psychological applications. Namely, that homogeneous variance appears to be the exception and not the rule. We noted that fixed and non-varying variance components are assumed to be the case in ANOVA and traditional hierarchical models, and thus also assumed when computing reliability. Our methodology not only relaxes this assumption, but it allows the ‘true" score and error variance to be a function of covariates, including individual variation therein. Hence, the reliability of a questionnaire, measurement device, or experimental task, can be *explained*. As we demonstrated, this can provide unique insights into reliability. These were made possible by merging ideas stemming from the variance modeling literature with classical test theory.

Practical Implications

There are several practical implications of this work. First, in educational research, we demonstrated the utility of modeling the variance structure for assessing standardized test scores. In particular, we demonstrated that both the “true" score and error variance can be the target of an explanatory model. This can inform study design, for example sampling strategies, with the goal of utilizing school-level information to obtain a specific level of reliability.

Second, we also demonstrated that average score reliability is equivalent to the “shrinkage" factor in a one-way random effects model. Hence, the Spearman-Brown equation in (11) (Brown, 1910; Spearman, 1910), which is used to forecast reliability when the test length is increased, can also be understood as predicting the degree of smoothing towards the overall average. In other words, the homogeneity of “true" scores can be predicted (Figure 1, panel D), as a function of covariates, with the goal of ensuring they are maximally different from one another. This is important for detecting cluster differences, which in the context of educational research, would translate into testing between-school differences in their average test scores. Note that this implication is not restricted to educational settings. It readily applies to all areas interested in studying cluster differences. For example, this is analogous to individual difference research in, say, human subjects.

Third, in cognitive psychology, there is recent debate surrounding the adequacy of commonly used experimental task for studying individual differences. The emerging consensus is that reliability is too low (i.e., "noisy" measures) to adequately study individual variation. However, to our knowledge, this discussion has revolved almost exclusively around the mean structure and avoided the within-person variance structure altogether. This is unfortunate, because reliability has routinely been computed from tasks that have heterogeneous error variance. But reliability indices assume a common error variance for each person (Equation 1). Therefore, from our perspective, a satisfactory answer to the question of individual differences in, say, the "Stroop effect," would require addressing the extreme heterogeneity in within-person variance (and thus reliability) that is apparently a defining feature of these kinds of tasks (Figure 2, panel A).

Moreover, recall that this work is partially motivated by recent findings demonstrating that homogeneous within-person variance appears to be the exception and not the rule in psychological applications, including cognitive tasks. This is not readily accounted for by current measurement theories. However, these insights have thus far been restricted to college age students Williams, Martin, and Rast (2019) and older adults Williams, Liu, et al. (2019). In other words, to date, it was not clear whether this general pattern would generalize to, say, within-infant variance. Hence, this work adds to the growing literature on within-person variance heterogeneity that now spans from infancy to adulthood. And these findings lend further credence to the notion that heterogeneous variance (and thus reliability) is a defining feature of commonly used cognitive tasks.

Importantly, we have done more than highlight this issue as it relates to reliability. We have provided a methodological solution, and importantly, a conceptual framework based upon the tradition of heterogeneous variance modeling (Cleveland et al., 2003; Hedeker et al., 2008, 2012; Lindley, 1969).

Methodological Implications

The primary aim of this work was to provide the necessary ingredients to extend classical test theory. Of course, various alternatives have been proposed, including generalizability theory (Brennan, 1992) and item response theory (DeMars, 2018). Our aim, however, was to explicitly stay within the CTT framework, but with the addition of modeling "true" score and error variance. Accordingly, the foundation for thought remains in tact, but with the possibility of predicting the variance components. Hence, rather than view reliability as a stable property of, say, an experimental task, we can fit explanatory models with the goal of finding avenues for improvement. For example, by identifying predictors of "true" score or error variance, this can be used to refine measurement. Such possibilities were demonstrated in the case studies, although these examples just scratched the surface of possibilities. For example, we focused on predicting "true" score variance, whereas identifying sources of within-group variance is also important to consider (i.e., error variance, Karch et al., 2019)

