Updated Guidelines on Selecting an ICC for Interrater Reliability: With Applications to Incomplete Observational Designs

ten Hove, D.; Jorgensen, T.D.; van der Ark, L.A.

DOI
10.1037/met0000516

Publication date
2022

Document Version
Author accepted manuscript

Published in
Psychological Methods

Link to publication

Citation for published version (APA):
ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Updated Guidelines on Selecting an ICC for Interrater Reliability: With Applications to Incomplete Observational Designs. *Psychological Methods*. https://doi.org/10.1037/met0000516

Updated Guidelines on Selecting an ICC for Interrater Reliability: With Applications to Incomplete Observational Designs

Debby ten Hove, Terrence D. Jorgensen, and L. Andries van der Ark
Research Institute of Child Development and Education, University of Amsterdam, the Netherlands, P. O. Box 15776, 1001 NG, Amsterdam.
D.tenHove@uva.nl

Updated Guidelines on Selecting an ICC for Interrater Reliability: With Applications to Incomplete Observational Designs

Debby ten Hove, Terrence D. Jorgensen, and L. Andries van der Ark

Research Institute of Child Development and Education

University of Amsterdam

Author Note

Debby ten Hove https://orcid.org/0000-0002-1335-4452 Terrence D. Jorgensen https://orcid.org/0000-0001-5111-6773 L. Andries van der Ark https://orcid.org/0000-0003-3131-7943

Preliminary results of this study have been presented at the 85th annual International Meeting of the Psychometric Society (online), and at the 9th European Congress of Methodology (online). All software code dedicated this paper is available on the Open Science Framework: https://doi.org/10.17605/OSF.IO/8J26U

This work was partly supported by the Dutch Research Council (NWO), project number 016.Veni.195.457, awarded to Dr. Terrence D. Jorgensen.

Correspondence regarding this article should be addressed to Debby ten Hove, Research Institute of Child Development and Education, University of Amsterdam, P. O. Box 15776, 1001 NG Amsterdam, The Netherlands. Email: D.tenHove@uva.nl

Abstract

Several intraclass correlation coefficients (ICCs) are available to assess the interrater reliability (IRR) of observational measurements. Selecting an ICC is complicated, and existing guidelines have three major limitations. First, they do not discuss incomplete designs, in which raters partially vary across subjects. Second, they provide no coherent perspective on the error variance in an ICC, clouding the choice between the available coefficients. Third, the distinction between fixed or random raters is often misunderstood. Based on Generalizability theory (GT), we provide updated guidelines on selecting an ICC for IRR, which are applicable to both complete and incomplete observational designs. We challenge conventional wisdom about ICCs for IRR by claiming that raters should seldom (if ever) be considered fixed. Also, we clarify how to interpret ICCs in the case of unbalanced and incomplete designs. We explain four choices a researcher needs to make when selecting an ICC for IRR, and guide researchers through these choices by means of a flowchart, which we apply to three empirical examples from clinical and developmental domains. In the discussion, we provide guidance in reporting, interpreting, and estimating ICCs, and propose future directions for research into the ICCs for IRR.

Keywords: Generalizability theory, incomplete designs, interrater reliability, intraclass correlation coefficients, observational research

Updated Guidelines on Selecting an ICC for Interrater Reliability: With Applications to Incomplete Observational Designs

Observational studies use raters to obtain information about attributes of subjects. For example, in an observational study by Zee et al. (2020) researchers (raters) assessed students' (subjects) emotional distance from their teachers (attribute). In subsequent analyses, emotional distance was used as a predictor of classroom related outcomes. The variance in the assessments' scores—generically called ratings—is due to both subject differences and rater differences. Researchers are typically interested in subject differences, whereas rater differences are typically considered noise. The interrater reliability (IRR) provides information about the ability to differentiate between subjects based on the ratings, and bounds the precision and validity of ratings (cf. Lord & Novick, 1968, p. 72). Using ratings with low IRR may result in biased estimates, loss of power in subsequent statistical analyses, and incorrect decisions in diagnostic settings. IRR is thus imperative in observational research.

