To cite this article:

Lin, T.-Y., Tuerlinckx, F., & Vanbelle, S. (in press). Reliability for multilevel data: A correlation approach. *Psychological Methods*. https://doi.org/10.1037/met0000738

Reliability for multilevel data: A correlation approach

Tzu-Yao Lin[1, 2], Francis Tuerlinckx[2], and Sophie Vanbelle[1]

[1]Department of Methodology and Statistics, Care and Public Health Research Institute, Maastricht University, the Netherlands

[2]Quantitative Psychology and Individual Differences Research Group, Faculty of Psychology and Educational Sciences, KU Leuven, Belgium

Abstract

Studying the reliability of a measurement instrument is essential. Despite the recognition of the importance of reliability in psychology and medicine and the various reliability coefficients that have been proposed, research on reliability for nested or multilevel data, ubiquitously in observational studies, remains limited. Two recent papers (Schönbrodt et al., 2022; ten Hove et al., 2022) address how to quantify reliability in multilevel settings based on generalizability theory. Specifically, ten Hove et al. (2022) defined between-cluster and within-cluster interrater intraclass correlation coefficients for multilevel designs where persons or raters are nested within clusters. Schönbrodt et al. (2022) also defined reliability coefficients at between-cluster and within-cluster (i.e., between-person) levels for designs where persons nested in couples are assessed numerous times daily over a number of days. Nevertheless, when applied to a common design, both approaches give inconsistent results regarding their definition of cluster-level reliability. In this paper, we propose an alternative approach to defining reliability coefficients for multilevel data that are based on calculating the expected correlation between repeated measurements. We will compare our approach with that of Schönbrodt et al. (2022) and ten Hove et al. (2022) and explain the differences between the three approaches in a number of common nested data structures: (1) raters crossed with both persons and clusters, but persons are nested within clusters, (2) raters nested within both persons and clusters, and (3) persons nested in clusters and crossed with raters and days.

Keywords: generalizability theory, intraclass correlation coefficient, hierarchical design, nested data, repeated measurements

Reliability for multilevel data: A correlation approach

Introduction

Precise measurement is a cornerstone of science. However, a perfectly precise measurement does not exist because any measurement is tainted by errors, thereby setting a limit on its precision. These principles hold for simple measurement devices (e.g., a kitchen scale) to very complex ones (e.g., a gravitational wave detector). For each measurement instrument, it is important to quantify its precision.

Also, in the behavioral sciences and related fields, measurement precision is of utmost importance. Psychometric research has a long and rich tradition of conceptualizing what precise measurement means when dealing with measurement instruments such as self-reports or observations. In this context, precision is often quantified through reliability, which can be considered a unitless indicator of precision. For psychometricians, a reliable (and thus precise) measurement instrument is a procedure that gives similar results when applied repeatedly under the same circumstances. In contrast, an unreliable measurement instrument means that, under the same circumstances, different results are obtained. Obviously, the validity of an inference based on an unreliable measurement instrument will be compromised (Shoukri, 2010; Shrout & Lane, 2012b; Streiner et al., 2015).

The core of classical test theory (Lord et al., 1968) deals with the definition and ways to estimate reliability for standardized tests: A sample of testees is measured with the same set of items under identical circumstances. However, researchers in the behavioral and health sciences also need to assess the reliability of measurement instruments in studies where the observed measurements have one or multiple identifiable sources of error. A generalization of classical test theory to deal with these situations is generalizability theory (Brennan, 2001; Cronbach, 1972; Shavelson et al., 1989), a statistical framework that recognizes and estimates the magnitude of multiple sources of measurement error.

Generalizability theory has been developed for several decades and has a much wider application range than classical test theory. However, it is not until very recently

that it has been applied to more complicated situations in which there is a clustering of units in groups, leading to a multilevel or hierarchical structure (Jeon et al., 2009; Schönbrodt et al., 2022; ten Hove et al., 2022). As an example, consider a study of student-teacher relationships, in which raters rate the students' drawings about their relationship with their teachers. Students (level 1) are nested within teachers (level 2), but raters are crossed with students and teachers (Zee et al., 2020). As another example, consider the study of daily motivational processes in couples in which repeated measurements (level 1) are nested within a participant (level 2), and each pair of participants is nested within a couple (level 3) (Zygar et al., 2018).

When dealing with multilevel data, the standard literature on generalizability theory offers little guidance to determine reliability in such cases. Given the existence of the various levels, it is natural to consider the reliability at each level in the hierarchy. Consider again the example of student-teacher relationships: One may wonder how reliable the raters' scores of individual students are (reliability at the level of students), but also how reliable the ratings at the level of the teachers (i.e., clusters) are. Unfortunately, the reliability coefficients in generalizability theory or other approaches are often not specifically defined at these different levels. As a result, there may be a confounding of cluster-level variances and person-level variances, leading to biased estimates of reliability (Geldhof et al., 2014; ten Hove et al., 2022).

Furthermore, despite the importance of having reliable measurement procedures, research on reliability in the presence of multilevel data structures is limited, although there are a few exceptions (e.g., Geldhof et al., 2014; Lai, 2021; Schönbrodt et al., 2022; ten Hove et al., 2022). While Geldhof et al. (2014) and Lai (2021) work in a latent variable framework, the approach by ten Hove et al. (2022) and Schönbrodt et al. (2022) is rooted in the linear mixed model framework, and that is the perspective we will also take in this paper. Both ten Hove et al. (2022) and Schönbrodt et al. (2022) decomposed the variance of the multilevel data based on generalizability theory and independently defined reliability coefficients at the various levels of the data.

Despite their commonalities, ten Hove et al. (2022) and Schönbrodt et al. (2022)

consider very different designs. On the one hand, ten Hove et al. (2022) is concerned with interrater reliability when ratees are nested within clusters and raters are nested within or crossed with clusters. They defined cluster-level and person-level (interrater) reliability in terms of intraclass correlation (ICC) for three types of multilevel observational designs. Schönbrodt et al. (2022), on the other hand, studied dyadic intensive longitudinal designs, where both participants (within-cluster level/between-person level) of a couple (between-cluster level) are assessed multiple times each day across several days (within-person level). Starting from an existing reliability model (Cranford et al., 2006; Shrout & Lane, 2012a) with additional temporal level and clustering structure, Schönbrodt et al. (2022) proposed to assess reliability for the between-cluster (i.e., cluster-level), within-cluster (i.e., person-level), and within-person levels for two different temporal scales (e.g., day-by-day, moment-by-moment).

Because of these widely different designs, it is not straightforward to judge the (dis)similarities of both approaches. However, we will demonstrate below that if we compare the approaches of ten Hove et al. (2022) and Schönbrodt et al. (2022) on a set of common designs, they agree in some situations but not in all. In situations where the approaches disagree, the problem arises because different variance components are being used as error sources (especially when defining cluster-level reliability).

Given the discrepancy between the approaches of ten Hove et al. (2022) and Schönbrodt et al. (2022), there is a need for a standard method that results in proper and standardized definitions of reliability in multilevel designs. First, we reiterate the well-known result that reliability can be obtained as the (population) correlation between repeated measurements under some assumptions in the classical test theory. Next, we show the results of Vangeneugden et al. (2005) and Molenberghs et al. (2007), who interpret reliability coefficients from generalizability theory (generalizability or dependability coefficient), as a (conditional) correlation. Using the correlation approach may provide us with a principled approach for defining reliability coefficients in multilevel designs, being applicable to a wide variety of designs. To the best of our

knowledge, the correlation approach has not been used to define level-specific reliability for multilevel data. Therefore, we extend this method to determine reliability for each hierarchical level and compare it with the approaches of ten Hove et al. (2022) and Schönbrodt et al. (2022).

The remainder of this article is divided into four sections. The next section Methodology briefly introduces the concepts of reliability based on classical test theory, generalizability theory, and the correlation approach. In the section Reliability for Multilevel Data Structures, we introduce the person-level and cluster-level reliability as defined by Schönbrodt et al. (2022) and ten Hove et al. (2022). Then, we illustrate how to construct distinct reliability at different levels based on the correlation approach and compare our results with the previous results. Finally, in the last section, we present our conclusion and discuss the strengths, limitations, and future challenges of our approach.

Methodology

Evolution of the Concept of Reliability

The concept of reliability was initially defined in classical test theory (CTT; Lord et al., 1968; McDonald, 1999). Consider a countably infinite population $\mathcal{P}$ of experimental units (e.g., persons) indexed by i. CTT assumes that the observation of a person i, denoted as Y_i, can be a linear combination of a person's true score T_i and a random measurement error ϵ_i (i.e., $Y_i = T_i + \epsilon_i$). The true score is defined as the expected score over an infinite number of independent replications (i.e., $T_i = \mathrm{E}[Y_i]$), and the measurement error is assumed to have mean zero and to be uncorrelated among the replications and with the true score. The random variables Y_i, T_i, and ϵ_i are defined in a person space, and their variances σ_Y^2, σ_T^2, and σ_ϵ^2 are assumed to have finite values.

In CTT, the reliability of a measurement instrument with respect to the population $\mathcal{P}$ represents the scaled precision. It is defined as the squared correlation between observed and true scores:

$$\rho_{YT}^2 = \mathrm{Cor}^2(Y_i, T_i) = \frac{\sigma_T^2}{\sigma_Y^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\epsilon^2}. \tag{1}$$

The reliability coefficient is therefore the ratio of the true score variance to the observed total score variance, which is the sum of the true score variance and the error variance.

Since the true score and its variance are (usually) unavailable, psychometricians have come up with methods to estimate reliability in practice. Among the several options, we consider here the method based on parallel measurements. Let us consider $Y_{ij} = T_{ij} + \epsilon_{ij}$ the observation of a person i at replication j, where j is the index of a countably infinite or finite population of measurements $\mathcal{J}$. The replication measurements j and j' are parallel if they have the same true score (i.e., $\mathrm{E}[Y_{ij}] = \mathrm{E}[Y_{ij'}] = T_i$) and the equal variance of the measurement error (i.e., $\mathrm{Var}(\epsilon_{ij}) = \mathrm{Var}(\epsilon_{ij'})$) among the replicates for each person ($i \in \mathcal{P}$).

Under these parallel measurement assumptions, reliability as defined in Equation 1 is equivalent to the correlation between two parallel measurements, Y_{ij} and $Y_{ij'}$ across the population of persons:

$$\rho_{YY'} = \mathrm{Cor}(Y_{ij}, Y_{ij'} \mid j, j') = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\epsilon^2}. \tag{2}$$

In Equation 2, we explicitly condition on the two parallel measurements j and j' but not on the person i. The correlation is taken over the population of persons but the parallel measurements j and j' are fixed. The proof is listed in Appendix A. This demonstrates that the reliability coefficient in CTT is actually an *intraclass correlation coefficient* (ICC) under a one-way ANOVA model, that is, the correlation between two repeated measurements taken from the same person or class (Bartko, 1966; McGraw & Wong, 1996; Shrout & Fleiss, 1979).

Nevertheless, the conditions under which the CTT is applicable are rather limited. Because in many cases there are more identifiable sources of variability, such as different test administrators, item difficulties, or time effects, at play than only between-person true scores and error variability, the estimated reliability based on CTT may not be fully representative of the ratio of the person systematic variance over the unsystematic variance. For these situations, alternative approaches have been formulated.

Generalizability Theory

The most well-known framework generalizing CTT is generalizability theory (GT). While in CTT an observed measurement is decomposed into a true score and a single undifferentiated random error, in GT the observed score is considered as the sum of the person's score and multiple sources of measurement error that contribute to the variability of the observed value, such as raters, occasions, time, etc. Reliability coefficients are then defined by considering ratios of (sums of) variances (Brennan, 2001; Cronbach, 1972; Shavelson & Webb, 1991). The notion of reliability is reinterpreted as generalizability—one can generalize from one observation to a universe of observations.

In GT (and our own correlation approach, see below), two different reliability coefficients can be defined, depending on the kind of decision to be made with the measurement instrument: relative or absolute. The *generalizability coefficient* (GC), defined by Cronbach (1972), is used for relative decisions. It concerns the error arising from variance components associated with the rank ordering of the persons. While the *dependability coefficient* (a.k.a. index of dependability; DC), defined by Brennan and Kane (1977a, 1977b), is applied to absolute decisions, in which all variance components associated with the score, except the person factor, are considered as errors. Under a two-way random effects ANOVA model, these definitions coincide with the (relative) consistency ICC and absolute agreement ICC, respectively. Compared to ICCs as defined in McGraw and Wong (1996), the generalizability and dependability coefficients (Brennan, 2001; Cronbach, 1972; Shavelson & Webb, 1991) are discussed in more complicated random or mixed effects ANOVA models (but see also Demetrashvili et al., 2016; Wang et al., 2011).

In order to simplify the proliferation of terminology, we have decided to partition reliability into two categories: *relative reliability* (denoted as ρ_{rel}) and *absolute reliability* (denoted as ρ_{abs}). These terms can be associated with the generalizability coefficient and dependability coefficient in GT, respectively, as well as with the consistency ICC and absolute agreement ICC, respectively.

To illustrate the mechanics of GT, let us consider a three-way *Person* × *Rater* ×

Day design with the sample size $n_S \times n_R \times n_D$ sampled from populations denoted as $\mathcal{P}$, $\mathcal{J}$, and $\mathcal{M}$, respectively. Suppose that each person[1] $i = 1, \ldots, n_S$ is rated by all the raters $j = 1, \ldots, n_R$ on different days $m = 1, \ldots, n_D$. These three effects are completely crossed and are assumed to be random effects. For this design, the observed score is defined as follows:

$$Y_{ijm} = \mu + S_i + R_j + D_m + SR_{ij} + SD_{im} + RD_{jm} + \epsilon_{ijm}. \tag{3}$$

Thus, the observed score is decomposed as a grand mean μ, main effects for persons S_i, raters R_j and days D_m, two-way interaction effects $SR_{ij}, SD_{im}, RD_{jm}$, and a measurement error ϵ_{ijm}. Note that there is only one observation in each cell of the design, such that we cannot distinguish the three-way interaction term SRD_{ijm} from ϵ_{ijm}. Each effect and error is assumed to be mutually uncorrelated and distributed with a mean of zero and a positive, finite variance. The associated sources of variance are denoted by σ^2_S for the person effect, σ^2_R for the rater effect, σ^2_D for the day effect, and so on.

In GT, we start by estimating the universe score for a person i (the GT analog of the CTT true score), which is the expected score for that person across (in this case) the raters and days:

$$\tau_i = \mathrm{E}_{\mathcal{J},\mathcal{M}}[Y_{ijm}] = \mu + S_i,$$

where the subscripts $\mathcal{J}$ and $\mathcal{M}$ for the expectation make clear that the average is taken over all elements of the rater and day populations (i.e., $\mathcal{J}$ and $\mathcal{M}$, respectively).

On the one hand, if we are only interested in the consistency of the relative ranking of participants, the error we are making is a relative error:

$$\delta_i = (Y_{ijm} - \mathrm{E}_{\mathcal{P}}[Y_{ijm}]) - (\tau_i - \mathrm{E}_{\mathcal{P},\mathcal{J},\mathcal{M}}[Y_{ijm}]) = SR_{ij} + SD_{im} + \epsilon_{ijm}.$$

This relative error gives for a person i by how much the observed deviation from the average of all persons for that rater and day, that is $Y_{ijm} - \mathrm{E}_{\mathcal{P}}[Y_{ijm}]$, differs from the universe deviation (across all raters and days), that is $\tau_i - \mathrm{E}_{\mathcal{P},\mathcal{J},\mathcal{M}}[Y_{ijm}]$. A relative

[1] The quantities related to persons are denoted by S to be typographically more distinct than P from the other quantities such as R and D, while avoiding the term "subject" at the same time.

reliability coefficient (i.e., a generalizability coefficient; Cronbach, 1972) is defined as the ratio of the variance of the universe score to itself plus the variance of the relative error:

$$\rho_{\text{rel}} = \frac{\text{Var}(\tau_i)}{\text{Var}(\tau_i) + \text{Var}(\delta_i)} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_{SR}^2 + \sigma_{SD}^2 + \sigma_\epsilon^2}. \tag{4}$$

On the other hand, if we care about the absolute agreement between observed scores, then we should consider an absolute reliability coefficient (i.e., a dependability coefficient in GT; Brennan & Kane, 1977a). This coefficient considers the absolute errors of the universe score τ_i in the population of persons. The absolute error measures to which extent the observed score differs from the universe score:

$$\Delta_i = Y_{ijm} - \tau_i = R_j + D_m + SR_{ij} + SD_{im} + RD_{jm} + \epsilon_{ijm}.$$

The absolute reliability coefficient is constructed as the ratio of the variance of the universe score to itself plus the variance of the absolute error:

$$\rho_{\text{abs}} = \frac{\text{Var}(\tau_i)}{\text{Var}(\tau_i) + \text{Var}(\Delta_i)} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_R^2 + \sigma_D^2 + \sigma_{SR}^2 + \sigma_{SD}^2 + \sigma_{RD}^2 + \sigma_\epsilon^2}. \tag{5}$$

The denominator of the absolute reliability coefficient contains more terms because there are more sources of variance that lead to an absolute error. Therefore, it holds that $\rho_{\text{abs}} \leq \rho_{\text{rel}}$.