Importantly, in our view, historical context should be taken into account when considering to expand a theory. CTT (and *G*-theory) is inextricably linked to ANOVA (Fisher, 1925)–"Often, CTT and analysis of variance (ANOVA) are viewed as the parents of G theory" (p. 7, Brennan, 2011)–with its origins going back to a time when the primary mode of transportation was horse and buggy. Just as the horse can limit possibilities for travel, so to can ANOVA (and traditional mixed models) for exploring measurement reliability. In this respect, our work can be seen as an extension to CTT; that is, a natural progression from classical to modern methodology. This brings full circle the arguments of

Lindley (1969) and Novick et al. (1971), where it was originally noted that Bayesian methods offer flexibility not possible with ANOVA.[4] The user-friendly R package `ICCier` can facilitate this transition towards richer models for investigating measurement reliability, all the while staying within the CTT framework.

An Alternative Perspective

It would be remiss of us to not offer an alternative perspective. It is customary to view the residuals as mere "noise" and perhaps measurement "error." For example, that trial to trial fluctuations are a nuisance to understanding the latent process. On the other hand, there is a large literature that views these same fluctuations as a key aspect of the construct. A good example is personality traits, that were customarily considered fixed, but now an active area of research revolves around within-person variability of these traits (i.e., the fluctuations; Fleeson, 2001; Hutteman, Back, Geukes, Küfner, & Nestler, 2016; Williams, Liu, et al., 2019). So rather than there being individual differences in reliability, the alternative perspective is to view these as individual differences in stability. That is, individuals with larger residual variance are relatively more volatile or inconsistent. Further, instead of using an adjective to label between-group variance "true," this perspective would view homogeneity of means as an important aspect of the construct. Consider age groups. An interesting research question is whether older adults are more or less homogeneous in, say, cognitive abilities. This is diametrically opposed to CTT, and thus the reliability literature, where measurements are construed as a "true" score plus error. However, in our opinion, this alternative perspective is worth considering.

[4] The R package `nlme` can be used for simple heterogeneous variance models. There are key disadvantages, however, including that the variances are not hierarchically modeled and that a measure of uncertainty is not readily available. Further, the models fitted in this paper failed to converge in `nlme`.

Future Directions

The proposed methodology provides a foundation for further quantitative advances. First, although we stayed within the CTT framework, the basic ideas readily apply to, say, G-theory. In this case, explanatory models could be fitted to several variance components. The model described in the section Model Formulation is can seamlessly accommodate this extension. Second, this methodology could be extended to consider test-retest reliability. This is especially important, because, to our knowledge, heterogeneous within-group variance has not been considered in test-retest situations. These ideas point towards our future work.

Conclusion

Measurement reliability has traditionally been considered a stable property of a measurement device or experimental task. This framework does not allow for the *explaining* reliability, because it assumes the “true” score and error variance are fixed and non-varying. However, heterogeneous variance modeling allows for investigating reliability with a fine-tooth comb. The illustrative examples highlighted such possibilities. That is, we demonstrated that there can be substantial variability in error variance, which necessarily implies the same for reliability, and that “true” score variance can be a function of covariates. The methodology that made these novel insights possible is implemented in the R package `ICCier`.

References

Aitkin, M. (1987, 6). Modelling Variance Heterogeneity in Normal Regression Using GLIM. *Applied Statistics*, *36*(3), 332. doi: 10.2307/2347792

Arnon, I. (2018). Do current statistical learning capture stable individual differences in children? An investigation of task reliability across modalities.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Bartko, J. J. (1966, 8). The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, *19*(1), 3–11. doi: 10.2466/pr0.1966.19.1.3

Betancourt, M. (2017, 1). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv*.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018, 6). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, *50*(3), 937–962. doi: 10.3758/s13428-017-0918-2

Bliese, P. D. (2000). *Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis* (1st ed.; K. J. Klein & S. W. Kozlowski, Eds.). San Francisco: Jossey-Bass.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. doi: 10.1080/08957347.2011.532417

Brown, W. (1910). SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES. *British Journal of Psychology, 1904[U+2010]1920*, *3*(3), 296–322. doi: 10.1111/j.2044-8295.1910.tb00207.x

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society. Series A: Statistics in*

Society, *180*(2), 551–568. doi: 10.1111/rssa.12205

Cleveland, W. S., Denby, L., & Liu, C. (2003). Random scale effects. (2), 33. Retrieved from stat.bell-labs.com

Colombo, J., & Fagen, J. (2014). *Individual differences in infancy: Reliability, stability, and prediction.* Psychology Press.