Reasons to Choose ICCs for IRR

Reasons for choosing between IRR coefficients are based on data characteristics such as measurement level and number of raters, but many coefficients are available for the same data characteristics. Depending on the choice of an IRR coefficient, the qualitative label of the IRR for a single dataset can range from poor to almost perfect (Ten Hove et al., 2018). This is confusing and may even lead to so called *researcher degrees of freedom* (Simmons et al., 2011): Researchers who want to defend or propose a specific rating protocol could simply search for the IRR coefficient that provides the most beneficial results. We therefore claim that a unique conceptualization and definition of IRR is required. The *choice* of an IRR coefficient should be guided by the objective of the ratings in the subsequent statistical analysis (c.f., Bartko, 1966; Shrout & Fleiss, 1979). Below, we show that IRR defined within the framework of classical test theory (CTT; see, e.g., Lord

& Novick, 1968) is equivalent to an ICC. Reliability within the framework of CTT is well understood, as its implications for statistical issues such as attenuation of correlation and measurement precision have been thoroughly investigated (e.g., Lord & Novick, 1968, p. 69). Therefore, this definition of IRR is our preferred choice, and an excellent candidate for the unique conceptualization of IRR.

In CTT (Lord & Novick, 1968), an observed score (X) is the sum of a true score (T; the expected score over independent replications) and random measurement error (E); that is, $X = T + E$. Reliability ($\rho_{XX'}$) is the proportion of observed-score variance (σ^2_X) that is due to true-score variance (σ^2_T). As E is random error, uncorrelated with T, reliability is defined as

$$\rho_{XX'} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}. \tag{1}$$

As the true-score variance and error variance can consist of multiple components, we use a more generic notation: σ^2_{true} for the true-score variance in the observation, and σ^2_{error}, for the error variance. Hence, in general, reliability is defined as

$$\rho_{XX'} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{true}} + \sigma^2_{\text{error}}}, \tag{2}$$

and CTT is the special case, where $\sigma^2_{\text{true}} = \sigma^2_T$ and $\sigma^2_{\text{error}} = \sigma^2_E$. For defining IRR, variances σ^2_{true} and σ^2_{error} must be specified in greater detail because error variance can include both random and systematic sources of variance. Generalizability theory (GT; Brennan, 2001a; Cronbach et al., 1963) can be used for further specifications of σ^2_{true} and σ^2_{error}. In GT, $\rho_{XX'}$ in Equation 2 is called a generalizability coefficient or an index of dependability. The true score is named *universe score*, denoted τ, and the error is named error (as in CTT), denoted ϵ. Generalizability coefficients and indices of dependability differ in their definition of ϵ, on which we will elaborate in the section Insights from Generalizability Theory.

Several ICCs for IRR are available, which are all defined as the ratio of true-score or universe-score variance, relative to itself plus the error variance and thus in line with CTT

(Equation 1) and GT:

$$\text{ICC} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{true}} + \sigma^2_{\text{error}}} = \frac{\sigma^2_{\tau}}{\sigma^2_{\tau} + \sigma^2_{\epsilon}}. \tag{3}$$

The true-score or universe-score variance in the ICC for IRR mostly includes the subject variance (see, e.g., McGraw & Wong, 1996). The error variance in the denominator includes the rater-related variance in the ratings. The ICCs for IRR can therefore be interpreted as the proportion of variance in subjects' scores that can be generalized over raters and attributed to differences across subjects. These ICCs are most flexible because they can be estimated for incomplete and complex nested designs for observational data collected through two or more raters for discrete or continuous data (Jorgensen, 2021; Ten Hove et al., 2021).

Current Guidelines for Selecting an ICC

Several papers discussed the choices a researcher has to make when selecting an ICC (e.g., Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). In these papers, selecting an ICC involves the choice between one-way or two-way designs, between an ICC for single or average ratings, between an ICC of interrater agreement or interrater consistency, and between treating rater effects as random or fixed effects. We believe that current guidelines for choosing the correct ICC have three major limitations.

First, current guidelines discuss only one-way and complete two-way observational designs whereas many large observational studies use incomplete two-way observational designs to disseminate the workload across raters (e.g., Majdandžić et al., 2021; Viswesvaran et al., 2005; Yuen et al., 2020; Zee et al., 2020). By lack of an alternative, researchers may treat data that were collected using an incomplete observational design as if these were collected under a one-way or a complete two-way design (e.g., Fürst, 2020), which is inappropriate and goes against Bartko's (1966, p. 3) advice that "use of the ICC however, should be restricted by the underlying model which most adequately describes the experimental situation".