Error variance in Average Scores

Previously, we considered the reliability of single scores Y_{ijm}. However, in practice, it is more common to work with average scores to infer participants' attributes, as these are more reliable than single scores. For example, participants are evaluated by multiple raters and/or are followed over a number of days, and the obtained scores are then averaged across raters and/or time points. Continuing with the *Person* × *Rater* × *Day* design, suppose that a researcher wants to use the person's average observed score over raters and days: $\bar{Y}_{i\cdot\cdot} = \frac{1}{n_R n_D} \sum_{j=1}^{n_R} \sum_{m=1}^{n_D} Y_{ijm}$. This means that $\bar{Y}_{i\cdot\cdot}$ is used to make inferences about the person's universe score τ_i and that the relative error and absolute error for the average scores should be taken over raters and days. For example, the relative error then becomes: $(\bar{Y}_{i\cdot\cdot} - \text{E}_{\mathcal{P}}[\bar{Y}_{i\cdot\cdot}]) - (\tau_i - \text{E}_{\mathcal{P},\mathcal{J},\mathcal{M}}[\bar{Y}_{i\cdot\cdot}]) =$

$\frac{1}{n_R}\sum_{j=1}^{n_R} SR_{ij} + \frac{1}{n_D}\sum_{m=1}^{n_D} SD_{im} + \frac{1}{n_R n_D}\sum_{j=1}^{n_R}\sum_{m=1}^{n_D} \epsilon_{ijm}$. As a result, the relative reliability for the average score is given by:

$$\rho_{\text{rel}} = \frac{\sigma_S^2}{\sigma_S^2 + \frac{\sigma_{SR}^2}{n_R} + \frac{\sigma_{SD}^2}{n_D} + \frac{\sigma_\epsilon^2}{n_S n_D}}. \tag{6}$$

Similarly, the absolute reliability for the average score is:

$$\rho_{\text{abs}} = \frac{\sigma_S^2}{\sigma_S^2 + \frac{\sigma_R^2}{n_R} + \frac{\sigma_D^2}{n_D} + \frac{\sigma_{SR}^2}{n_R} + \frac{\sigma_{SD}^2}{n_D} + \frac{\sigma_{RD}^2}{n_R n_D} + \frac{\sigma_\epsilon^2}{n_R n_D}}. \tag{7}$$

The different various components that make up the relative and absolute error variances in Equations 6 and 7 are reduced proportionally to the number of raters and/or days. This implies that as the number of raters and/or days increases, the precision of the measurements also increases, as can be expected. Equations such as Equations 6 and 7 can then be used in a so-called Decision (D) study to optimize the number of levels of each factor (i.e., n_R and n_D) to reach a desired reliability level (Brennan, 2001; Shavelson & Webb, 1991; Strube, 2000).

While reliability for average scores is important, we will now shift our focus to the reliability for single scores. Reliability for single scores can be seen as a special case of the reliability for average scores when the numbers of raters and days are set to one. It is worth noting that we want to emphasize the structure of the reliability coefficient, which components are involved, and compare different coefficients defined by our correlation approach and others.

The Correlation Approach

Besides the ratio of variances construction outlined above, reliability coefficients can also be obtained through correlations. For simplicity, we focus on the reliability of single scores. The case of average scores will be dealt with in Appendix E. As already explained, the reliability coefficient for the CTT model (one-way ANOVA) can be obtained by correlating two parallel measurements as discussed in the previous section (see also Appendix A for the derivation). Similarly, the coefficients of relative reliability and absolute reliability can be expressed as correlation coefficients under a more complicated model, such as a two-way or three-way crossed random effects ANOVA model.

According to Vangeneugden et al. (2005) and Molenberghs et al. (2007), the relative reliability (i.e., the generalizability coefficient in Equation 4) for the *Person* × *Rater* × *Day* design is the same as the conditional correlation between two parallel observations:

$$\rho_{\text{rel}} = \text{Cor}(Y_{ijm}, Y_{ij'm'} \mid j, j', m, m'), \tag{8}$$

where we explicitly denote that we calculate the correlation across the population of persons but fix two raters (j and j') and two days (m and m'). The formal derivation of this correlation is presented in Appendix B. However, intuitively it can be seen that this is a relative reliability measure with respect to raters and days because the correlation is not affected by the constant deviation between two fixed raters or two fixed days. Stated differently, if rater j is, in general, more strict or lenient compared to rater j' (i.e., $R_j \neq R_{j'}$), this does not affect the relative reliability. An analogue reasoning holds for the day effects.

It is important to note that the conditioning terms used in the correlation are shorthand notations because we are conditioning on the random variables whose indices are included in those terms, as well as their combinations. For example, the conditioning terms in Equation 8 are j, j', m, and m'. Essentially, this implies that the random effects R_j, $R_{j'}$, D_m, $D_{m'}$, RD_{jm}, and $RD_{j'm'}$ are turned into fixed values and do not cause variation anymore when calculating the correlation.

An alternative interpretation of this conditional correlation is that we calculate the correlation between two pure person-related scores. The pure person-related score is the observed score without the effects that are not related to the person. In this case, the pure person-related score is defined as
$Y_{ijm} - R_j - D_m - RD_{jm} = \mu + S_i + SR_{ij} + SD_{im} + \epsilon_{ijm}$, is the observed score after subtracting the rater, day, and their interaction effects. The conditioning correlation, Equation 8, used for relative reliability can be interpreted as:

$$\rho_{\text{rel}} = \text{Cor}(Y_{ijm} - R_j - D_m - RD_{jm}, Y_{ij'm'} - R_{j'} - D_{m'} - RD_{j'm'}), \tag{9}$$

the correlation between the two pure person-related scores for the same person i from

the two raters j, j' at two days m, m', separately. It measures the degree of generalization for the pure person-related scores from one situation to another.

Conversely, absolute reliability (i.e., the dependability coefficient in Equation 5) is identical to a (marginal) correlation between two parallel observations:

$$\rho_{\text{abs}} = \text{Cor}(Y_{ijm}, Y_{ij'm'}). \tag{10}$$

The absolute reliability in Equation 10 does not condition on anything in the correlation, compared to Equation 8. It means that the correlation is taken across the population of persons, each measured by two different raters on two different days. Note that for every person, (j, j') and (m, m') may refer to different raters and days, and those differences will add uncertainty and reduce the correlation value. We assume throughout the paper that the same set of raters evaluate the persons. However, when considering relative reliability, two raters (j and j') are picked at random from this set and the correlation between the scores of these raters is calculated. For absolute reliability, for each person, two raters are selected from that same set (and hence, the selected pair may differ from person to person).

Multiple Types of Reliability

The correlation approach permits to define finer-grained reliability coefficients than GT when treating all facets as random. . Within the context of the *Person* $\times$ *Rater* $\times$ *Day* design, there are three scenarios to obtain parallel measurements from a person: considering scores (1) from rater j on day m and another rater j' on a different day m'; (2) from the same rater j but on different days m and m'; and (3) on the same day m from two different raters j and j'. The first scenario is the reliability as we have discussed before and, depending on what we condition on, it leads to the quantities ρ_{rel} and ρ_{abs}. This can be seen as an overall reliability, because raters examine how a person's score in a situation generalizes to another distinctive situation, which may have more than one source of variation at the same time. The second and third scenarios give rise to reliabilities that are simpler because there is only one aspect over which to generalize (either raters or days). The second scenario is commonly referred to as test-retest reliability, and the third scenario as interrater reliability.

Again, depending on what we condition on, both test-retest reliability and interrater reliability could also be discussed for the relative and absolute decisions leading to the following equations (Molenberghs et al., 2007; Vangeneugden et al., 2005):

$$\rho_{\text{test-retest, rel}} = \text{Cor}(Y_{ijm}, Y_{ijm'} \mid j, m, m'), \qquad \rho_{\text{test-retest, abs}} = \text{Cor}(Y_{ijm}, Y_{ijm'}),$$
$$\rho_{\text{interrater, rel}} = \text{Cor}(Y_{ijm}, Y_{ij'm} \mid j, j', m), \qquad \rho_{\text{interrater, abs}} = \text{Cor}(Y_{ijm}, Y_{ij'm}).$$

The test-retest reliability coefficient, which focuses exclusively on the consistency from one day to another, expresses how much we can generalize the observations of the person i from the rater j on two different days m, m'. On the other hand, the interrater reliability coefficient, which measures how consistent different raters are, expresses how much we can generalize the observations of the person i on the same day m but from different raters j, j'. Regarding the differentiation between relative and absolute reliability, the former still needs to condition on rater and day effects but the latter does not. It is possible to obtain test-retest and inter-rater reliability coefficients in GT by treating days and raters, respectively, as fixed. But as an inevitable consequence, the generalizability coefficients obtained in this way are only relevant for particular days or raters.

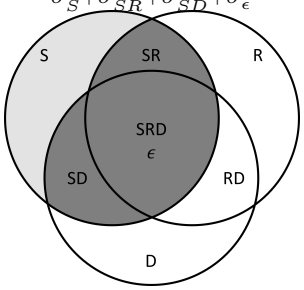
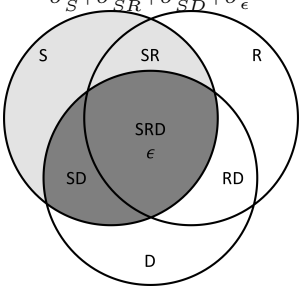
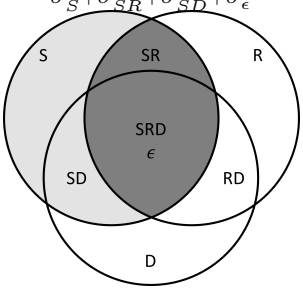
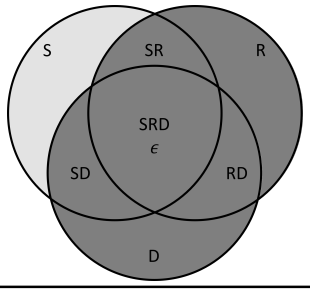
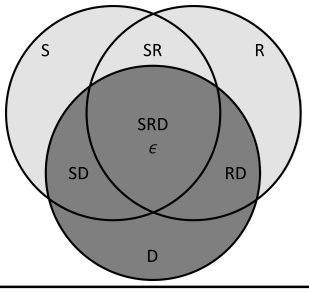
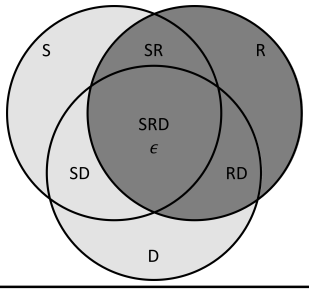
Comparison of the Approaches

Table 1 summarizes the relative and absolute reliability coefficients for single scores defined in GT and with the correlation approach in the three-way *Person* × *Rater* × *Day* design. The approaches coincide in defining the overall reliability but the correlation approach adds test-retest and interrater reliability. In addition, the reliabilities for average scores under the same design are presented in Table E1.

To support the interpretation and comparison, each coefficient is also represented by a Venn diagram illustrating the role of the variance components in the reliability coefficients. Each circle represents a variance component associated with a specific factor (including the person factor). The intersection between circles represents the variance of an interaction effect. The whole area of these circles is the total variance of an observed score. Light gray areas denote systematic variance (terms common to

Table 1

Reliability coefficients for single scores in a three-way ANOVA crossed design defined in GT and by the correlation approach

Reliability	Overall	Test-retest	Interrater
ρ_{rel}	$\text{Cor}(Y_{ijm}, Y_{ij'm'} \mid j, j', m, m') = \frac{\sigma^2_S}{\sigma^2_S+\sigma^2_{SR}+\sigma^2_{SD}+\sigma^2_\epsilon}$ 	$\text{Cor}(Y_{ijm}, Y_{ijm'} \mid j, m, m') = \frac{\sigma^2_S+\sigma^2_{SR}}{\sigma^2_S+\sigma^2_{SR}+\sigma^2_{SD}+\sigma^2_\epsilon}$ 	$\text{Cor}(Y_{ijm}, Y_{ij'm} \mid j, j', m) = \frac{\sigma^2_S+\sigma^2_{SD}}{\sigma^2_S+\sigma^2_{SR}+\sigma^2_{SD}+\sigma^2_\epsilon}$ 
ρ_{abs}	$\text{Cor}(Y_{ijm}, Y_{ij'm'}) = \frac{\sigma^2_S}{\left(\begin{array}{r}\sigma^2_S+\sigma^2_R+\sigma^2_D+\sigma^2_{SR}+\\ \sigma^2_{SD}+\sigma^2_{RD}+\sigma^2_\epsilon\end{array}\right)}$ 	$\text{Cor}(Y_{ijm}, Y_{ijm'}) = \frac{\sigma^2_S+\sigma^2_R+\sigma^2_{SR}}{\left(\begin{array}{r}\sigma^2_S+\sigma^2_R+\sigma^2_D+\sigma^2_{SR}+\\ \sigma^2_{SD}+\sigma^2_{RD}+\sigma^2_\epsilon\end{array}\right)}$ 	$\text{Cor}(Y_{ijm}, Y_{ij'm}) = \frac{\sigma^2_S+\sigma^2_D+\sigma^2_{SD}}{\left(\begin{array}{r}\sigma^2_S+\sigma^2_R+\sigma^2_D+\sigma^2_{SR}+\\ \sigma^2_{SD}+\sigma^2_{RD}+\sigma^2_\epsilon\end{array}\right)}$ 

Note: The Venn diagram illustrates the total variance of the score, which is divided into distinct variance components. Components that exist in both the numerator and denominator of the reliability coefficient are represented in light gray, while those that only appear in the denominator are represented in dark gray.

the numerator and denominator in each reliability coefficient). Dark gray areas refer to terms included in the unsystematic error variance (terms only included in the denominator).

Venn diagrams are a crucial visual aid that enables us to quickly distinguish between different reliability coefficients. For example, from Table 1 it can be seen that

the relative reliability coefficients merely involves the circle belonging to the person because these relative coefficients only consider sources of variance that influence the order of the persons. In contrast, the absolute reliability coefficients contain all the random factor areas because each of them contributes to the variation of the scores. We refer the reader to Cardinet et al. (2009) and Shavelson and Webb (1991) for Venn diagram representations of other ANOVA models.

Comparing the GT and the correlation approach, the latter has several advantages compared to the former. Firstly, various types of reliability coefficients can be defined by similar correlation formulas. Interrater, test-retest, or overall reliability can be determined from the correlation of two observed scores that vary by the raters, days, or both raters and days, respectively. As for the reliability of relative or absolute decisions, it is decided by whether the rater and day effects are fixed or remain random. Secondly, the correlation is easy to interpret. The correlation reflects the similarity of scores for the same person under various conditions. Furthermore, the reliability coefficient obtained by calculating the correlation under ANOVA models can be expressed as a ratio of variances (see Table 1). It provides insight into which sources of variance contribute to the systematic variance (the terms common to the numerator and the denominator of the ratio), and which ones serve as the unsystematic error variance (the terms only included in the denominator).

In this section, the definition of reliability has been reviewed from three different but strongly related perspectives: CTT, GT, and the correlation approach. The idea of calculating the correlation between two parallel measurements originally arose in the context of CTT (where the model simply separates the person's true score and measurement error from the observed score), but it also turned out suitable for defining relative and absolute reliability under the GT model (thereby recognizing multiple sources of variance in the observed score). In the next section, we will explore how the correlation approach can establish reliability for the nested data structure which had been discussed in two recent papers (Schönbrodt et al., 2022; ten Hove et al., 2022) using different approaches.

Reliability for Multilevel Data Structures

This section extends the reliability indices introduced earlier to account for multilevel or hierarchical structures, where observed units are nested within higher-level groups. This kind of data is commonly encountered in observational studies. For example, in education, exam scores do not only depend on the student's ability but also on classroom properties, such as the teacher's teaching experience. In the domain of medical evaluation, the postoperative performance of patients (nested in hospitals) needs to be assessed by several experienced doctors, but both patient characteristics (e.g., physical condition) as well as higher-level hospital characteristics (e.g., quality of medical personnel) and aspects specific to doctors (e.g., years of experience) may play a role in the generation of the final assessment score. To examine the consistency or agreement for the replicated measurements (e.g., the exam results on multiple occasions or the scores given by a group of juries) at different hierarchical levels, such as the person level (e.g., students or patients) and the cluster level (e.g., classrooms or hospitals) in the above examples, it is inevitable to define level-specific reliability for multilevel data structures.

Recently, ten Hove et al. (2022) and Schönbrodt et al. (2022) introduced level-specific reliability coefficients with the help of GT, which decomposes different sources of variance from the nested, hierarchical data. Their reliability coefficients possess a similar ratio structure to what we have discussed in the previous section:

$$\rho_{\text{decision}}^{\text{target level}} = \frac{\sigma^2_{\text{target level}}}{\sigma^2_{\text{target level}} + \sigma^2_{\text{errors}}}, \tag{11}$$

where $\sigma^2_{\text{target level}}$ refers to the isolated systematic variance at the target level (e.g., the person level or the cluster level) and σ^2_{errors} is the unsystematic error variance, which can be either the variance of the relative error, denoted as $\sigma^2_{\text{rel.errors}}$, or of the absolute error, denoted as $\sigma^2_{\text{abs.errors}}$. The subscript of ρ is "rel" or "abs", depending on whether relative or absolute errors are considered, respectively. Depending on the target level, the usage of the measurement instrument, and the hierarchical study design, different sources of variance obtained within the GT are included in the numerator and denominator.

The designs considered by ten Hove et al. (2022) and Schönbrodt et al. (2022) to

illustrate their coefficients differ markedly, and this makes it difficult to compare the similarities and possible differences between their approaches. Therefore, we will apply their methods, as well as our correlation approach, to a series of common designs. This way, possible differences between the approaches can be studied.

The three specific designs we consider are the following: (1) persons nested in clusters and crossed with raters, (2) persons nested in clusters with raters nested in both persons and clusters, and (3) persons nested in clusters and crossed with raters and days. In the first two designs, we will only study the (interrater) reliability at different hierarchical levels, but in the final design, the overall, test-retest, and inter-rater reliability will be investigated (because there are two kinds of parallel replicates (raters and days) with respect to the same person). Note that for simplicity we will focus the main derivations on the reliability coefficients of single scores, but that we will also present the formulas for the average scores in Appendix E.