Cronbach, L. J. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, 1–33.

de Boer, H., Timmermans, A. C., & van der Werf, M. P. C. (2018, 4). The effects of teacher expectation interventions on teachers' expectations and student achievement: narrative review and meta-analysis. *Educational Research and Evaluation*, *24*(3-5), 180–200. doi: 10.1080/13803611.2018.1550834

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t- test. *International Review of Social Psychology*, *30*(1), 92–101. doi: 10.5334/irsp.82

DeMars, C. E. (2018, 2). Classical Test Theory and Item Response Theory. In *The wiley handbook of psychometric testing* (pp. 49–73). Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/9781118489772.ch2

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver & Bond.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, *80*(6), 1011–27.

Gelfand, A., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal Of The American Statistical Association*, *85*(410), 398–409. doi: 10.2307/2289776

Gelman, A., Hill, J., & Yajima, M. (2012, 4). Why We (Usually) Don't Have to Worry

About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. doi: 10.1080/19345747.2011.618213

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A Multilevel Analysis of School Examination Results. *Oxford Review of Education*, *19*(4), 425–433. doi: 10.1080/0305498930190401

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the Frontiers of Modeling Intensive Longitudinal Data: Dynamic Structural Equation Models for the Affective Measurements from the COGITO Study. *Multivariate Behavioral Research*, *53*(6), 820–841. doi: 10.1080/00273171.2018.1446819

Haveman, R., & Smeeding, T. (2006). The role of higher education in social mobility. *The Future of children*, *16*(2), 125–50.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*(2), 627–634. doi: 10.1111/j.1541-0420.2007.00924.x

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336. doi: 10.1002/sim.5338

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. doi: 10.3758/s13428-017-0935-1

Hoffman, L. (2007). Multilevel Models for Examining Individual Differences in Within-Person Variation and Covariation Over Time. *Multivariate Behavioral Research*, *42*(4), 609–629. doi: 10.1080/00273170701710072

Hutteman, R., Back, M. D., Geukes, K., Küfner, A. C., & Nestler, S. (2016). Trait personality and state variability: Predicting individual differences in within- and cross-context fluctuations in affect, self-evaluations, and behavior in everyday life.

Journal of Research in Personality, *69*, 124–138. doi: 10.1016/j.jrp.2016.06.003

Karch, J. D., Filevich, E., Wenger, E., Lisofsky, N., Becker, M., Butler, O., … Kühn, S. (2019, 10). Identifying predictors of within-person variance in MRI-based brain volume estimates. *NeuroImage*, *200*, 575–589. doi: 10.1016/J.NEUROIMAGE.2019.05.030

Kelley, T. L. (1927). *The Interpretation of Educational Measurements.* New York: World Book.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation.* Thousand Oaks: Sage.

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, *39*(5), 307–332.

Lindley, D. (1969, 12). A BAYESIAN ESTIMATE OF TRUE SCORE THAT INCORPORATES PRIOR INFORMATION. *ETS Research Bulletin Series*, *1969*(2), i-8. doi: 10.1002/j.2333-8504.1969.tb00754.x

Logan, J. R., & Burdick-Will, J. (2017, 11). School Segregation and Disparities in Urban, Suburban, and Rural Areas. *The Annals of the American Academy of Political and Social Science*, *674*(1), 199. doi: 10.1177/0002716217733936

Lord, F. M. (1955). Estimating test reliability. *Educational and psychological measurement*(8), 30.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Oxford: IAP.