Second, current guidelines do not provide a clear perspective on the error variance in an ICC for IRR. Different ICCs include different rater-related variance components in the denominator as error, yet the reasoning behind these differences is rarely elucidated. A clear perspective on the error variance would guide the choice between an ICC of interrater agreement or interrater consistency. This is especially true for incomplete designs, in which the correct definition of the error variance is essential for computing the IRR. Brennan (2001a) discussed the implications of missing data for both generalizability coefficients and indices of dependability (hence ICCs), but these insights have been largely neglected in the IRR and ICC literature. As one of the few studies that considered missing ratings, Putka et al. (2008) proposed estimating ICCs under incomplete designs, but guidelines for selecting the correct error variance terms under incomplete designs are currently unavailable.

Third, current guidelines do not provide enough information for an informative choice between an ICC for random raters and an ICC for fixed raters. Shrout and Fleiss (1979) proposed an ICC for agreement given random effects and an ICC of consistency given fixed effects. Although McGraw and Wong (1996) extended their work with ICCs for agreement given fixed effects and ICCs for consistency given random effects, several methodological papers still confuse the topic of fixed versus random raters with the issue of interrater agreement and interrater consistency (e.g., De Vet et al., 2017; Revelle & Condon, 2019). From our experience in consultation, we learned that these limitations hinder researchers to choose the appropriate ICC.

Contributions of this Paper

This paper extends current guidelines and challenges the conventional wisdom about which ICC should be selected to estimate IRR, by synthesizing the literature about GT, ICCs, and missing data. The paper is structured as follows. First, we explain the definition of ICCs for IRR in the one- and two-way designs that have been discussed by

Bartko (1966), Shrout and Fleiss (1979), and McGraw and Wong (1996). Second, for various observational designs and various applications of ratings, we use GT to define σ^2_{true} and σ^2_{error}, the two key elements of an ICC (Equation 3). Using GT and CTT, we also claim that treating raters as fixed is rarely—if ever—appropriate in an IRR study. Third, we use GT to describe the differences between the proposed ICCs for one- and two-way designs, and we present generalizations of these ICCs for incomplete two-way designs. Fourth, we discuss a four-step approach to guide researchers in the process of selecting an ICC, which we visualized in a flowchart. Fifth, we guide researchers through this flowchart using three empirical examples from clinical and developmental domains. We end with a discussion on estimation and reporting of the ICCs and suggestions for further research on ICCs for IRR. Applied researchers who seek guidance in selecting an ICC but are not necessarily interested in the theory behind the ICCs, can turn directly to the section Step-Wise Procedure to Selecting an ICC.

Intraclass Correlation Coefficients for Interrater Reliability

Variance Decomposition

In a complete *two-way design* (Figure 1, Design b), S subjects are each rated by R raters in a completely crossed design. Let y_{sr} be the realization of random variable Y_{sr}, which is the rating of subject s $(s = 1, \ldots, S)$ by rater r $(r = 1, \ldots, R)$ on an attribute. Let μ denote the average rating, let μ_s denote the effect of subject s, let μ_r denote the effect of rater r, and let μ_{sr} be the interaction effect of subject s with rater r. Because each rater rates each subject once in this design, μ_{sr}, also includes the random error (Cronbach et al., 1963). It is assumed that these effects are uncorrelated and that y_{sr} can be decomposed as

$$y_{sr} = \mu + \mu_s + \mu_r + \mu_{sr}. \tag{4}$$

Let $\sigma^2_{y_{sr}}$ denote the variance of the observed ratings, let σ^2_s and σ^2_r denote the variance of the subject effects and rater effects, respectively, and let σ^2_{sr} denote the remaining variance which includes the variance of the subject-by-rater effects and the indistinguishable random-error variance. Because the effects are assumed to be uncorrelated, $\sigma^2_{y_{sr}}$ can be decomposed into orthogonal variance components:

$$\sigma^2_{y_{sr}} = \sigma^2_s + \sigma^2_r + \sigma^2_{sr}. \tag{5}$$

In a *one-way design* (Figure 1, Design a), the raters are nested within the subjects. Let the notation $x : y$ indicate that x is nested in y, then $y_{r:s}$ denotes the rating of subject s by rater r in a one-way design. Because the raters are nested within subjects, rater effect μ_r cannot be disentangled from subject-by-rater effect μ_{sr} (Cronbach et al., 1963). Let $\mu_{r:s}$ denote combined rater and subject-by-rater effects, and random error. The decomposition of $y_{r:s}$ therefore is

$$y_{r:s} = \mu + \mu_s + \mu_{r:s}. \tag{6}$$

Let $\sigma^2_{y_{r:s}}$ denote the variance of the observed ratings in a one-way design, and let $\sigma^2_{r:s}$ denote the variance of the combined rater and subject-by-rater effects, and the random error. Analogous to $\sigma^2_{y_{sr}}$ of the two-way design (Equation 5), $\sigma^2_{y_{r:s}}$ can be decomposed into orthogonal variance components:

$$\sigma^2_{y_{r:s}} = \sigma^2_s + \sigma^2_{r:s}. \tag{7}$$

Intraclass Correlation Coefficients

The variance decompositions in Equations 5 and 7 are used for several varieties of IRR (Table 1). Each definition is an ICC that expresses the proportion of subject variance relative to the subject variance plus error variance (cf. Equation 1).