Design 1: Persons nested in clusters and crossed with raters

Consider a design with persons nested in clusters and all crossed with raters (i.e., *Person(Cluster)* $\times$ *Rater* or Case 1 in ten Hove et al., 2022). In this design, $Y_{i(k)j}$ is a score of person $i(k) = 1, \dots, n_S$, nested within cluster $k = 1, \dots, n_C$, obtained from rater $j = 1, \dots, n_R$. The notation $a(b)$ denotes a factor at level a nested within another factor at level b. For the sake of simplicity, we assume a balanced design where all clusters have the same number of persons and each rater only gives one rating for each person. Based on GT, $Y_{i(k)j}$ can be decomposed as:

$$Y_{i(k)j} = \mu + C_k + S(C)_{i(k)} + R_j + CR_{kj} + \epsilon_{i(k)j}, \tag{12}$$

with the grand mean μ, three main effects–the cluster effect C_k, the person effect $S(C)_{i(k)}$, and the rater effect R_j, the interaction effect of cluster and rater CR_{kj}, and the measurement error $\epsilon_{i(k)j}$. Note that the person and rater interaction effect $S(C)R_{i(k)j}$ and the measurement error can only be distinguished when every person is observed more than once by every rater. As this is not the case in most applications, we do not add this term to Equation 12. Each effect and error is assumed to be mutually uncorrelated and distributed with zero mean and a positive, finite variance. Hence, the

total variance decomposition corresponding to Equation 12 is given by

$$\text{Var}(Y_{i(k)j}) = \sigma^2_C + \sigma^2_{S(C)} + \sigma^2_R + \sigma^2_{CR} + \sigma^2_\epsilon.$$

The corresponding reliability coefficients defined by ten Hove et al. (2022), Schönbrodt et al. (2022), and with the correlation approach are given at the person (S) and cluster levels (C) for relative (rel) and absolute (abs) decisions in Table 3. These four types of reliability are denoted as $\rho^{\text{S}}_{\text{rel}}$, $\rho^{\text{S}}_{\text{abs}}$, $\rho^{\text{C}}_{\text{rel}}$, and $\rho^{\text{C}}_{\text{abs}}$. Note that Schönbrodt et al. (2022) did not consider relative decisions and only discussed reliability for average scores. For the sake of comparison, we demonstrate reliability for single scores to highlight the discrepancy between these approaches. In each cell of the table, we present each reliability coefficient with a Venn diagram, where the light gray and dark gray areas indicate $\sigma^2_{\text{target level}}$ and σ^2_{errors} corresponding to Equation 11, respectively. In addition, we also present our correlation formula with respect to the type of reliability in the correlation approach column. Derivations under the correlation approach can be found in Appendix C.

The Approaches by ten Hove et al. (2022) and Schönbrodt et al. (2022)

Let us first discuss the reliability defined by ten Hove et al. (2022) in the second column of Table 3 and by Schönbrodt et al. (2022) in the third column.

Person-Level Reliability. At the person level (see the first and second rows in Table 3), within GT, the target variance $\sigma^2_{\text{target level}}$ is always the person variance $\sigma^2_{S(C)}$ because it is the only source of systematic variance at that level. The relative error variance $\sigma^2_{\text{rel.errors}}$ does not include the rater variance σ^2_R while the absolute error variance $\sigma^2_{\text{abs.errors}}$ does. Error terms exclude variance components at the cluster level (i.e., σ^2_C and σ^2_{CR}) because scores of persons within a cluster are independent of cluster-related effects. Therefore, both ten Hove et al. (2022) and Schönbrodt et al. (2022) exclude the higher-level components (i.e., cluster-related variances) in the lower-level reliability (i.e., person-level reliability).

Cluster-Level Reliability. At the cluster level (see third and fourth rows in Table 3), when defining reliability within GT, the target variance $\sigma^2_{\text{target level}}$ is the cluster variance σ^2_C. Just like for the person-level case, Schönbrodt et al. (2022) do not define a

cluster-level relative reliability, while ten Hove et al. (2022) do. Hence, a direct comparison is not possible.

When turning to the cluster-level absolute reliability, it can be seen that there is a significant difference in the definition of the absolute error variance between Schönbrodt et al. (2022) and ten Hove et al. (2022). In Schönbrodt et al. (2022), $\sigma^2_{\text{abs.errors}}$ includes all relevant sources of variances (except the target variance) that are lower than or equal to the focal level (see Equation 4[2]in their paper). In ten Hove et al. (2022), however, $\sigma^2_{\text{abs.errors}}$ merely includes the rater variance and the rater by cluster interaction variance but excludes any person-related variances, i.e., $\sigma^2_{S(C)}$ and σ^2_ϵ (see Equation 13 in their paper). Originally, ten Hove et al. (2022) defined the cluster-level absolute reliability for average scores as $\text{ICC}_C(A, n_R) = \frac{\sigma^2_C}{\sigma^2_C + (\sigma^2_R + \sigma^2_{CR})/n_R}$, with the subscript C denoting the cluster-level, and A being for agreement reliability of an average score based on n_R raters. In our case, $n_R = 1$ for single scores, so that $\text{ICC}_C(A, 1)$ is actually the cluster-level absolute reliability in this table. Ten Hove et al. (2022) claim that $\text{ICC}_C(A, n_R)$:

> "... does not contain any subject-related (i.e., person-related) variance components because the ordering of clusters across raters is independent of subject-related effects" (ten Hove et al., 2022, p. 655).

However, this rule may be undesirable because the cluster-level reliability coefficients cannot be defined for some nested data. More specifically, if we obey this rule directly, in some designs, the cluster-level reliability turns out be meaningless. For example, in the $Rater(Person(Cluster))$ design, the denominator of the cluster-level reliability reduces to σ^2_C because the terms related to the persons ($\sigma^2_{S(C)}$ and σ^2_ϵ) are excluded, resulting in a cluster-level reliability coefficient of $\sigma^2_C/\sigma^2_C = 1$ (see the examples in Tables 4 and D1 in this paper or Table 3 in ten Hove et al., 2022).

[2] In order to make the comparison with the result presented here, we ignored the variances of main and interaction effect related to the fixed factors in both the numerator and denominator and focus on the reliability for a single measurement, so that the number of all factor levels used in there formula, that is $j, k, \& \ l$, equal one.

The Correlation Approach

Let us move on to the results obtained by the correlation approach in the fourth and fifth column of Table 3. Here, we list the universal correlation formula for the definition of person- and two viewpoints of cluster-level reliability.

Person-Level Reliability. Within the correlation approach, we define the person-level relative reliability as the correlation between the scores given by two fixed raters j and j' on the same person $i(k)$ randomly selected from a specific cluster k:

$$\rho^{\text{S}}_{\text{rel}} = \text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j'). \tag{13}$$

The person $i(k)$ remains random but a specific cluster k is fixed in the above equation. This leads to the result applicable to the whole person population and eliminates variation between different clusters when considering the person-level reliability. Furthermore, conditioning on the raters leads to a relative measure because rater j can be more severe or less severe than rater j' by a constant without affecting reliability.

On the other hand, person-level absolute reliability is obtained by only conditioning on a specific cluster:

$$\rho^{\text{S}}_{\text{abs}} = \text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k). \tag{14}$$

It is clear that such a correlation measures absolute agreement because differences in severity between raters will lower the correlation.

Cluster-Level Reliability. The cluster-level reliability can be viewed from two different perspectives. The key difference between the two perspectives is the role of the person variance $\sigma^2_{S(C)}$, leading to what we call the manifest cluster-level reliability and latent cluster-level reliability.

Manifest cluster-level reliability. The first perspective is to consider the person variance $\sigma^2_{S(C)}$ as one of the sources of unsystematic error variance σ^2_{errors}. To better understand this viewpoint, we first talk about the reliability for average scores (Appendix E will discuss in detail) and then move on to the case of single scores. For example, suppose that a researcher collects well-being scores from students. For simplicity, we will ignore the school level and focus on the nesting of students in classes.

The researcher wants to evaluate the well-being scores at the class level. Therefore, she samples several classes and a number of students in each class. The students are asked to complete a short well-being questionnaire consisting of five items. Thus, in this case, the students (person-level) are nested within the classes (cluster-level), and the items (or raters) are crossed with the students and the classes. After the questionnaire, the researcher calculates each class' well-being score by averaging all students within that class and their responses to the five questionnaire items. Then, she ranks all the classes according to their well-being scores. However, at the same time, the researcher is wondering how reliable well-being scores at the class level are. In other words, she wonders about the extent to which the ranking of the classes would change if she samples different students from the classes and uses five different (parallel) items from the well-being questionnaire the next time.

In this particular example, it makes sense to include the person variance $\sigma^2_{S(C)}$ as a source of unsystematic or error variance. If student variability in well-being scores within classes is very large, then it is very difficult to say something reliably about between-class differences in well-being. Returning to the case of reliability for single scores, we are assessing in this example the consistency between two observed scores for two different persons $i(k)$ and $i'(k)$ under the same cluster k but given by two raters j and j'. We named this reliability *the manifest cluster-level reliability* because we check the relation of the cluster directly from the observed single (or average) score. Based on the correlation approach, we can further distinguish between manifest cluster-level reliability for relative or absolute decisions, as before. The manifest cluster-level relative reliability coefficient is defined as:

$$\rho^{\text{C, man}}_{\text{rel}} = \text{Cor}(Y_{i(k)j}, Y_{i'(k)j'} \mid j, j') = \frac{\sigma^2_C}{\sigma^2_C + \sigma^2_{S(C)} + \sigma^2_{CR} + \sigma^2_\epsilon}. \tag{15}$$

The superscript "man" after C represents the manifest cluster-level reliability. Likewise, the manifest cluster-level absolute reliability coefficient is defined as:

$$\rho^{\text{C, man}}_{\text{abs}} = \text{Cor}(Y_{i(k)j}, Y_{i'(k)j'}) = \frac{\sigma^2_C}{\sigma^2_C + \sigma^2_{S(C)} + \sigma^2_R + \sigma^2_{CR} + \sigma^2_\epsilon}, \tag{16}$$

It is worth noting that Equation 16 corresponds to the cluster-level absolute reliability

defined by Schönbrodt et al. (2022) (see Table 3).

Latent Cluster-Level Reliability. Another perspective is to exclude the person variance $\sigma^2_{S(C)}$ from the source of unsystematic error variance σ^2_{errors}. Again, we will illustrate a situation with an example. Suppose again that there are multiple classes (ignoring the school level), and, in each class, only one teacher is responsible for the students. A researcher is interested in the student-teacher interactive quality at the teacher level, so he randomly selects classes and students from these classes and invites several educational experts to provide a continuous interactive score for each student-teacher pair by watching a recording of their classroom interactions. Again, this is a similar design as before, in which the students (person-level) are nested within the teachers (cluster-level), and the educational experts are crossed with both the students and the teachers. To evaluate the interactive quality at the teacher level, the researcher averages student-teacher interactive scores across the students and educational experts to determine each teacher's interactive score. In this situation, the within-teacher variability (i.e., the variability across students) should not influence the reliability of the teacher scores. The students are necessary when evaluating the student-teacher interactive interactions, but whether there is more or less variability among the students should not play a role in the reliability of the scores at the teacher level. Partialling the within-teacher student variability out, can be understood as follows: We imagine that the same students could hypothetically participate in other classes and interact with different teachers. Because the sample of students does not change, the variability due to students disappears. Therefore, under this situation, teacher-level (cluster-level) reliability means measuring how consistent the teacher's interactive scores, which eliminate the student variation, would be if the study were repeated on the same specific students (who can show up in different classes) and randomly selected experts.

To reflect the above perspective in the correlation approach, the cluster-level reliability can be defined as the correlation between two (average) scores (e.g., provided by two raters) of the same person showing up in a randomly selected cluster. Because in this reliability we only care about the (randomly varying) cluster scores but without the

person effect, which can not be directly observed, we name this reliability the *latent cluster-level reliability*. Thus, in the case of reliability for single scores at the cluster level, we define the latent cluster-level reliability for relative decision (i.e., latent cluster-level relative reliability) as the correlation between two observed scores for the same person $i(k)$ from a randomly selected cluster k but given by two different raters j and j':

$$\rho_{\text{rel}}^{\text{C, lat}} = \text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j') = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{CR}^2 + \sigma_\epsilon^2}, \tag{17}$$

where it conditions on two specific raters j and j' and on a specific individual $i(k)$ in cluster k. The superscript "lat" after C represents the latent cluster-level reliability. It should be emphasized that, though the notation for the person who comes from a cluster is $i(k)$, the above equation does merely condition on the person but not on the cluster, which still remains random. Here, we conceptually act as if a specific person could exist in different clusters, so no matter what cluster is randomly selected, we still could fix the same person in the hypothetical replication. On the one side, we condition on that person $i(k)$ to eliminate the person variance in the observed score. On the other side, we keep the clusters random to make the result applicable to the whole cluster population. Likewise, we also fix the same pair of raters (j, j') to get the relative measurement.

Similarly, latent cluster-level absolute reliability is again obtained by mere conditioning on the person $i(k)$ (but not the cluster):

$$\rho_{\text{abs}}^{\text{C, lat}} = \text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k)) = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_R^2 + \sigma_{CR}^2 + \sigma_\epsilon^2}. \tag{18}$$

By conditioning on the person effect, the correlation excludes the contribution of person variance in the denominator (compared to the manifest cluster-level reliability coefficients in Equations 15 and 16) and focuses on the cluster-level variance. Without conditioning on raters, the coefficient measures absolute agreement because differences between raters would impact the variability of the observed values.

An Alternative Perspective. Applying the correlation approach to obtain latent cluster-level reliability starts from conditioning on a person $i(k)$, but not on the

cluster. That may sound counterintuitive or possibly contrived. For this reason, it may be helpful to consider another perspective on the conditional correlations presented thus far. Hence, an alternative way to understand the conditional correlations in Equations 13-14, 17-18 is to consider as reliability measure the correlation between two pure level-specific scores. As discussed before (see Equation 9), a pure level-specific score, such as pure person-level score and pure cluster-level score, is an observed score from which the unrelated fixed effect(s) have been removed. Therefore, the pure level-specific score is a latent score, which cannot be directly observed from the measurement instrument.

For instance, the person-level relative reliability coefficient (Equation 13) and absolute reliability coefficient (Equation 14) can be calculated by:

$$\rho^{\mathrm{S}}_{\mathrm{rel}} = \mathrm{Cor}(Y_{i(k)j} - C_k - R_j - CR_{kj}, Y_{i(k)j'} - C_k - R_{j'} - CR_{kj'}), \tag{19}$$

$$\rho^{\mathrm{S}}_{\mathrm{abs}} = \mathrm{Cor}(Y_{i(k)j} - C_k, Y_{i(k)j'} - C_k), \tag{20}$$

where $Y_{i(k)j} - C_k - R_j - CR_{kj} = \mu + S(C)_{i(k)} + \epsilon_{i(k)j}$ is a pure person-level score for the relative decision and $Y_{i(k)j} - C_k = \mu + S(C)_{i(k)} + R_j + CR_{kj} + \epsilon_{i(k)j}$ for the absolute decision. On the other side, the cluster-level relative reliability coefficient (Equation 17) and absolute reliability coefficient (Equation 18) can be written as:

$$\rho^{\mathrm{C,\ lat}}_{\mathrm{rel}} = \mathrm{Cor}(Y_{i(k)j} - S(C)_{ik} - R_j, Y_{i(k)j'} - S(C)_{ik} - R_{j'}), \tag{21}$$

$$\rho^{\mathrm{C,\ lat}}_{\mathrm{abs}} = \mathrm{Cor}(Y_{i(k)j} - S(C)_{ik}, Y_{i(k)j'} - S(C)_{ik}), \tag{22}$$

where $Y_{i(k)j} - S(C)_{ik} - R_j = \mu + C_k + \epsilon_{i(k)j}$ is a pure cluster-level score for the relative decision and $Y_{i(k)j} - S(C)_{ik} - R_j = \mu + C_k + \epsilon_{i(k)j}$ for the absolute decision.

The pure person-level (or cluster-level) score for the relative decision is the observed score removing the cluster (or person) effect and the rater effect (and their interaction if it exists). In other words, the pure person-level (or cluster-level) score merely includes the grand mean, the person effect (or the cluster effect), and measurement error. Then the person-level (or cluster-level) relative reliability is defined as the correlation between two parallel pure person-level (or cluster-level) scores. The perspective of considering pure level-specific scores and correlating them is actually

similar to the reliability definition in the CTT under the parallel assumption (see Equation 2). For example, we may re-parameterize $\mu + S(C)_{i(k)}$ as T_i. The pure level-specific scores for the absolute decision can be interpreted likewise. The only difference is that only the rater effect is subtracted in that case.

Summary. In short, we define six kinds of level-specific reliability—person-level relative reliability, person-level absolute reliability, manifest cluster-level relative reliability, manifest cluster-level absolute reliability, latent cluster-level relative reliability, and latent cluster-level absolute reliability—based on (conditional) correlations. To clarify, Table 2 collects Equations 13-18 together, highlighting the similarities and differences between them. All relative reliability coefficients are defined as the correlation between two observed scores given by two different raters (or other types of replicates) j, j'. However, the person-level reliability and latent cluster-level reliability focus on the same person $i(k)$ from cluster k, but manifest cluster-level reliability targets to the two different persons $i(k), i'(k)$ from the same cluster k.

In fact, there are two rules to decide which effect should be conditioned on. First, relative reliability requires the correlation to be conditioned on two rater values, while absolute reliability does not. Second, person-level reliability is conditioned on the same cluster, latent cluster-level reliability is conditioned on the same person, while manifest cluster-level reliability does not need to be conditioned on persons. These rules can also be applied to other nested designs, as we will see in the next paragraphs.

Note that the correlation approach is also applicable to person- and cluster-level reliability of average scores. The correlation formula is similar to the single score cases, as shown in Table 2, except that the correlation between two average scores is measured instead of between two single scores. The average score is either taken over raters for each person for person-level reliability or over raters and persons for each cluster for cluster-level reliability. Therefore, the conditioning terms should include all involved targets used in the average scores. For example, in relative reliability, the correlation needs to be conditioned on two groups of raters, from which the raters will be sampled at the two measurement occasions, instead of only two different raters. This will lead to

Table 2

*The Correlation Representation for Person-Level and Cluster-Level Reliabilities**

	Person-level	Manifest Cluster-level	Latent Cluster-level
Relative reliability	$\text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j')$	$\text{Cor}(Y_{i(k)j}, Y_{i'(k)j'} \mid j, j')$	$\text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j')$
Absolute reliability	$\text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k)$	$\text{Cor}(Y_{i(k)j}, Y_{i'(k)j'})$	$\text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k))$

*we condition on the person in the latent cluster-level reliability while this is not the case in the manifest cluster-level reliability

the division of some variance component(s) by the number of replicate items, such as rater or person (see Appendix E for the details). Therefore, the reliability for a single score can be viewed as a special case of the reliability for an average score, when the number of replicated items is set to one.