McCulloch, C. E. (2003). Chapter 2: Linear mixed models (LMMs). In *Generalized linear mixed models* (Vol. Volume 7, pp. 9–20). Beechwood OH and Alexandria VA: Institute of Mathematical Statistics and American Statistical Association.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. doi:

10.1037/1082-989X.1.1.30

Morris, C. N., & Lysy, M. (2012). Shrinkage Estimation in Multilevel Normal Models. *Statistical Science*, *27*(1), 115–134. doi: 10.1214/11-sts363

Myles, J., Price, G., Hunter, N., Day, M., & Duffy, S. (2003). A potentially useful distribution model for dietary intake data. *Public Health Nutrition*, *6*(5), 513–519. doi: 10.1079/phn2003459

Novick, M. R. (1965). The axioms and principal results of classical test theory. *ETS Research Bulletin Series*, *1965*(1), i–31.

Novick, M. R., Jackson, P. H., & Thayer, D. T. (1971, 9). Bayesian inference and the classical test theory model: Reliability and true scores. *Psychometrika*, *36*(3), 261–288. doi: 10.1007/BF02297848

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata.* STATA press.

Rast, P., & Ferrer, E. (2018). A Mixed-Effects Location Scale Model for Dyadic Interactions. , 1–63. doi: 10.1080/00273171.2018.1477577

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling Individual Differences in Within-Person Variation of Negative and Positive Affect in a Mixed Effects Location Scale Model Using BUGS/JAGS. *Multivariate Behavioral Research*, *47*(2), 177–200. doi: 10.1080/00273171.2012.658328

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks: Sage Publications.

Robert, C., & Casella, G. (2011, 8). A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, *26*(1), 102–115. doi: 10.1214/10-STS351

Rose, S. A., Feldman, J. F., & Wallace, I. F. (1988). Individual differences in infants' information processing: Reliability, stability, and prediction. *Child Development*, 1177–1197.

Ross-Sheehy, S., Schneegans, S., & Spencer, J. P. (2015). The infant orienting with attention task: Assessing the neural basis of spatial attention in infancy. *Infancy*, *20*(5), 467–506.

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. *PsyArXiv*, 1–37.

Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology*, *8*(1), 1–11. doi: 10.1027/1614-2241/a000034

Sandy, J., & Duncan, K. (2010, 9). Examining the achievement test score gap between urban and suburban students. *Education Economics*, *18*(3), 297–315. doi: 10.1080/09645290903465713

Shavelson, R. J., & Webb, N. M. (2010). Generalizability Theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (Vol. 24, pp. 27–34). Thousand Oaks: Wiley Online Library.

Shavelson, R. J., & Webb, N. M. (2012). Generalizability Theory. In J. Green, G. Camilli, & P. Elmore (Eds.), *Handbook of complementary methods in education research* (3rd ed.). Routledge.

Shieh, G. (2016). Choosing the best index for the average score intraclass correlation coefficient. *Behavior Research Methods*, *48*(3), 994–1003. doi: 10.3758/s13428-015-0623-y

Sirin, S. R. (2005, 9). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. doi: 10.3102/00346543075003417

Spearman, C. (1910, 10). CORRELATION CALCULATED FROM FAULTY DATA. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Stan Development Team. (2016). *Rstan: the R interface to Stan.* Retrieved from http://mc-stan.org/

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the {t}hird {b}erkeley {s}ymposium on {m}athematical {s}tatistics and {p}robability, 1954–1955, vol. {i}* (pp. 197–206). University of California Press, Berkeley and Los Angeles.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* The Science Press.