Comparisons between the various approaches

For the *Person(Cluster)* × *Rater* design, the person-level relative reliability coefficients are the same in both ten Hove et al.'s (2022) and the correlation approaches. However, regarding the person-level absolute reliability, the correlation approach led to a slightly different result compared to the results of Schönbrodt et al. (2022) and ten Hove et al. (2022). From Table 3, it can be observed that $\rho^{\text{S}}_{\text{abs}}$ defined by the correlation approach does include an extra term, the variance of the cluster and rater interaction σ^2_{CR}. To separate each reliability defined by ten Hove et al. (2022), Schönbrodt et al. (2022), and our correlation approach, we subscript ρ^S (and ρ^C) with "ten Hove", "Schönbrodt", or "cor", respectively. Thus, we have $\rho^{\text{S}}_{\text{abs, cor}} \leq \rho^{\text{S}}_{\text{abs,ten Hove}} = \rho^{\text{S}}_{\text{abs, Schönbrodt}}$.

The three approaches lead to some notable differences in reliability coefficients at the cluster level. Compared to Schönbrodt et al. (2022) and manifest correlation

approach (the fourth column), the latent correlation approach (the fifth column) excludes the person variance $\sigma^2_{S(C)}$ in the denominator, and ten Hove et al. (2022) excludes person variance $\sigma^2_{S(C)}$ as well as the error measurement variance σ^2_ϵ. As a result, we have the following two order relationships for the relative and absolute decisions: $\rho^{\text{C, man}}_{\text{rel, cor}} \leq \rho^{\text{C, lat}}_{\text{rel, cor}} \leq \rho^{\text{C}}_{\text{rel, ten Hove}}$ and $\rho^{\text{C}}_{\text{abs, Schönbrodt}} = \rho^{\text{C, man}}_{\text{abs, cor}} \leq \rho^{\text{C, lat}}_{\text{abs, cor}} \leq \rho^{\text{C}}_{\text{abs, ten Hove}}$, respectively.

Three key differences arise between the correlation and the GT approaches. First, the term σ^2_{CR} appears in the denominator of the person-level absolute reliability $\rho^{\text{C}}_{\text{abs}}$ under the correlation approach. Although not intuitive at first sight, we think it is necessary to include σ^2_{CR} in the computation of the person-level absolute reliability. Given that the correlation approach defines a correlation, we may imagine a hypothetical table with two columns between which the correlation is computed (over an infinite number of rows). The rows of the table are the sampled persons, but given the conditioning information, the cluster from which they are sampled is kept fixed. The two columns refer to two raters. However, because there is no conditioning on specific raters, for each person a new pair of raters is sampled. Owing to the randomly selected raters, the $(CR)_{jk}$'s remain random as well (despite the fixing of k), which ultimately affects the scores being observed. As such, σ^2_{CR} contributes to the absolute reliability at the person level.

The second and third difference concern the inclusion of $\sigma^2_{S(C)}$ and σ^2_ϵ in cluster-level (relative and absolute) reliability. We have already mentioned whether including $\sigma^2_{S(C)}$ or not depends on what perspective (manifest vs. latent cluster scores) on cluster-level reliability one is taking. As for the σ^2_ϵ, we think it is necessary to include it in the denominator because $\epsilon_{i(k)j}$ is an unsystematic measurement error and is random all the time. Therefore, σ^2_ϵ distorts precision at all levels. Even though we condition on the person for cluster-level reliability, every measurement for the same person (even under identical conditions) still fluctuates because of the existence of the measurement error. In addition, ignoring the measurement error variance in cluster-level reliability can result in a lack of definability of the cluster-level reliability in

some nested designs σ^2_ϵ (e.g., see the next nested design example or Appendix D).

Design 2: Raters nested in persons nested in clusters

Let us now consider a design where persons are nested within clusters and raters are nested within both persons and clusters (i.e., *Rater(Person(Cluster))* or Case 3 in ten Hove et al., 2022). Each rater only rates a person belonging to a specific cluster once. For example, consider an experiment to evaluate the quality of products (i.e., the person level) in different factories (i.e., the cluster level). Each product is assessed by a number of raters, different for each product.

The corresponding model for a *Rater(Person(Cluster))* design is given by:

$$Y_{j(i(k))} = \mu + C_k + S(C)_{i(k)} + \epsilon_{j(i(k))}, \tag{23}$$

where μ is the grand mean, C_k is the cluster effect, $S(C)_{i(k)}$ is the person effect, and $\epsilon_{j(i(k))}$ is the measurement error. Each effect and error is assumed to be mutually uncorrelated and distributed with zero mean and a positive, finite variance. Thus, the variance of the observed score can be decomposed as $\text{Var}(Y_{j(i(k))}) = \sigma^2_C + \sigma^2_{S(C)} + \sigma^2_\epsilon$.

Reliability coefficients obtained with the three approaches are given in Table 4. Note that there is no difference between the relative and absolute reliabilities because the raters assess every person only once. Consequently, we can not separate the rater effect and the random measurement error. If we ignore the cluster structure, the so obtained reliability is virtually the same as the ICC(1) defined in a one-way ANOVA design (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

For the *Rater(Person(Cluster))* design, the person-level reliability coefficients ρ^{S} defined by the three approaches are identical; however, the cluster-level reliability coefficients ρ^{C} are quite different as can be seen in Table 4. Especially, ρ^{C} cannot be defined in ten Hove et al.'s (2022) approach because, under their principles, the cluster-level reliability coefficient should ignore any variance related to the persons ($\sigma^2_{S(C)}$ and σ^2_ϵ), thus $\rho^{\text{C}}_{\text{ten Hove}}$ is always equal to 1 and therefore meaningless.

ρ^{C} can nevertheless be defined with Schönbrodt et al.'s (2022) approach and our correlation approach. $\rho^{\text{C}}_{\text{Schönbrodt}}$ contains all variances in the denominator. It can also be derived by the correlation approach using the manifest cluster perspective $\rho^{\text{C, man}}_{\text{cor}}$.

However, $\rho_{\text{cor}}^{\text{C, let}}$ from the perspective of the latent cluster scores, $\sigma^2_{S(C)}$ is exluded from the denominator. Thus, the order relation between the cluster reliability coefficients is as follows: $\rho^{\text{C}}_{\text{Schönbrodt}} = \rho_{\text{cor}}^{\text{C, man}} \leq \rho_{\text{cor}}^{\text{C, let}} \leq \rho^{\text{C}}_{\text{ten Hove}} = 1$.

Design 3: Persons nested in clusters and crossed with raters and days

To give a more comprehensive comparison of the nested structure and fully crossed structure, we consider a more complicated design: *Person(Cluster)* × *Rater* × *Day*. It is actually a multilevel version of the *Person* × *Rater* × *Day* design presented earlier (see Table 1) because now we consider persons to be nested within different clusters. In addition, this design is also a simplified version of the design discussed in Schönbrodt et al. (2022) (where moment effects are additionally considered). Here, each person belonging to one of the clusters will receive scores from a set of raters on different days, but each rater only gives one score for each person on each day. So there is no replication in each condition. The corresponding model is given by

$$Y_{i(k)jm} = \mu + C_k + S(C)_{i(k)} + R_j + D_m + CR_{kj} + CD_{km} + CRD_{kjm} + S(C)R_{i(k)j} + S(C)D_{i(k)m} + \epsilon_{i(k)jm}, \quad (24)$$

where μ is the grand mean, C_k is the cluster effect, $S(C)_{i(k)}$ is the person effect, R_j is the rater effect, D_m is the day effect, and the remaining terms are the two-way or three-way interactions and measurement error $\epsilon_{i(k)jm}$. Each effect and error is assumed to be mutually uncorrelated and distributed with a mean of zero and a positive, finite variance, such as σ^2_C, $\sigma^2_{S(C)}$, and so on, over their population. The total variance is the sum of these effect variances,

$$\text{Var}(Y_{i(k)jm}) = \sigma^2_C + \sigma^2_{S(C)} + \sigma^2_R + \sigma^2_D + \sigma^2_{CR} + \sigma^2_{CD} + \sigma^2_{S(C)R} + \sigma^2_{S(C)D} + \sigma^2_\epsilon.$$

The rules that we have discussed in the correlation approach to define reliability are still feasible under this more complex nested design. For the previous two designs, we have focused on the interrater reliability since that allowed for a direct comparison with the approaches of ten Hove et al. (2022) and Schönbrodt et al. (2022). However, for the *Person(Cluster)* × *Rater* × *Day* design, we will not make a comparison with other approaches as it does not add anything new. Instead, we demonstrate how the

correlation approach can be used to derive various kinds of reliability by calculating a correlation between two parallel scores. As explained previously, depending on the definition of the parallel measurements, there are three types of reliability: (1) overall reliability (generalizing across raters and days); (2) test-retest reliability (generalizing only across days); and (3) interrater reliability (generalizing only across raters). Moreover, each reliability has a relative or an absolute version (determined by what is conditioned on: days and/or raters). Beyond that, reliability can be established at the person level (conditioned on a cluster), the manifest cluster score, or at the cluster level in the latent cluster score (conditioned on a person). Combining all choices results in 18 different reliability coefficients in Table 4. In this table, the three columns determine the type of reliability coefficient, and the six rows are the combinations of the target level and the kind of decision. To save space, each reliability coefficient is only represented by the coloring of the Venn diagrams: the light gray areas make up the numerator, and both light and dark gray areas make up the denominator. From the colored Venn diagrams, the ratio of variances formula can be derived easily.

The type of reliability (and thus the parallel measurements for which the correlation is considered) essentially determines which factor should be regarded as differentiating or as the source of error. As an example, take the person-level reliabilities (the first and second rows of Table 5). The person-day interaction $S(C)D$ is part of the target-level variance in the interrater reliability but for the overall and test-retest reliability, it is part of the unsystematic error variance. The same situation also arises when assessing cluster-level reliability, regardless of whether the decision is relative or absolute.

For the *Person(Cluster) × Rater × Day* design, all three types of reliability (overall, test-retest, and interrater) have a version at the person and two versions at the cluster level. Based on our reliability definition using the correlation approach, the person-level relative reliability coefficients (the first row in Table 5) merely concern the person circle. The manifest cluster-level relative reliability coefficients (the third row) take all the components in the cluster circle as well as the person circles. Nevertheless,

the latent cluster-level relative reliability coefficients (the fifth row) take the components in the cluster circle and the measurement error variance σ^2_ϵ into account.

A counterintuitive result is that the person-level absolute reliability coefficients (the second row) contain the cluster-related interaction variances (i.e., $\sigma^2_{CR}, \sigma^2_{CD}, \sigma^2_{CRD}$). But, imagine once more we are calculating a correlation between two columns of scores. The rows are formed by randomly sampled participants from a given cluster. The first column is a random rater assessing the participant on a randomly sampled day and the second column is another random rater assessing the same participant on a random day. Despite the fact that the cluster is predetermined for the person-level reliability, CR_{kj}, CD_{km}, and CRD_{kjm} still influence the value of the observed scores if we selected different pairs of raters and days every time. Thus, these interaction variances should also be considered in the reliability coefficients. It is similar to the case that we discussed in the *Person(Cluster)* $\times$ *Rater* design where the rater is also crossed with the cluster, the cluster and rater interaction variance, σ^2_{CR} is one of the components in the person-level absolute reliability coefficient. A similar reasoning is applicable for the latent cluster-level absolute reliability coefficients (the last row) that contain the person-related interaction variances (i.e., $\sigma^2_{S(C)D}, \sigma^2_{S(C)R}$) and the measurement error variance.

Conclusions and Discussion

In this paper, we have investigated the various faces of reliability in multilevel data. First, reliability can be defined at the person or cluster level, depending on the hierarchical level of interest. Second, relative and absolute reliability can be evaluated, depending on whether the measurement instrument will only be used for ranking purposes or if one is interested in the measure itself given by the measurement instrument. Two approaches (ten Hove et al. (2022) and Schönbrodt et al. (2022)) based on generalizability theory (Brennan, 2001; Cronbach, 1972; Shavelson et al., 1989) were recently identified in the literature to evaluate these reliabilities.

When applied to some common designs with a hierarchical structure, we found remarkable differences between the reliability coefficients as obtained by applying the

rules from ten Hove et al. (2022) and Schönbrodt et al. (2022). The main difference is found in the cluster-level reliability. The approach by ten Hove et al. (2022) excludes variances of person-related components and therefore also measurement error variance in the reliability coefficient, while Schönbrodt et al.'s (2022) approach includes them.

Similar differences in definitions have been found in other places in the literature. For example, in the context of multilevel confirmatory factor analysis, both Geldhof et al. (2014) and Lai (2021) defined within- and between-level composite reliability coefficients, which are the factor analysis analogues of the person- and cluster-level reliability in our case. Cluster-level (relative) reliability in ten Hove et al. (2022) or between-level reliability in Geldhof et al. (2014) can be defined as the squared correlation of the cluster effect and the unobserved, latent cluster true score, which is composed of the cluster-related components in the model structure and ignores the measurement error. In case of the *Persons(Clusters)* $\times$ *Rater* design, this becomes: $\text{Cor}^2(C_k, \mu + C_k + CR_{kj})$. However, ignoring the measurement error may result in the fact that the cluster-level reliability is meaningless in some nested designs, where raters are nested within clusters, such as is demonstrated by the examples in Section and Appendix D. The estimated cluster-level reliability is always one regardless of how fluctuating the observed scores are, either for (ten Hove et al., 2022) in the context of generalizability theory or for Geldhof et al. (2014) in the context of multilevel confirmatory factor analysis. On the other hand, Schönbrodt et al. 2022 and Lai 2021 defined reliability based on the manifest observed scores. For example, manifest cluster-level reliability (or between-level reliability) for absolute decisions can be written as the square correlation of the cluster effect and the observed cluster mean score, which is the average over person scores given by all rates under the same cluster. That is $\text{Cor}^2(C_k, \frac{1}{n_S n_R}\sum_{i=1}^{n_S}\sum_{j=1}^{n_R} Y_{i(k)j})$ for the *Person(Cluster)* $\times$ *Rater* design. Under this definition, cluster-level reliability involves variability coming from measurement error as well as the variability between persons. Note that the square correlation is another way to construct manifest cluster-level reliability, compared to Equation 16 which we defined as a correlation between two observed scores from the two persons under the same

cluster but given by two raters.

Because the inconsistency between the definitions of Schönbrodt et al. (2022) and ten Hove et al. (2022) causes confusion about whose definition should be used in practice, it prompts us to adopt a simple, consistent, and extendable approach—the correlation approach—to defining reliability. It can also be applied to nested data structures not considered in ten Hove et al. (2022) and Schönbrodt et al. (2022) (see Design 3). The correlation approach was originally used in classical test theory under the parallel assumption (Lord et al., 1968) and was generalized by Molenberghs et al. (2007) and Vangeneugden et al. (2004, 2005) to more complex designs. We extend the approach to define person- and two types of cluster-level reliability for relative and absolute decisions.

The person-level reliability is defined unequivocally as the correlation between two observed scores conditional on a specific cluster. This correlation is equivalent to the correlation between two (latent) pure person-level scores, which are the observed scores removing the cluster effect. As a consequence, the person-level reliability coefficient excludes the cluster variance as a source of unsystematic error variance.

Regarding the cluster-level reliability, there are two perspectives on the role of the lower level person variance: (1) consider it as a relevant source of unsystematic error variance or (2) consider it as irrelevant fluctuation when discriminating between clusters. When taking the former perspective, the reliability coefficient is defined directly as a correlation between two manifest observed scores, without conditioning on the person. We have called this the manifest cluster-level reliability. For the latter, the reliability coefficient is defined as a correlation between two observed scores with conditioning on the same person. Formally, this is the same as a correlation between two latent pure cluster-level scores with the person effect removed from the observed scores. Therefore, we have called this the latent cluster-level reliability, and this reliability coefficient excludes the person variance from the coefficient. The question to which extent person-level variance should be taken into account in the cluster level reliability has been a topic of discussion in the literature (Bliese, 2000; Jak et al., 2023;

Klein & Kozlowski, 2000; Stapleton et al., 2016). While it is tempting to choose the latent cluster-level reliability because it is larger than the manifest cluster-level reliability by definition, this former is more appropriate when the construct at cluster level is not affected by changes at the person level.

While the correlation approach was used here to define reliability specifically at the cluster and person levels, it can also be used to define a reliability coefficient that merges several levels (e.g., persons and clusters). This is similar to what Vangeneugden et al. (2005) did. For example, in the $Person(Cluster) \times Rater$ design, the combined relative reliability can be defined through $\text{Cor}(Y_{i(k)j}, Y_{i(k)j'}|j, j') = \frac{\sigma^2_{S(C)}+\sigma^2_C}{\sigma^2_{S(C)}+\sigma^2_C+\sigma^2_\epsilon}$ and the combined absolute reliability through $\text{Cor}(Y_{i(k)j}, Y_{i(k)j'}) = \frac{\sigma^2_{S(C)}+\sigma^2_C}{\sigma^2_{S(C)}+\sigma^2_C+\sigma^2_R+\sigma^2_{CR}+\sigma^2_\epsilon}$. These types of reliability coefficients can be used when the aim is to compare scores of different persons across the clusters. An example of this situation (suggested to us by a reviewer) is the case where students work in groups. Ultimately, the students are graded individually and their grades are compared across the groups. In this situation, both the person (nested in cluster) variance term and the cluster variance term show up in the numerator.

The significant advantage of the correlation approach is its intelligibility because it is based on a clear definition. Reliability is defined as the degree to which an observed score can be generalized from one occasion to another, where the occasion can be understood broadly (e.g., rater, moment, etc.). Second, the correlation approach allows researchers to obtain easily and in a unified manner different types of reliability, such as overall reliability, test-retest reliability, or interrater reliability under the multi-way crossed designs, such as the $Person \times Rater \times Day$ design, and person-level reliability or cluster-level reliability under nested designs, such as the $Person(Cluster) \times Rater$ design, for example. Third, cluster-level reliability (including the latent and manifest versions) is always properly defined (and hence would not be estimated as one by default). Fourth, the correlation approach is flexible and can be applied to a variety of designs. The same principle can be used for completely crossed designs (Vangeneugden et al., 2005) or nested designs (discussed in this paper) based on the GT models to

decompose the source of variance. Vangeneugden et al. (2005) and Molenberghs et al. (2007) further discuss how to build reliability under (generalized) linear mixed-effects models. Last but not least, the effect of changing the sample size of different factors (e.g., clusters, persons, raters) can be studied similarly to what is done in D studies in GT, since the reliability coefficients can also be defined for average scores.

Despite variations in how to obtain the reliability coefficients, the results are all based on a ratio of (sums of) variance components. Consequently, the same estimation methods can be used for statistical inference regardless of the approach taken. One can use frequentist methods, such as mean square estimation (Brennan, 2001; Cronbach et al., 1963) or (restricted) maximum likelihood estimation (Marcoulides, 1990), to estimate the variance components and derive the reliability coefficients. Alternatively, Bayesian inference methods can also be employed. These methods can deal more flexibly with complex situations, such as missing data, unbalanced design, or data that does not follow the normality assumption (LoPilato et al., 2015; ten Hove et al., 2022). For detailed guidance on using Bayesian methods, we refer the reader to ten Hove et al. (2022).

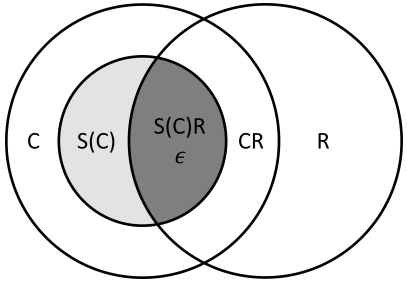
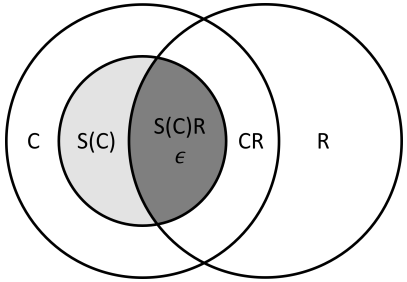
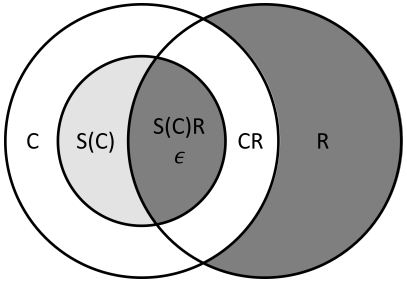
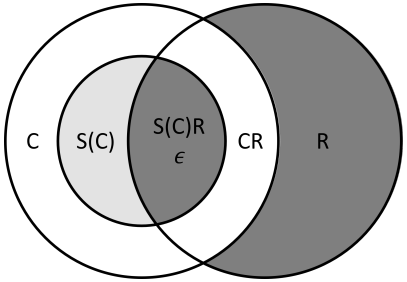
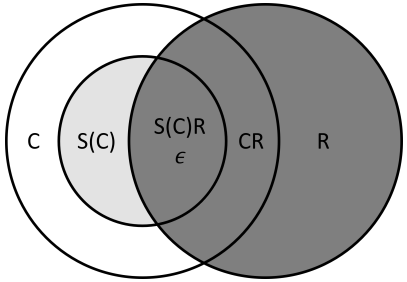
Obviously, the discussed reliability coefficients are derived from a set of modeling assumptions that may be violated. The most important assumption of our approach lingering in the background is that the model is correctly specified. If not, the misspecification can potentially influence the estimate of the variance components and reliability. In the GT model, we assume that all potential sources of variability (such as Equations 3, 12, and 23) are included in the model and that all effects are linearly additive and mutually uncorrelated. In addition, measurement errors must have an equal variance and are also assumed to be independent. These assumption is easily violated for time series data. For example, data obtained from the experience sampling method show dependence across time, and this is not taken into account in the standard GT formulations. However, as previously mentioned, the correlation approach is quite flexible in dealing with different kinds of model assumptions. For example, Molenberghs et al. (2007) has demonstrated how to define the reliability coefficient from

the generalized linear mixed effect model with serial correlation for longitudinal data with binary (or non-Gaussian) outcomes. A similar generalization can be made to account for time dependencies.

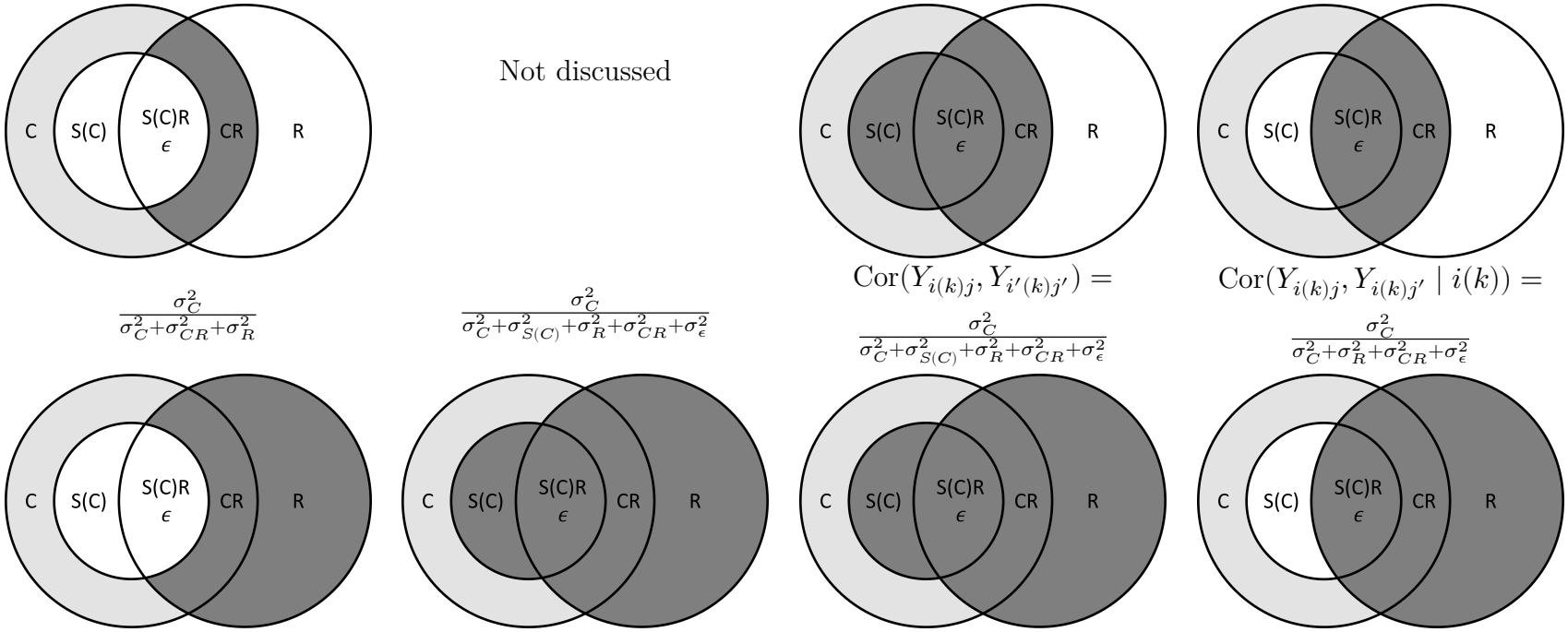
This paper provides a starting point for discussing the possibility of adopting the correlation approach to define reliability for multilevel data. Currently, we have only discussed the case of random effect models. Future research should consider the role of fixed effects when we discuss reliability. For example, in the two-way mixed-effect model, each person (randomly selected from the population) is rated by the same set of raters. Because they are the only raters of interest, the rater is seen as a fixed effect. Under this case, the absolute reliability defined in McGraw and Wong (1996) (i.e., absolute agreement ICC) includes the variance between the rater effects. However, the absolute reliability defined in generalizability theory (i.e., dependent coefficient Brennan, 2001) does not take the systematic shifts by raters into account. Additional studies to better understand the differences in the case of fixed effects are required.

Table 3

Reliability coefficients for the Person(Cluster) × Rater design

Reliability	Ten Hove et al. (2022)	Schönbrodt et al. (2022)	Correlation approach
$\rho^{\mathrm{S}}_{\mathrm{rel}}$	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_\epsilon}$ 	Not discussed	$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j') = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_\epsilon}$ 
$\rho^{\mathrm{S}}_{\mathrm{abs}}$	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_R+\sigma^2_\epsilon}$ 	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_R+\sigma^2_\epsilon}$ 	$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_R+\sigma^2_{CR}+\sigma^2_\epsilon}$ 

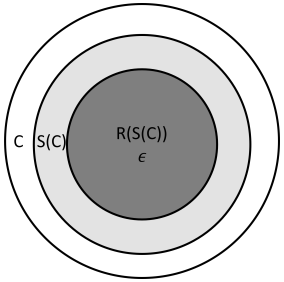
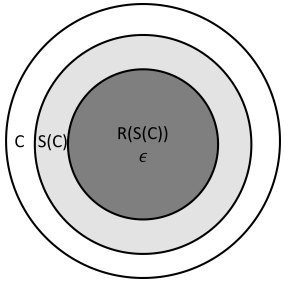
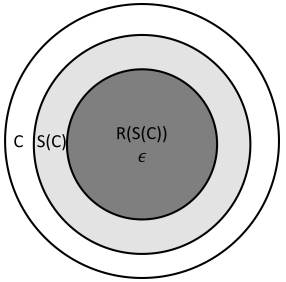
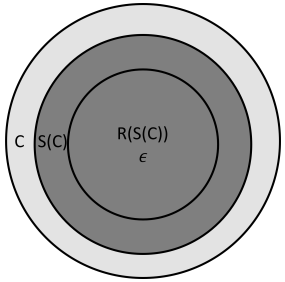
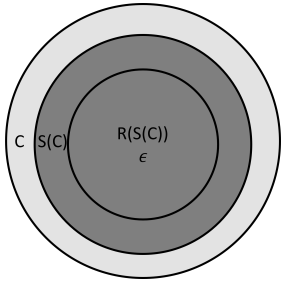
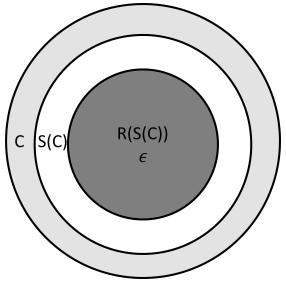
Reliability	Ten Hove et al. (2022)	Schönbrodt et al. (2022)	Correlation approach	
			$\text{Cor}(Y_{i(k)j}, Y_{i'(k)j'} \mid j, j') =$ $\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_{CR}+\sigma^2_\epsilon}$	$\text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j') = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{CR}+\sigma^2_\epsilon}$
$\rho^{\text{C}}_{\text{rel}}$	$\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{CR}}$	Not discussed		
$\rho^{\text{C}}_{\text{abs}}$	$\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{CR}+\sigma^2_R}$	$\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_R+\sigma^2_{CR}+\sigma^2_\epsilon}$	$\text{Cor}(Y_{i(k)j}, Y_{i'(k)j'}) =$ $\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_R+\sigma^2_{CR}+\sigma^2_\epsilon}$	$\text{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k)) =$ $\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_R+\sigma^2_{CR}+\sigma^2_\epsilon}$



Note: The Venn diagram illustrates the total variance of the score, which is divided into distinct variance components. Components that exist in both the numerator and denominator of the reliability coefficient are represented in light gray, while those that only appear in the denominator are represented in dark gray.

Table 4

Reliability coefficients for the Rater(Person(Cluster)) design

Reliability	Ten Hove et al. (2022)	Schönbrodt et al. (2022)[a]	Correlation approach
ρ^{S}	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_\epsilon}$ 	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_\epsilon}$ 	$\mathrm{Cor}(Y_{j(i(k))}, Y_{j'(i(k))} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_\epsilon}$ 
ρ^{C}	Cannot be defined.	$\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_\epsilon}$ 	$\mathrm{Cor}(Y_{j(i(k))}, Y_{j'(i'(k))}) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_\epsilon}$  $\mathrm{Cor}(Y_{j(i(k))}, Y_{j'(i(k))} \mid i(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_\epsilon}$ 

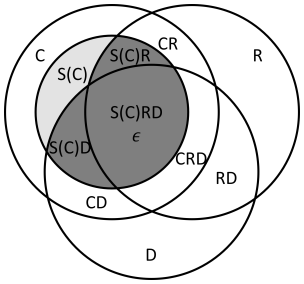
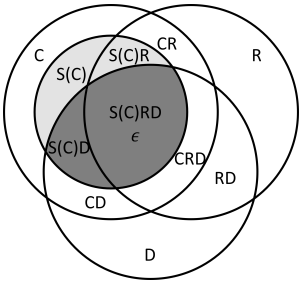
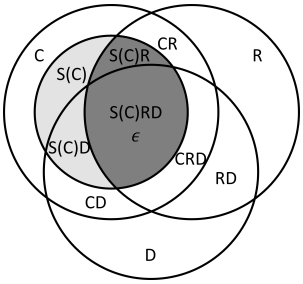
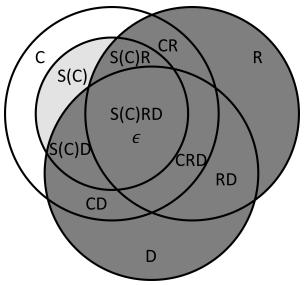
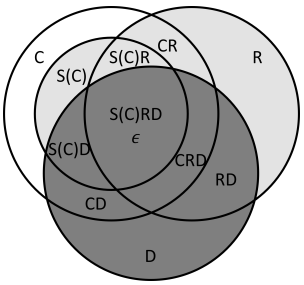
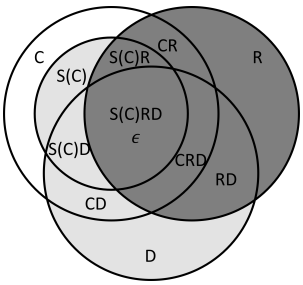
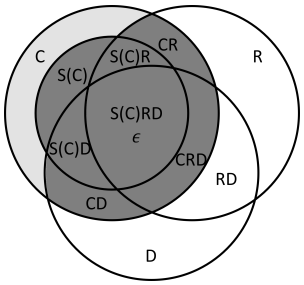
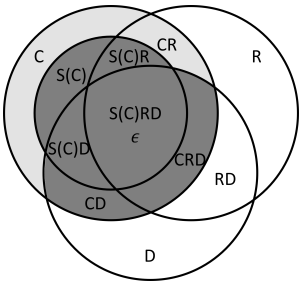
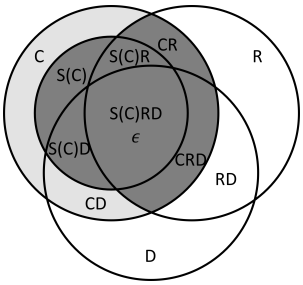
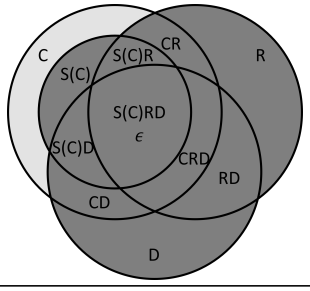
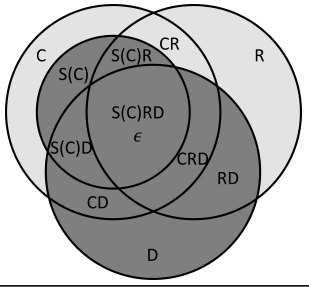
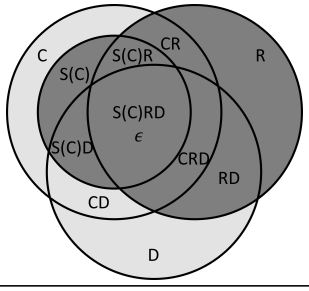
Reliability	Ten Hove et al. (2022)	Schönbrodt et al. (2022)[a]	Correlation approach

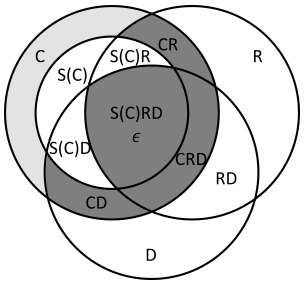
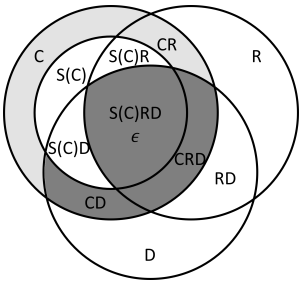
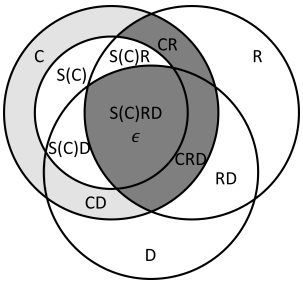
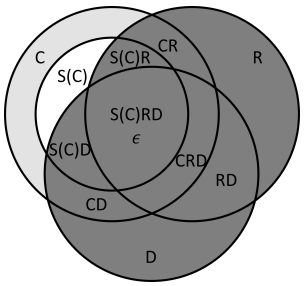
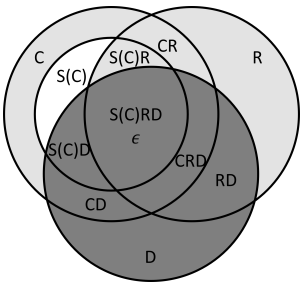
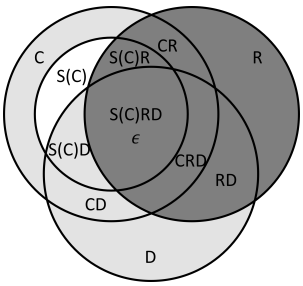
Note: The Venn diagram illustrates the total variance of the score, which is divided into distinct variance components. Components that exist in both the numerator and denominator of the reliability coefficient are represented in light gray, while those that only appear in the denominator are represented in dark gray.