Thouless, R. H. (1936). Test Unrealiability and Function Fluctuation. *British Journal of Psychology*, *26*(4), 325.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, *16*, 8–13.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Watts, A., Walters, R. W., Hoffman, L., & Templin, J. (2016). Intra-individual variability of physical activity in older adults with and without mild Alzheimer's disease. *PLoS ONE*, *11*(4). doi: 10.1371/journal.pone.0153898

Williams, D. R., Liu, S., Martin, S. R., & Rast, P. (2019). Bayesian Multivariate Mixed-Effects Location Scale Modeling of Longitudinal Relations among Affective Traits, States, and Physical Activity. doi: 10.31234/OSF.IO/4KFJP

Williams, D. R., Martin, S. R., & Rast, P. (2019). Putting the Individual into Reliability: Bayesian Testing of Homogeneous Within-Person Variance in Hierarchical Models. *PsyArXiv*. doi: 10.31234/OSF.IO/HPQ7W

Williams, D. R., Rouder, J., & Rast, P. (2019). Beneath the Surface: Unearthing Within-Person Variability and Mean Relations with Bayesian Mixed Models. *PsyArXiv*. doi: 10.31234/OSF.IO/GWATQ

Williams, D. R., Zimprich, D. R., & Rast, P. (2019, 5). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods*, 1–19. doi: 10.3758/s13428-019-01255-9