[a] See also Bonito et al. (2012), Raudenbush et al. (1991), and Shrout and Lane (2012a), having the same definition under this case.

Table 5

Reliability coefficients for the Person(Cluster) × Rater × Day design

Reliability	Overall	Test-retest	Interrater
$\rho^{\mathrm{S}}_{\mathrm{rel}}$	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)j'm'} \mid k, j, j', m, m')$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)jm'} \mid k, j, m, m')$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)j'm} \mid k, j, j', m)$ 
$\rho^{\mathrm{S}}_{\mathrm{abs}}$	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)j'm'} \mid k)$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)jm'} \mid k)$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)j'm} \mid k)$ 
$\rho^{\mathrm{C,\ man}}_{\mathrm{rel}}$	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i'(k)j'm'} \mid j, j', m, m')$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)jm'} \mid j, m, m')$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i(k)j'm} \mid j, j', m)$ 
$\rho^{\mathrm{C,\ man}}_{\mathrm{abs}}$	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i'(k)j'm'})$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i'(k)jm'})$ 	$\mathrm{Cor}(Y_{i(k)jm}, Y_{i'(k)j'm})$ 

Reliability	Overall	Test-retest	Interrater
$\rho^{\text{C, lat}}_{\text{rel}}$	$\text{Cor}(Y_{i(k)jm}, Y_{i(k)j'm'} \mid i(k), j, j', m, m')$ 	$\text{Cor}(Y_{i(k)jm}, Y_{i(k)jm'} \mid i(k), j, m, m')$ 	$\text{Cor}(Y_{i(k)jm}, Y_{i(k)j'm} \mid i(k), j, j', m)$ 
$\rho^{\text{C, lat}}_{\text{abs}}$	$\text{Cor}(Y_{i(k)jm}, Y_{i(k)j'm'} \mid i(k))$ 	$\text{Cor}(Y_{i(k)jm}, Y_{i(k)jm'} \mid i(k))$ 	$\text{Cor}(Y_{i(k)jm}, Y_{i(k)j'm} \mid i(k))$ 

Note: The Venn diagram illustrates the total variance of the score, which is divided into distinct variance components. Components that exist in both the numerator and denominator of the reliability coefficient are represented in light gray, while those that only appear in the denominator are represented in dark gray.

References

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass/Wiley.

Bonito, J. A., Ruppel, E. K., & Keyton, J. (2012). Reliability estimates for multilevel designs in group research. *Small Group Research*, *43*(4), 443–467. https://doi.org/10.1177/1046496412437614

Brennan, R. L. (2001). *Generalizability theory*. Springer New York. https://doi.org/10.1007/978-1-4757-3456-0

Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*(3), 277–289.

Brennan, R. L., & Kane, M. T. (1977b). Signal/noise ratios for domain-referenced tests. *Psychometrika*, *42*(4), 609–625. https://doi.org/10.1007/BF02295983

Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying generalizability theory using EduG*. Routledge. https://doi.org/10.4324/9780203866948

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Johon Wilely & Sons.

Demetrashvili, N., Wit, E. C., & van den Heuvel, E. R. (2016). Confidence intervals for intraclass correlation coefficients in variance components models. *Statistical Methods in Medical Research*, *25*(5), 2359–2376. https://doi.org/10.1177/0962280214522787

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91. https://doi.org/10.1037/a0032138

Jak, S., Jorgensen, T. D., ten Hove, D., & Nevicka, B. (2023). Modeling Cluster-Level Constructs Measured by Individual Responses: Configuring a Shared Approach. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231182319. https://doi.org/10.1177/25152459231182319

Jeon, M.-J., Lee, G., Hwang, J.-W., & Kang, S.-J. (2009). Estimating reliability of school-level scores using multilevel and generalizability theory models. *Asia Pacific Education Review*, *10*, 149–158. https://doi.org/10.1007/s12564-009-9014-3

Klein, K. J., & Kozlowski, S. W. J. (Eds.). (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. Jossey-Bass/Wiley.

Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, *26*(1), 90–102. https://doi.org/10.1037/met0000287

LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management*, *41*(2), 692–717. https://doi.org/10.1177/0149206314554215

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

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, *66*(2), 379–386. https://doi.org/10.2466/pr0.1990.66.2.379

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Molenberghs, G., Laenen, A., & Vangeneugden, T. (2007). Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics*, *17*(4), 595–627. https://doi.org/10.1080/10543400701329448

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

1011 Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model
1012 for studying school climate with estimation via the EM algorithm and
1013 application to U.S. high-school data. *Journal of Educational Statistics*, *16*(4),
1014 295–330. https://doi.org/10.3102/10769986016004295
1015 Schönbrodt, F. D., Zygar-Hoffmann, C., Nestler, S., Pusch, S., & Hagemeyer, B. (2022).
1016 Measuring motivational relationship processes in experience sampling: A
1017 reliability model for moments, days, and persons nested in couples. *Behavior*
1018 *Research Methods*, *54*(4), 1869–1888.
1019 https://doi.org/10.3758/s13428-021-01701-7
1020 Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE
1021 Publications, Inc.
1022 Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory.
1023 *American Psychologist*, *44*(6), 922–932.
1024 https://doi.org/10.1037/0003-066X.44.6.922
1025 Shoukri, M. M. (2010). *Measures of interobserver agreement and reliability* (2nd ed.).
1026 CRC Press. https://doi.org/10.1201/b10433
1027 Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater
1028 reliability. *Psychological Bulletin*, *86*(2), 420–428.
1029 https://doi.org/10.1037/0033-2909.86.2.420
1030 Shrout, P. E., & Lane, S. P. (2012a). Psychometrics. In M. R. Mehl & T. S. Conner
1031 (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). The
1032 Guilford Press. https://psycnet.apa.org/record/2012-05165-017
1033 Shrout, P. E., & Lane, S. P. (2012b). Reliability. In H. Cooper, P. M. Camic,
1034 D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of*
1035 *research methods in psychology, vol 1: Foundations, planning, measures, and*
1036 *psychometrics.* (pp. 643–660). American Psychological Association.
1037 https://doi.org/10.1037/13619-034

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. https://doi.org/10.3102/1076998616646200

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales*. Oxford University Press. https://doi.org/10.1093/med/9780199685219.001.0001

Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding MORE multivariate statistics* (pp. 23–66). American Psychological Association.

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, *27*(4), 650–666. https://doi.org/10.1037/met0000391

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, *25*(1), 13–30. https://doi.org/10.1016/j.cct.2003.08.009

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, *61*(1), 295–304. https://doi.org/10.1111/j.0006-341X.2005.031040.x

Wang, L., Keen, K. J., & Holland, B. (2011). Estimation of reliability in a three-factor model. *Statistics in Medicine*, *30*(11), 1254–1265. https://doi.org/10.1002/sim.4182

Zee, M., Rudasill, K. M., & Roorda, D. L. (2020). "Draw Me a Picture": Student-teacher relationship drawings by children displaying externalizing, internalizing, or prosocial behavior. *The Elementary School Journal*, *120*(4), 636–666. https://doi.org/10.1086/708661

Zygar, C., Hagemeyer, B., Pusch, S., & Schönbrodt, F. D. (2018). From motive dispositions to states to outcomes: An intensive experience sampling study on

1066 communal motivational dynamics in couples. *European Journal of Personality*,
1067 *32*(3), 306–324. https://doi.org/10.1002/per.2145

Appendix A

Reliability in CTT through Parallel measurements

In this appendix, we will formally derive Equation 2. Although this is a well-known result and a derivation may therefore seem unnecessary, we think a step-by-step derivation is useful to fully understand the derivations of the reliability coefficients in more complicated models. In addition, we will make use of a more precise notation in the appendices (which we do not introduce in the main text for reasons of readability), and the derivation of a well-known result allows a gentle introduction.

Let Y_{ij} be a random variable defined over a population $\mathcal{P}$ of persons and a population $\mathcal{J}$ of measurement instruments or, in short, measurements (e.g., tests, raters, test items), meaning that an observed score of a person $i \in \mathcal{P}$ is generated by measurement $j \in \mathcal{J}$. Theoretically, $\mathcal{P}$ and $\mathcal{J}$ may contain a (countably) infinite number of elements or a finite number of elements. Based on the classic test theory (Novick, 1966), the true score T_{ij} of a person i at measurement j is defined as the expected value of the observed score; that is $T_{ij} = \mathrm{E}[Y_{ij}]$, a constant value. The measurement error ϵ_{ij} is defined as the difference between the observed score and the true score; that is $\epsilon_{ij} = Y_{ij} - T_{ij}$, a random variable with mean zero and variance $\sigma^2_{\epsilon_{ij}}$.

The above definitions are for a given person and a given measurement instrument. However, in psychometrics, we usually would like to generalize the results to a population of persons using the same measurement instrument. (Note that it can also generalize the result to a population of measurements, e.g., a population of raters or of items in a test.) To do this, we *randomly select* a person i under a specified measurement j (not random) in their respective population. As a result, when calculating the variance (or expectation) of the true score T_{ij} and random measurement error ϵ_{ij} over the population, caution is needed because there are two kinds of random sampling mechanisms: one is from a random sampling of the persons, and the other one is from the error as such, which takes different realizations in the hypothetical replications. Thus, the variance of the true scores is only random between the persons, while the variance of the measurement error is determined by these two mechanisms. If

we let $\sigma^2_{T_j}$ and $\sigma^2_{\epsilon_j}$ denote the variances of T_{ij} and ϵ_{ij}, respectively, then we have:

$$\sigma^2_{T_j} \coloneqq \mathrm{Var}(T_{ij} \mid j) = \mathrm{Var}_{\mathcal{P}}(T_{ij} \mid j), \tag{A1}$$

$$\sigma^2_{\epsilon_j} \coloneqq \mathrm{Var}(\epsilon_{ij} \mid j) = \mathrm{Var}_{\mathcal{P}}(\mathrm{E}[\epsilon_{ij} \mid i, j]) + \mathrm{E}_{\mathcal{P}}[\mathrm{Var}(\epsilon_{ij} \mid i, j)] = \mathrm{E}_{\mathcal{P}}[\sigma^2_{\epsilon_{ij}}]. \tag{A2}$$

Note that the conditioning on j (but not on i) stresses the fact that the measurement j is fixed while variance is taken over the population of persons (and hence the index i is random). In Equations A1 and A2, the populations over which the expectation or variance is computed (and later also the covariance) are sometimes implicitly but sometimes also explicitly denoted. For example, in Equations A1 and A2, the expectation or variance without subscription means that the expectation or variance of the random variable is taken over all possible populations that make it random. The expectation or variance with the subscript $\mathcal{P}$ means that it takes the expectation or variance over the person population only.

Consider two specific measurements $j, j' \in \mathcal{J}$. We aim to determine the correlation between two observed scores on two different measurements j, j' for a randomly selected person i, that is $\mathrm{Cor}(Y_{ij}, Y_{ij'} \mid j, j')$. Based on the classical linear theory model, the two observed scores are:

$$Y_{ij} = T_{ij} + \epsilon_{ij} \text{ and } Y_{ij'} = T_{ij'} + \epsilon_{ij'}.$$

Under the above setting, one could derive three properties: (1) $\mathrm{E}[\epsilon_{ij} \mid j] = \mathrm{E}[\epsilon_{ij'} \mid j'] = 0$, (2) $\mathrm{Cov}(\epsilon_{ij}, T_{ij} \mid j) = \mathrm{Cov}(\epsilon_{ij'}, T_{ij'} \mid j') = 0$, and (3) $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{ij'} \mid j, j') = 0$ for a randomly selected person $i \in \mathcal{P}$ and $j \neq j' \in \mathcal{J}$. These properties are actually the axioms of CTT (Novick, 1966).

Furthermore, under the parallel assumption, any two measurements j, j' in $\mathcal{J}$ have the same true scores (i.e., $T_{ij} = T_{ij'}$) and equal error variances (i.e., $\sigma^2_{\epsilon_{ij}} = \sigma^2_{\epsilon_{ij'}}$) for every person. According to the definition in Equations A1 and A2, these also lead to equal variances of true score $\sigma^2_{T_j} = \sigma^2_{T_{j'}} = \sigma^2_T$ and equal variances of measurement error $\sigma^2_{\epsilon_j} = \sigma^2_{\epsilon_{j'}} = \sigma^2_\epsilon$ over the person population regardless which measurement is considered.

Based on the above setting, the correlation between two measurements j and j'

for a randomly selected person i is:

$$\text{Cor}(Y_{ij}, Y_{ij'} \mid j, j') = \frac{\text{Cov}(Y_{ij}, Y_{ij'} \mid j, j')}{\sqrt{\text{Var}(Y_{ij} \mid j)\,\text{Var}(Y_{ij'} \mid j')}}. \tag{A3}$$

The conditioning on j, j' also stresses it merely considers these two measurements. The numerator in Equation A3 is as follows:

$$\begin{aligned}
\text{Cov}(Y_{ij}, Y_{ij'} \mid j, j') &= \text{Cov}_{\mathcal{P}}(\text{E}[Y_{ij} \mid i, j], \text{E}[Y_{ij'} \mid i, j']) + \text{E}_{\mathcal{P}}[\text{Cov}(Y_{ij}, Y_{ij'} \mid i, j, j')] \\
&= \text{Cov}_{\mathcal{P}}(\text{E}[T_{ij} + \epsilon_{ij} \mid i, j], \text{E}[T_{ij'} + \epsilon_{ij'} \mid i, j']) \\
&\quad + \text{E}_{\mathcal{P}}[\text{Cov}(T_{ij} + \epsilon_{ij}, T_{ij'} + \epsilon_{ij'} \mid i, j, j')] \\
&= \text{Cov}_{\mathcal{P}}(T_{ij} + 0, T_{ij'} + 0 \mid j, j') + \text{E}_{\mathcal{P}}[\text{Cov}(T_{ij}, T_{ij'} \mid i, j, j') \\
&\quad + \text{Cov}(T_{ij}, \epsilon_{ij'} \mid i, j, j') + \text{Cov}(\epsilon_{ij}, T_{ij'} \mid i, j, j') + \text{Cov}(\epsilon_{ij}, \epsilon_{ij'} \mid i, j, j')] \\
&= \text{Var}_{\mathcal{P}}(T_{ij}) + \text{E}_{\mathcal{P}}[0] = \sigma_T^2.
\end{aligned}$$

We get the third equation because the expectation of measurement error is zero. In the same equation, there are two seemingly similar terms, $\text{Cov}_{\mathcal{P}}(T_{ij}, T_{ij'} \mid j, j')$ and $\text{Cov}(T_{ij}, T_{ij'} \mid i, j, j')$, but with rather different meanings. The former does not condition on person i, so T_{ij} and $T_{ij'}$ are random variables. Furthermore, because of the parallel assumption $T_{ij} = T_{ij'}$, they are the same quantity and the covariance of them is actually the variance of T_{ij} over the person population (see the fourth equation). The latter term, however, does fix a specific person i as well as measurements j, j', so T_{ij} and $T_{ij'}$ are the same constant and the covariance of them turns out to be zero. In addition, due to the CTT axioms, the covariances between errors and between error and true score are also uncorrelated. Therefore, the four covariances in the expectation without subscript $\mathcal{P}$ in the third equation is zero.

In the denominator in Equation A3, we further have:

$$\begin{aligned}
\text{Var}(Y_{ij} \mid j) &= \text{Var}(T_{ij} + \epsilon_{ij} \mid j) \\
&= \text{Var}(T_{ij} \mid j) + \text{Var}(\epsilon_{ij} \mid j) \\
&= \sigma_T^2 + \sigma_\epsilon^2
\end{aligned}$$

In the second to the third equations, we follow the definition in Equations A1 and A2. Similarly, we also have: $\text{Var}(Y_{ij'} \mid j') = \sigma_T^2 + \sigma_\epsilon^2$.

Thus, combining the above result into Equation A3, the correlation under the parallel assumptions is:

$$\text{Cor}(Y_{ij}, Y_{ij'} \mid j, j') = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\epsilon^2}. \tag{A4}$$

Comparing Equation A4 with Equation 1, we get $\text{Cor}(Y_{ij}, Y_{ij'} \mid j, j') = \text{Cor}^2(Y_i, T_i)$. Thus, the correlation between parallel measurements is equal to the squared correlation between observed scores and true scores, which is the definition of the reliability of a test/measurement instrument (Lord et al., 1968).

Appendix B

Derivation of the reliability for a three-way crossed design

We can use the correlation approach from Appendix A to define the reliability in a more complicated model. Consider a three-way *Person* × *Rater* × *Day* design, and let Y_{ijm} be a random variable of the observed score for the person $i \in \mathcal{P}$ given by the rater $j \in \mathcal{J}$ on the day $m \in \mathcal{M}$. $\mathcal{P}$, $\mathcal{J}$, and $\mathcal{M}$ are the populations of persons, raters, and days, respectively. The observed score can be decomposed as follows:

$$Y_{ijm} = \mu + S_i + R_j + D_m + SR_{ij} + SD_{im} + RD_{jm} + \epsilon_{ijm},$$

where μ is the overall mean, S_i is the person effect, R_j is the rater effect, D_m is the day effect, SR_{ij}, SD_{im}, and RD_{jm} are their interaction effects, and ϵ_{ijm} is the measurement error. Each effect and error has mean zero and variance, such as σ_S^2, σ_R^2, and so on, over their population, and all effects are mutually uncorrelated.

Based on the correlation approach from Molenberghs et al. (2007) and Vangeneugden et al. (2005), the overall reliability is defined as the conditional correlation between two observed scores for the same person i but from two fixed raters j and j' and two fixed days m and m'. That is:

$$\mathrm{Cor}(Y_{ijm}, Y_{ij'm'} \mid j, j', m, m') = \frac{\mathrm{Cov}(Y_{ijm}, Y_{ij'm'} \mid j, j', m, m')}{\sqrt{\mathrm{Var}(Y_{ijm} \mid j, m)\,\mathrm{Var}(Y_{ij'm'} \mid j', m')}}. \tag{B1}$$

Note that the person i is randomly selected from $\mathcal{P}$, not fixed.

The numerator in Equation B1 is as follows:

$$\begin{aligned}
&\mathrm{Cov}(Y_{ijm}, Y_{ij'm'} \mid j, j', m, m') \\
&\quad = \mathrm{Cov}_{\mathcal{P}}(\mathrm{E}[Y_{ijm} \mid i, j, m], \mathrm{E}[Y_{ij'm'} \mid i, j', m']) + \mathrm{E}_{\mathcal{P}}[\mathrm{Cov}(Y_{ijm}, Y_{ij'm'} \mid i, j, j', m, m')] \\
&\quad = \mathrm{Cov}_{\mathcal{P}}(\mu + S_i + R_j + D_m + SR_{ij} + SD_{im} + RD_{jm}, \\
&\qquad\qquad \mu + S_i + R_{j'} + D_{m'} + SR_{ij'} + SD_{m'} + RD_{j'm'} \mid j, j', m, m') + \mathrm{E}_{\mathcal{P}}[0] \\
&\quad = \mathrm{Var}_{\mathcal{P}}(S_i) = \sigma_S^2.
\end{aligned}$$

We further have in the denominator:

$$\begin{aligned}\mathrm{Var}(Y_{ijm} \mid j, m) &= \mathrm{Var}_{\mathcal{P}}(\mathrm{E}[Y_{ijm} \mid i, j, m) + \mathrm{E}_{\mathcal{P}}[\mathrm{Var}(Y_{ijm} \mid i, j, m)] \\ &= \mathrm{Var}_{\mathcal{P}}(\mu + S_i + R_j + D_m + SR_{ij} + SD_{im} + RD_{jm} \mid j, m) + \mathrm{E}_{\mathcal{P}}[\sigma_\epsilon^2] \\ &= \sigma_S^2 + \sigma_{SD}^2 + \sigma_{SM}^2 + \sigma_\epsilon^2.\end{aligned}$$

Similarly, we have: $\mathrm{Var}(Y_{ij'm'}) = \sigma_S^2 + \sigma_{SD}^2 + \sigma_{SM}^2 + \sigma_\epsilon^2$. Thus,

$$\mathrm{Cor}(Y_{ijm}, Y_{ij'm'} \mid j, j', m, m') = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_{SD}^2 + \sigma_{SM}^2 + \sigma_\epsilon^2},$$

is the same as the generalizability coefficient in GT (Cronbach et al., 1963).

The absolute overall reliability, which is defined in Molenberghs et al. (2007) and Vangeneugden et al. (2005), can also be proved to be the same as the dependability coefficient in GT (Brennan & Kane, 1977a, 1977b) by the similar deriving process as above. That is $\mathrm{Cor}(Y_{ijm}, Y_{ij'm'}) = \frac{\sigma_S^2}{\sigma_S^2+\sigma_R^2+\sigma_D^2+\sigma_{SR}^2+\sigma_{SD}^2+\sigma_{RD}^2+\sigma_\epsilon^2}$. However, it should be carefully noted that the correlation here doesn't condition on rater and day anymore, meaning that person i as well as rater j and day m are randomly selected from their population. Thus, when calculating the (co)variance of Y_{ijm} and $Y_{ij'm'}$, one should consider that there are four types of random mechanisms: randomly selecting a person, a rater, and a day and random measurement error.

Appendix C

Reliability for multilevel data by the correlation approach

Consider the *Person(Cluster)* $\times$ *Rater* design in Section for illustration, and let $Y_{i(k)j}$ be a random variable of the observed score for the person $i \in \mathcal{P}$ coming from the cluster $k \in \mathcal{C}$ given by the rater $j \in \mathcal{J}$. $\mathcal{P}$, $\mathcal{K}$, and $\mathcal{J}$ are the populations of persons, clusters, and days, respectively. The notation $i(k)$ expresses the person i is nested within the cluster k. The observed score can be decomposed by

$$Y_{i(k)j} = \mu + C_k + S(C)_{i(k)} + R_j + CR_{kj} + \epsilon_{i(k)j},$$

where μ is the grand mean, C_k is the cluster effect, $S(C)_{i(k)}$ is the person effect, R_j is the rater effect, CR_{kj} is the cluster and rater interaction effect, and $\epsilon_{i(k)j}$ is the measurement error. Each effect including the error term has mean zero and variance, such as σ^2_C, $\sigma^2_{S(C)}$, and so on, over their population, and all effects are mutually uncorrelated. Note that there is only one observation under every combination of these factors, so the model does not contain the person and rater interaction effect (i.e., $S(C)R_{i(k)j}$).

We propose to use conditional correlation to define person-level and (manifest and latent) cluster-level reliability for relative and absolute decisions. The derivations below also follow the technique used in the previous proof in Appendix A; in this case, we demonstrate how to derive person- and latent cluster-level relative reliability, but not for the absolute ones and manifest cluster-level reliability for relative and absolute one. However, the derivation for the remainder is quite similar.

Person-level relative reliability

The person-level relative reliability coefficient (Equation 13) is defined as the correlation coefficient between two scores from the same person i randomly selected from a fixed cluster k and fixed two raters j and j':

$$\rho^{\mathrm{S}}_{\mathrm{rel}} = \mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j') = \frac{\mathrm{Cov}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j')}{\sqrt{\mathrm{Var}(Y_{i(k)j} \mid k, j)\,\mathrm{Var}(Y_{i(k)j'}, \mid k, j')}}. \tag{C1}$$

Note that the person i is randomly selected from the subsection of $\mathcal{P}$ that belongs to the cluster k, not a fixed person.

The numerator in Equation C1 is as follows:

$$\begin{aligned}\mathrm{Cov}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j') &= \mathrm{Cov}_{\mathcal{P}}(\mathrm{E}[Y_{i(k)j} \mid i, k, j], \mathrm{E}[Y_{i(k)j'} \mid i, k, j']) \\ &\quad + \mathrm{E}_{\mathcal{P}}[\mathrm{Cov}(Y_{i(k)j}, Y_{i(k)j'} \mid i, k, j, j')] \\ &= \mathrm{Cov}_{\mathcal{P}}(\mu + C_k + S(C)_{i(k)} + R_j + (CR)_{kj}, \\ &\quad \mu + C_k + S(C)_{i(k)} + R_{j'} + (CR)_{kj'} \mid k, j, j') + \mathrm{E}_{\mathcal{P}}[0] \\ &= \mathrm{Var}_{\mathcal{P}}(S(C)_{i(k)} \mid k) = \sigma^2_{S(C)}.\end{aligned}$$

We further have in the denominator:

$$\begin{aligned}\mathrm{Var}(Y_{i(k)j} \mid k, j) &= \mathrm{Var}_{\mathcal{P}}(\mathrm{E}[Y_{i(k)j} \mid i, k, j]) + \mathrm{E}_{\mathcal{P}}[\mathrm{Var}(Y_{i(k)j} \mid i, k, j)] \\ &= \mathrm{Var}_{\mathcal{P}}(\mu + C_k + S_{i(k)} + R_j + (CR)_{kj} \mid k, j) + \mathrm{E}_{\mathcal{P}}[\sigma^2_\epsilon \mid k, j] \\ &= \sigma^2_{S(C)} + \sigma^2_\epsilon.\end{aligned}$$

In the same way, we have: $\mathrm{Var}(Y_{i(k)j'} \mid k, j') = \sigma^2_{S(C)} + \sigma^2_\epsilon$. Thus,

$$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j') = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)} + \sigma^2_\epsilon}.$$

Cluster-level relative reliability

Conversely, the latent cluster-level relative reliability coefficient (Equation 17) is defined as the correlation coefficient between two scores from the same fixed person i belonging to a randomly selected cluster k from $\mathcal{K}$ and fixed two raters j and j'.

$$\rho^{\mathrm{C}}_{\mathrm{rel}} = \mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j') = \frac{\mathrm{Cov}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j')}{\sqrt{\mathrm{Var}(Y_{i(k)j} \mid i(k), j)\,\mathrm{Var}(Y_{i(k)j'} \mid i(k), j')}}. \quad \text{(C2)}$$

It is noteworthy that, in contrast to person-level reliability, the cluster k, in this case, is chosen at random, but the individual $i(k)$ is predetermined. Put differently, we hypothetically set the individual to show up in multiple clusters, but we always select the same individual regardless of the cluster that is chosen at random.

The numerator in Equation C2 is as follows:

$$\begin{aligned}\mathrm{Cov}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j') &= \mathrm{Cov}_{\mathcal{K}}(\mathrm{E}[Y_{i(k)j} \mid i(k), k, j], \mathrm{E}[Y_{i(k)j'} \mid i(k), k, j']) \\ &\quad + \mathrm{E}_{\mathcal{K}}[\mathrm{Cov}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), k, j, j'] \\ &= \mathrm{Cov}_{\mathcal{K}}(\mu + C_k + S(C)_{i(k)} + R_j + (CR)_{kj}, \\ &\quad \mu + C_k + S(C)_{i(k)} + R_{j'} + (CR)_{kj'}) + \mathrm{E}_{\mathcal{K}}[0] \\ &= \mathrm{Var}_{\mathcal{K}}(C_k) = \sigma_C^2.\end{aligned}$$

We further have in the denominator:

$$\begin{aligned}\mathrm{Var}(Y_{i(k)j} \mid i(k), j) &= \mathrm{Var}_{\mathcal{K}}(\mathrm{E}[Y_{i(k)j} \mid i(k), k, j]) + \mathrm{E}_{\mathcal{K}}[\mathrm{Var}(Y_{i(k)j} \mid i(k), k, j)] \\ &= \mathrm{Var}_{\mathcal{K}}(\mu + C_k + S(C)_{i(k)} + R_j + (CR)_{kj} \mid i(k), j) + \mathrm{E}_{\mathcal{K}}[\sigma_\epsilon^2] \\ &= \sigma_C^2 + \sigma_{CR}^2 + \sigma_\epsilon^2.\end{aligned}$$

The same derivation for: $\mathrm{Var}(Y_{i(k)j'} \mid i(k), j') = \sigma_C^2 + \sigma_{CR}^2 + \sigma_\epsilon^2$. Thus, we get:

$$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j') = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{CR}^2 + \sigma_\epsilon^2}.$$

Similar techniques can be applied to derive the coefficients of person-level absolute reliability (Equation 14), latent cluster-level absolute reliability (Equation 18). Yet, it needs to additionally deal with the random selection between raters. It is also applied to derive the coefficients of manifest cluster-level relative (Equation 15) and absolute reliability (Equation 16), but they should focus more on the random selection between persons.

Appendix D

Another example of reliability for multilevel structure with Venn diagrams

Consider a design where persons and raters are nested within clusters, and persons and rater are crossed with each other (i.e., *Person(Cluster)* $\times$ *Rater(Cluster)*; Case 2 in ten Hove et al., 2022). In this design, $Y_{i(k)j(k)}$ is a score of person $i(k)$ obtained from rater $j(k)$ under the cluster k. We also assume a balanced design where all clusters have the same number of persons and rater, and each rater only gives one rating for each person. Based on GT, $Y_{i(k)j(k)}$ can be decomposed as

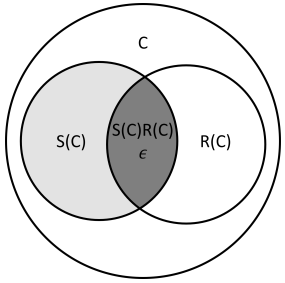
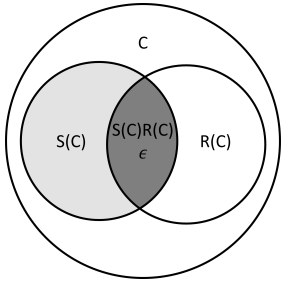
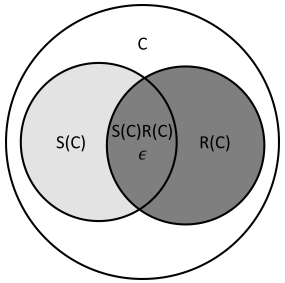
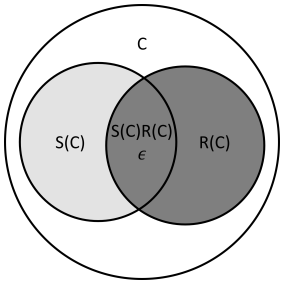
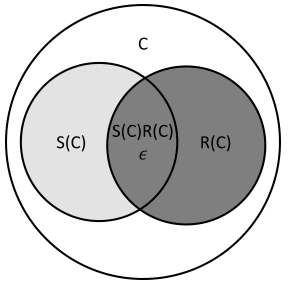
$$Y_{i(k)j(k)} = \mu + C_k + S(C)_{i(k)} + R(C)_{j(k)} + \epsilon_{i(k)j(k)},$$

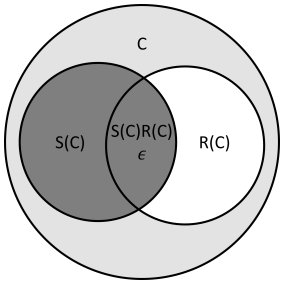
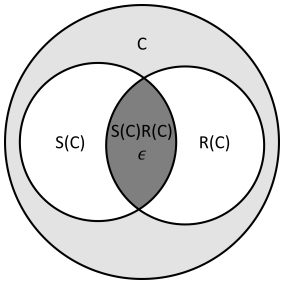
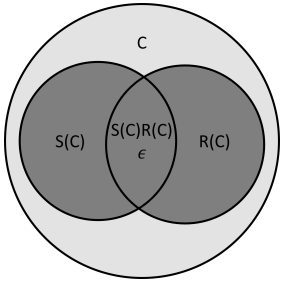
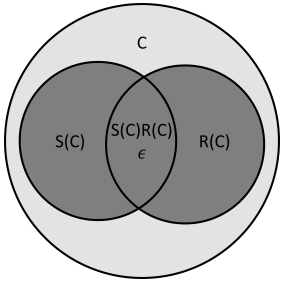
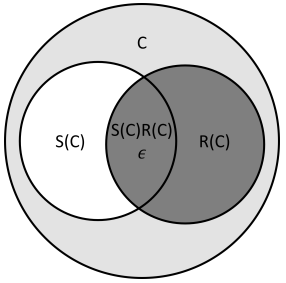
where μ is the grand mean, C_k is the cluster effect, $S(C)_{i(k)}$ is the person effect, $R(C)_{j(k)}$ is the rater effect, and $\epsilon_{i(k)j(k)}$ is the measurement error. Each effect and error is assumed to be mutually uncorrelated and distributed with a mean of zero and a positive, finite variance. Thus, the variance of the observed score can be decomposed as $\text{Var}(Y_{i(k)j(k)}) = \sigma^2_C + \sigma^2_{S(C)} + \sigma^2_{R(C)} + \sigma^2_\epsilon$.

Table D1 presents the person- and (manifest and latent) cluster-level reliability coefficients for relative and absolute decisions obtained with the three approaches. It is worth noting that, under this case, the cluster-level reliability coefficients for both relative and absolute decisions cannot be defined under the ten Hove et al.'s 2022 approach, and the person- and cluster-level reliability coefficients for relative decision were not discussed by Schönbrodt et al. (2022). Nevertheless, the correlation approach could define these four types of reliability coefficients using the rules we described in Section .

Table D1

Reliability coefficients for Person(Cluster) × Rater(Cluster) design

Reliability	Ten Hove et al. (2022)	Schönbrodt et al. (2022)	Correlation approach
$\rho^{\mathrm{S}}_{\mathrm{rel}}$	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{\epsilon}}$ 	Not discussed	$\mathrm{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid k, j(k), j'(k)) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{\epsilon}}$ 
$\rho^{\mathrm{S}}_{\mathrm{abs}}$	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{R(C)}+\sigma^2_{\epsilon}}$ 	$\frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{R(C)}+\sigma^2_{\epsilon}}$ 	$\mathrm{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{R(C)}+\sigma^2_{\epsilon}}$ 

Reliability	Ten Hove et al. (2022)	Schönbrodt et al. (2022)	Correlation approach	
			$\text{Cor}(Y_{i(k)j(k)}, Y_{i'(k)j'(k)} \mid j(k), j'(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_\epsilon}$	$\text{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid i(k), j(k), j'(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_\epsilon}$
$\rho^{\text{C}}_{\text{rel}}$	Cannot be defined	Not discussed		
		$\frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_{R(C)}+\sigma^2_\epsilon}$	$\text{Cor}(Y_{i(k)j(k)}, Y_{i'(k)j'(k)}) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_{R(C)}+\sigma^2_\epsilon}$	$\text{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid i(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{R(C)}+\sigma^2_\epsilon}$
$\rho^{\text{C}}_{\text{abs}}$	Cannot be defined			

Note: The Venn diagram illustrates the total variance of the score, which is divided into distinct variance components. Components that exist in both the numerator and denominator of the reliability coefficient are represented in light gray, while those that only appear in the denominator are represented in dark gray.

Appendix E

Reliability for Average Scores

Most reliability coefficients that we have presented in the main body of the paper are for single scores. However, in practical applications, researchers usually use averages or sum scores obtained from multiple replications over raters, items, and/or time points, to represent study targets' attributes or abilities. Aggregating is used to increase the reliability compared to individual scores provided by different raters, items, or time points. By averaging the scores, the random error variance (and some other variances) are divided by the number of replications in the reliability coefficient. In the following, we will demonstrate how to use the correlation approach to define reliability under the fully crossed designs and hierarchical designs.

Reliability in Fully Crossed Designs

In Section Error variance in Average Scores, we have briefly discussed relative reliability (i.e., the generalizability coefficient) and absolute reliability (i.e., the dependency coefficient) for average scores using the generalizability theory. Indeed, the correlation approach is also approachable to average scores in a similar manner. Let us consider the *Person* $\times$ *Rater* $\times$ *Day* design again and let Y_{ijm} be a single observed score for person $i \in \{1, \dots, n_S\} \subset \mathcal{P}$ given by rater $j \in \{1, \dots, n_R\} \subset \mathcal{R}$ on day $m \in \{1, \dots, n_D\} \subset \mathcal{D}$. We can calculate the average score across raters and days for each person, that is $\bar{Y}_{i\cdot\cdot} = \frac{1}{n_R n_D} \sum_{j=1}^{n_R} \sum_{k=1}^{n_D} Y_{ijm}$, where the subscript $\cdot$'s represent the average terms.