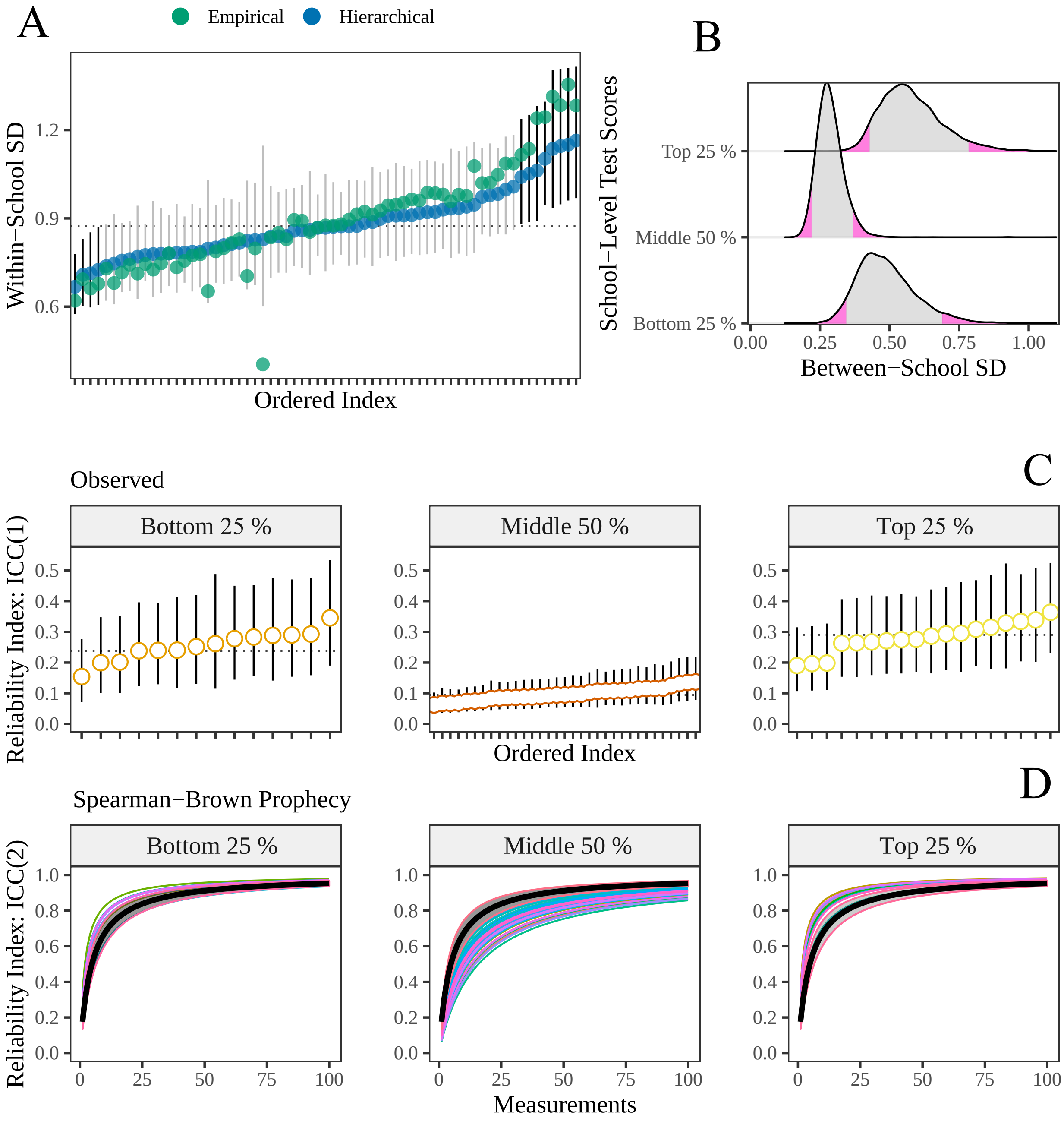

Figure 1. Results for Case 1: Educational Research. A) Within-school standard deviations (SD). The error bars are 90% credible intervals (CrI). The black error bars excluded the dotted line that denotes the average, or fixed effect, within-school SD. The blue points are the shrunken estimates that have been smoothed towards the fixed effect. The green points are the empirical, non-shrunken, within-school SDs. B) Posterior densities for between-school ("true" score) variability according to school-level rankings. The grey area is the 90% highest density interval. C) School-specific reliability that are separated according to school-level rankings. ICC(1) is for single observations.The dotted lines denote the average reliability for each group. D) School-specific reliability predicted with the Spearman-Brown equation (11). ICC(2) is for mean scores. Each school is represented by a colored line. The black line is the classic "CTT model" that assumes the variance components (i.e., "true" score and error variance) are fixed and non-varying. The grey region is the 90% CrI for the "CTT model."