As for reliability, researchers need to consider which aspect they want this average score to generalize over. For example, overall reliability is concerned with the generalization over all aspects of replications, including raters and days. Therefore, we look at the consistency between repeated measurements for person i averaged over newly sampled raters $j' \in \{1, \dots, n_R\}' \subset \mathcal{R}$ and days $m' \in \{1, \dots, n_D\}' \subset \mathcal{D}$, i.e., $\bar{Y}_{i\cdot'\cdot'} = \frac{1}{n_R n_D} \sum_{j'=1}^{n_R} \sum_{m'=1}^{n_D} Y_{ij'm'}$. Analogue to the overall reliability for single scores (see Equations 8 and 10), the overall reliability for relative decision is the correlation

between these two average scores conditioned on all sampled raters and days involved:

$$\rho_{\text{rel}} = \text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot'\cdot'} \mid \forall j, \forall j', \forall m, \forall m'), \tag{E1}$$

where $\forall j$ means that all sampled raters are involved in calculating the sample average. Similar definitions hold for $\forall j'$, $\forall m$, and $\forall m'$. The overall reliability for absolute decision does not need to be conditioned on them:

$$\rho_{\text{abs}} = \text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot'\cdot'}). \tag{E2}$$

Similarly, interrater reliability generalizes on different sets of raters but for the same person on the same days, and test-retest reliability only generalizes on different sets of days but for the same person and raters. Relative and absolute decisions can also be distinguished in that case. The summary of these various reliability coefficients for average scores in the *Person* × *Rater* × *Day* design is shown in Table E1, which can be compared to Table 1 for single scores. The difference between single scores and average scores is that the variances used in average are divided by the number of replications.

Table E1

Reliability coefficients for average scores in Person × Rater × Day design

Reliability	Overall	Test-retest	Interrater
ρ_{rel}	$\text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot'\cdot'} \mid \forall j, \forall j', \forall m, \forall m') = \frac{\sigma_S^2}{\sigma_S^2+\frac{\sigma_{SR}^2}{n_R}+\frac{\sigma_{SD}^2}{n_D}+\frac{\sigma_\epsilon^2}{n_R n_D}}$	$\text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot\cdot'} \mid \forall j, \forall m, \forall m') = \frac{\sigma_S^2+\frac{\sigma_{SR}^2}{n_R}}{\sigma_S^2+\frac{\sigma_{SR}^2}{n_R}+\frac{\sigma_{SD}^2}{n_D}+\frac{\sigma_\epsilon^2}{n_R n_D}}$	$\text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot'\cdot} \mid \forall j, \forall j', \forall m) = \frac{\sigma_S^2+\frac{\sigma_{SD}^2}{n_D}}{\sigma_S^2+\frac{\sigma_{SR}^2}{n_R}+\frac{\sigma_{SD}^2}{n_D}+\frac{\sigma_\epsilon^2}{n_R n_D}}$
ρ_{abs}	$\text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot'\cdot'}) = \frac{\sigma_S^2}{\left(\sigma_S^2+\frac{\sigma_R^2}{n_R}+\frac{\sigma_D^2}{n_D}+\frac{\sigma_{SR}^2}{n_R}+\frac{\sigma_{SD}^2}{n_D}+\frac{\sigma_{RD}^2}{n_R n_D}+\frac{\sigma_\epsilon^2}{n_R n_D}\right)}$	$\text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot\cdot'}) = \frac{\sigma_S^2+\frac{\sigma_R^2}{n_R}+\frac{\sigma_{SR}^2}{n_R}}{\left(\sigma_S^2+\frac{\sigma_R^2}{n_R}+\frac{\sigma_D^2}{n_D}+\frac{\sigma_{SR}^2}{n_R}+\frac{\sigma_{SD}^2}{n_D}+\frac{\sigma_{RD}^2}{n_R n_D}+\frac{\sigma_\epsilon^2}{n_R n_D}\right)}$	$\text{Cor}(\bar{Y}_{i\cdot\cdot}, \bar{Y}_{i\cdot'\cdot}) = \frac{\sigma_S^2+\frac{\sigma_D^2}{n_D}+\frac{\sigma_{SD}^2}{n_D}}{\left(\sigma_S^2+\frac{\sigma_R^2}{n_R}+\frac{\sigma_D^2}{n_D}+\frac{\sigma_{SR}^2}{n_R}+\frac{\sigma_{SD}^2}{n_D}+\frac{\sigma_{RD}^2}{n_R n_D}+\frac{\sigma_\epsilon^2}{n_R n_D}\right)}$

Reliability for Multilevel Designs

In Section The Correlation Approach, three level-specific reliabilities were discussed–person-level reliability, manifest cluster-level reliability, and latent

cluster-level reliability. Let us consider the *Person(Cluster)* × *Rater* design for example and tell how these reliabilities are used for average scores. Suppose that $Y_{i(k)j}$ is an single observed score for person $i \in \{1, \dots, n_S\}$ coming from cluster $k \in \{1, \dots, n_C\}$ given by rater $j \in \{1, \dots, n_C\}$. The average scores adopted at the person level and cluster level are different.

On the one hand, at the person level, researchers consider the person's average scores that average over raters' responses, $\bar{Y}_{i(k)\cdot} = \frac{1}{n_R} \sum_{j=1}^{n_R} Y_{i(k)j}$, as an evaluate for that person i under the cluster k. The person-level reliability 2expresses how a person's average score $\bar{Y}_{i(k)\cdot}$ is consistent across two occasions, over the person population under a specific cluster k. The average score on the second occasion is $\bar{Y}_{i(k)\cdot'} = \frac{1}{n_R} \sum_{j'=1}^{n_R} Y_{i(k)j'}$, where scores are given by a different set of raters $j' \in \{1, \dots, n_R\}'$. As we discuss person-level reliability for single scores (Equations 13 and 14), the person-level relative reliability coefficient for average scores is defined as:

$$\rho^{\mathrm{S}}_{\mathrm{rel}} = \mathrm{Cor}(\bar{Y}_{i(k)\cdot}, \bar{Y}_{i(k)\cdot'} \mid k, \forall j, \forall j'), \tag{E3}$$

and, the person-level absolute reliability coefficient for average scores is:

$$\rho^{\mathrm{S}}_{\mathrm{abs}} = \mathrm{Cor}(\bar{Y}_{i(k)\cdot}, \bar{Y}_{i(k)\cdot'} \mid k). \tag{E4}$$

On the other hand, at the cluster level, there are two viewpoints to define cluster-level reliability in Section Cluster-Level Reliability that need to be distinguished. In short, manifest cluster-level reliability considers the person variance as a source of unsystematic error variance, but latent cluster-level reliability does not. First, the manifest cluster-level reliability focuses on the consistency between a cluster's average score, $\bar{Y}_{\cdot(k)\cdot} = \frac{1}{n_S n_R} \sum_{i=1}^{n_S} \sum_{j=1}^{n_R} Y_{i(k)j}$, over persons $i \in \{1, \dots, n_S\}$ and raters $j \in \{1, \dots, n_R\}$ and another cluster's average score, $\bar{Y}_{\cdot'(k)\cdot'} = \frac{1}{n_S n_R} \sum_{i'=1}^{n_S} \sum_{j'=1}^{n_R} Y_{i'(k)j'}$, over different samples persons $i' \in \{1, \dots, n_S\}'$ and raters $j' \in \{1, \dots, n_R\}'$ but both sets of persons are coming from the same cluster k. Thus, the person variance is involved because of the random sample of persons from the cluster. Following the correlation approach (Equations 15 and 16 for single scores), the manifest cluster-level relative and

absolute reliability for average scores are:

$$\rho_{\text{rel}}^{\text{C, man}} = \text{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot'(k)\cdot'} \mid \forall j, \forall j'), \tag{E5}$$

and

$$\rho_{\text{abs}}^{\text{C, man}} = \text{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot'(k)\cdot'}), \tag{E6}$$

respectively. Compared to the absolute reliability, the relative one should additionally condition on two sets of raters.

Second, the latent cluster-level reliability, however, focuses on the consistency between two cluster's average scores $\bar{Y}_{\cdot(k)\cdot} = \frac{1}{n_S n_R} \sum_{i=1}^{n_S} \sum_{j=1}^{n_R} Y_{i(k)j}$ and $\bar{Y}_{\cdot(k)\cdot'} = \frac{1}{n_S n_R} \sum_{i=1}^{n_S} \sum_{j'=1}^{n_R} Y_{i(k)j'}$. Both of them average over the same selected set of persons $i \in \{1, \ldots, n_S\}$ for each cluster k but over different sets of raters $j \in \{1, \ldots, n_R\}$ and $j' \in \{1, \ldots, n_R\}'$.

Compared to the cluster's average scores used in manifest cluster-level reliability, where a new sample of persons is sampled at the two measurement occasions from the cluster, the selected persons remain the same at the two measurement occasions and as if they could show up in different clusters in latent cluster-level reliability. In addition, to eliminate the person variation, we further condition on these selected persons in the correlation equation. Thus, following the correlation approach (Equations 17 and 18 for single scores), the latent cluster-level relative and absolute reliability for average scores are:

$$\rho_{\text{rel}}^{\text{C, lat}} = \text{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot(k)\cdot'} \mid \forall i(k), \forall j, \forall j'), \tag{E7}$$

and

$$\rho_{\text{abs}}^{\text{C, lat}} = \text{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot(k)\cdot'} \mid \forall i(k)), \tag{E8}$$

respectively.

Table E2 summarizes level-specific reliability for single scores as well as average scores for the above example (the *Person(Cluster)* × *Rater* design). Furthermore, it also collects the results in the *Person(Cluster)* × *Rater(Cluster)* design and the *Rater(Person(Cluster))* design. The difference between the reliability of single scores and average scores is that the variances used for average scores are divided by the

number of replications. The structure of the reliability coefficient (the variance components involved in the coefficient) does not change in both cases.

Table E2

Person- and Cluster-level Reliability Coefficients for Single and Average Scores in Nested Designs

Reliability	Single score	Average score
	Person(Cluster) × Rater	
$\rho^{\mathrm{S}}_{\mathrm{rel}}$	$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k, j, j') = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{i(k)\cdot}, \bar{Y}_{i(k)\cdot'} \mid k, \forall j, \forall j') = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\frac{\sigma^2_{\epsilon}}{n_R}}$
$\rho^{\mathrm{S}}_{\mathrm{abs}}$	$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{R}+\sigma^2_{CR}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{i(k)\cdot}, \bar{Y}_{i(k)\cdot'} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\frac{\sigma^2_{R}}{n_R}+\frac{\sigma^2_{CR}}{n_R}+\frac{\sigma^2_{\epsilon}}{n_R}}$
$\rho^{\mathrm{C,\ man}}_{\mathrm{rel}}$	$\mathrm{Cor}(Y_{i(k)j}, Y_{i'(k)j'} \mid j, j') = \frac{\sigma^2_{C}}{\sigma^2_{C}+\sigma^2_{S}+\sigma^2_{CR}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot'(k)\cdot'} \mid \forall j, \forall j') = \frac{\sigma^2_{C}}{\sigma^2_{C}+\frac{\sigma^2_{S}}{n_S}+\frac{\sigma^2_{CR}}{n_R}+\frac{\sigma^2_{\epsilon}}{n_S n_R}}$
$\rho^{\mathrm{C,\ man}}_{\mathrm{abs}}$	$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'})) = \frac{\sigma^2_{C}}{\sigma^2_{C}+\sigma^2_{R}+\sigma^2_{CR}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot(k)\cdot'} \mid \forall i(k)) = \frac{\sigma^2_{C}}{\sigma^2_{C}+\frac{\sigma^2_{S}}{n_S}+\frac{\sigma^2_{R}}{n_R}+\frac{\sigma^2_{CR}}{n_R}+\frac{\sigma^2_{\epsilon}}{n_S n_R}}$
$\rho^{\mathrm{C,\ lat}}_{\mathrm{rel}}$	$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k), j, j') = \frac{\sigma^2_{C}}{\sigma^2_{C}+\sigma^2_{CR}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot(k)\cdot'} \mid \forall i(k), \forall j, \forall j') = \frac{\sigma^2_{C}}{\sigma^2_{C}+\frac{\sigma^2_{CR}}{n_R}+\frac{\sigma^2_{\epsilon}}{n_S n_R}}$
$\rho^{\mathrm{C,\ lat}}_{\mathrm{abs}}$	$\mathrm{Cor}(Y_{i(k)j}, Y_{i(k)j'} \mid i(k)) = \frac{\sigma^2_{C}}{\sigma^2_{C}+\sigma^2_{R}+\sigma^2_{CR}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{\cdot(k)\cdot}, \bar{Y}_{\cdot(k)\cdot'} \mid \forall i(k)) = \frac{\sigma^2_{C}}{\sigma^2_{C}+\frac{\sigma^2_{R}}{n_R}+\frac{\sigma^2_{CR}}{n_R}+\frac{\sigma^2_{\epsilon}}{n_S n_R}}$
	Person(Cluster) × Rater(Cluster)	
$\rho^{\mathrm{S}}_{\mathrm{rel}}$	$\mathrm{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid k, j(k), j'(k)) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{i(k)\cdot(k)}, \bar{Y}_{i(k)\cdot'(k)} \mid k, \forall j(k), \forall j'(k)) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\frac{\sigma^2_{\epsilon}}{n_R}}$
$\rho^{\mathrm{S}}_{\mathrm{abs}}$	$\mathrm{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_{R(C)}+\sigma^2_{\epsilon}}$	$\mathrm{Cor}(\bar{Y}_{i(k)\cdot(k)}, \bar{Y}_{i(k)\cdot'(k)} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\frac{\sigma^2_{R(C)}}{n_R}+\frac{\sigma^2_{\epsilon}}{n_R}}$

Reliability	Single score	Average score
$\rho^{\text{C, man}}_{\text{rel}}$	$\text{Cor}(Y_{i(k)j(k)}, Y_{i'(k)j'(k)} \mid j(k), j'(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_\epsilon}$	$\text{Cor}(\bar{Y}_{\cdot(k)\cdot(k)}, \bar{Y}_{\cdot'(k)\cdot'(k)} \mid \forall j(k), \forall j'(k)) = \frac{\sigma^2_C}{\sigma^2_C+\frac{\sigma^2_{S(C)}}{n_S}\frac{\sigma^2_\epsilon}{n_S n_R}}$
$\rho^{\text{C, man}}_{\text{abs}}$	$\text{Cor}(Y_{i(k)j(k)}, Y_{i'(k)j'(k)}) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_{R(C)}+\sigma^2_\epsilon}$	$\text{Cor}(\bar{Y}_{\cdot(k)\cdot(k)}, \bar{Y}_{\cdot(k)\cdot'(k)}) = \frac{\sigma^2_C}{\sigma^2_C+\frac{\sigma^2_{S(C)}}{n_S}+\frac{\sigma^2_{R(C)}}{n_R}+\frac{\sigma^2_\epsilon}{n_S n_R}}$
$\rho^{\text{C, lat}}_{\text{rel}}$	$\text{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid i(k), j(k), j'(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_\epsilon}$	$\text{Cor}(\bar{Y}_{\cdot(k)\cdot(k)}, \bar{Y}_{\cdot(k)\cdot'(k)} \mid \forall i(k), \forall j(k), \forall j'(k)) = \frac{\sigma^2_C}{\sigma^2_C+\frac{\sigma^2_\epsilon}{n_S n_R}}$
$\rho^{\text{C, lat}}_{\text{abs}}$	$\text{Cor}(Y_{i(k)j(k)}, Y_{i(k)j'(k)} \mid i(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{R(C)}+\sigma^2_\epsilon}$	$\text{Cor}(\bar{Y}_{\cdot(k)\cdot(k)}, \bar{Y}_{\cdot(k)\cdot'(k)} \mid \forall i(k)) = \frac{\sigma^2_C}{\sigma^2_C+\frac{\sigma^2_{R(C)}}{n_R}+\frac{\sigma^2_\epsilon}{n_S n_R}}$
	Rater(Person(Cluster))	
ρ^{S}	$\text{Cor}(Y_{j(i(k))}, Y_{j'(i(k))} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\sigma^2_\epsilon}$	$\text{Cor}(\bar{Y}_{\cdot(i(k))}, \bar{Y}_{\cdot'(i(k))} \mid k) = \frac{\sigma^2_{S(C)}}{\sigma^2_{S(C)}+\frac{\sigma^2_\epsilon}{n_R}}$
$\rho^{\text{C, man}}$	$\text{Cor}(Y_{j(i(k))}, Y_{j'(i'(k))}) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_{S(C)}+\sigma^2_\epsilon}$	$\text{Cor}(\bar{Y}_{\cdot(\cdot(k))}, \bar{Y}_{\cdot'(\cdot'(k))}) = \frac{\sigma^2_C}{\sigma^2_C+\frac{\sigma^2_S}{n_S}+\frac{\sigma^2_\epsilon}{n_S n_R}}$
$\rho^{\text{C, lat}}$	$\text{Cor}(Y_{j(i(k))}, Y_{j'(i(k))} \mid i(k)) = \frac{\sigma^2_C}{\sigma^2_C+\sigma^2_\epsilon}$	$\text{Cor}(\bar{Y}_{\cdot(\cdot(k))}, \bar{Y}_{\cdot'(\cdot(k))} \mid \forall i(k)) = \frac{\sigma^2_C}{\sigma^2_C+\frac{\sigma^2_\epsilon}{n_S n_R}}$

Note: The symbol ρ represents reliability. The superscript S represents person-level, C, man represents cluster-level for manifest scores, and C, lat represents cluster-level for latent cluster scores. The subscript rel represents relative decisions, and abs represents absolute decisions.