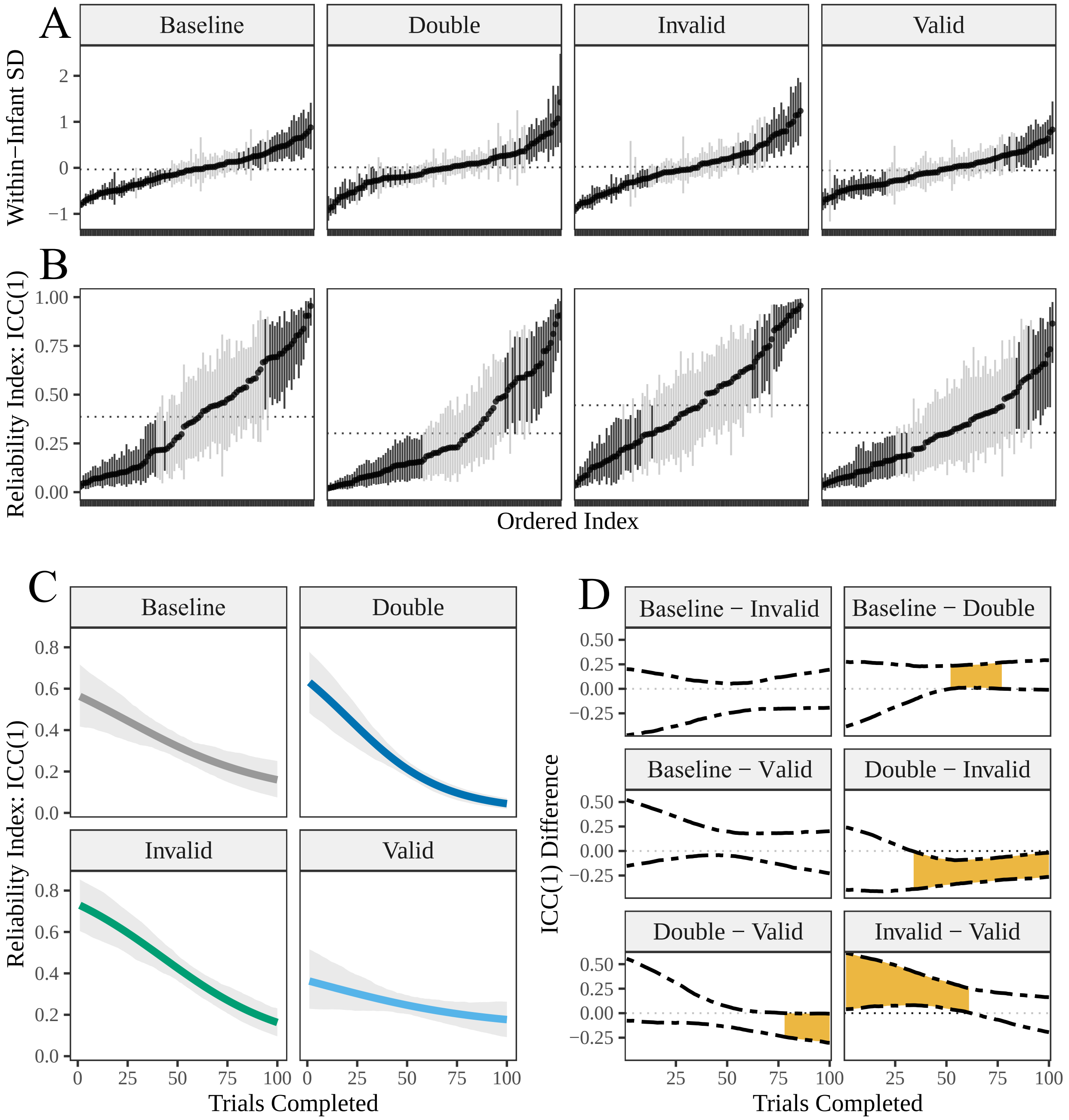

Figure 2. Results for Case 2: Developmental Psychology. A) Within-infant standard deviations (*SD*) on the log scale. The error bars are 90% credible intervals (CrI). The black error bars excluded the dotted line that denotes the average, or fixed effect, within-infant *SD*. B) Infant-specific reliability that are separated according to experimental condition. ICC(1) is for single observations. The dotted lines denote the average reliability for each condition. The black error bars excluded the dotted line that denotes the average, or fixed effect, infant-specific reliability. C) Predicted reliability as a function of the number of trials completed. The shaded bands correspond to the posterior *SD* of the predicted reliability. This is analogous to one standard error. D) Predicted reliability differences between conditions. The dotted line is the “null” value of zero. The shaded regions correspond to areas in which the difference in predicted reliability excluded zero for that value of the covariate (trials completed). The confidence bands correspond to 90% credible intervals.

Appendix

Supplementary Results

Table A1

Sub-model estimates for predicting "true" score variance

	Estimates			
Parameter	M	SD	90% CrI	$p(\kappa_1 < 0 \mid \mathbf{Y})$
$\exp[\kappa_0^{(control)}]$	0.37	0.05	[0.31, 0.45]	–
$\exp[\kappa_0 + (\kappa_1 \cdot 2.5)]$	0.30	0.06	[0.21, 0.41]	0.97
$\exp[\kappa_0^{(double)}]$	0.32	0.05	[0.26, 0.39]	–
$\exp[\kappa_0 + (\kappa_1 \cdot 2.5)]$	0.21	0.06	[0.14, 0.29]	0.99
$\exp[\kappa_0^{(invalid)}]$	0.45	0.05	[0.38, 0.53]	–
$\exp[\kappa_0 + (\kappa_1 \cdot 2.5)]$	0.35	0.07	[0.27, 0.46]	0.99
$\kappa_0^{(valid)}$	0.38	0.05	[0.31, 0.46]	–
$\exp[\kappa_0 + (\kappa_1 \cdot 2.5)]$	0.34	0.07	[0.25, 0.45]	0.87

Note. Posterior mean (M) and standard deviation (SD). κ_0 is the intercept and the "true" score variance for the average number of completed trials. κ_1 is the regression weight. $p(\kappa_1 < 0 \mid \mathbf{Y})$ denotes the posterior probability of a negative effect. To ease interpretation, we computed the "true" score variance for an increase of 25 trials completed. This corresponds to $\exp[\kappa_0 + (\kappa_1 \cdot 2.5)]$.