Two-Way Designs

For a two-way design, an ICC of interrater agreement for the average rating across k raters per subject equals

$$\mathrm{ICC}(\mathrm{A}, k) = \frac{\sigma^2_s}{\sigma^2_s + \frac{\sigma^2_r + \sigma^2_{sr}}{k}}. \tag{8}$$

Other varieties of IRR for two-way designs are determined by removing terms from Equation 8. ICCs can be described by three characteristics: ’agreement versus consistency’, ’average versus single ratings’, and ’random versus fixed raters’ (Bartko, 1966; McGraw & Wong, 1996; Shrout & Fleiss, 1979). This classification results in 8 ICCs for two-way designs (Table 1, top rows). We briefly describe the differences in ICC varieties in Table 1, but delay discussion of the choices between these varieties until the section Implications of GT on ICC Selection.

Agreement versus Consistency. The ICC in Equation 8 is an ICC of interrater agreement—denoted by a capital *A*—because it includes the variance of the rater effects, σ_r^2, in the denominator. An ICC of interrater consistency—denoted by a capital *C*—does not include the variance of the rater effects in the denominator (McGraw & Wong, 1996).

Average versus Single Ratings. The ICC in Equation 8 is an ICC of average ratings—denoted by a *k*—because all rater-related variance components in the denominator are divided by the number of raters per subject, *k*. An ICC of single ratings—denoted by a 1—does not divide the rater-related variance components by the number of raters per subject because *k* would be 1 and disappears from the equation (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

Random versus Fixed Raters. The ICC in Equation 8 is an ICC for random raters. ICCs for fixed raters are defined by subtracting a portion of the subject-by-rater interaction variance from the subject variance in the numerator of the ICC (Table 1; McGraw & Wong, 1996; Shrout & Fleiss, 1979). ICCs for fixed raters therefore require that the subject-by-rater interaction effect is isolated from the random error, so these ICCs can be estimated only if either the subject-by-rater interaction effect is assumed absent or if each rater assessed each subject multiple times so that the interaction effects can be distinguished from random error (McGraw & Wong, 1996).

One-Way Designs

For one-way designs, the most elaborated ICC in Table 1 equals

$$\text{ICC}(k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{r:s}^2}{k}}. \tag{9}$$

As for two-way designs, ICCs for average ratings and ICCs for single ratings can be distinguished, by (not) dividing the rater-related component by k. Because σ_r^2 and σ_{sr}^2 are confounded in $\sigma_{r:s}^2$, agreement and consistency can not be distinguished, and ICCs for fixed raters cannot be defined (Bartko, 1966; McGraw & Wong, 1996; Shrout & Fleiss, 1979). This results in two ICCs for one-way designs (Table 1, bottom row).

Insights from Generalizability Theory

In this section, we first discuss some general GT definitions that are useful for understanding GT for estimating IRR. Next, we discuss five GT topics that are of interest for selecting an ICC for IRR: universe-score variance, absolute and relative error variance, error variance in average ratings, error variance in incomplete designs, and random versus fixed rater effects. We focus on a two-way observational design. In the subsection on error variance in incomplete designs, we explain how a one-way design is a special case of an incomplete two-way design.

Terminology

In contrast to CTT, GT recognizes that a single attribute can be measured under several conditions. A single *observation* (e.g., the rating of a student's emotional distance to their teacher) is therefore considered a sample from a *universe of admissible observations.* The specific characteristics of an observation are called *facets* (e.g., raters, subjects). These facets can be divided into (a) sources of theoretical interest (termed *facets of differentiation*), such as subjects, or (b) sources of nuisance variability (termed *facets of generalization*), such as raters, items, or measurement occasions (Brennan, 2001a;

Cronbach et al., 1963; Vangeneugden et al., 2005). GT considers two types of studies on the relative contribution of facets of generalization to the observed scores: generalizability studies and decision studies. A *generalizability study* investigates the contribution of the facets of generalization and differentiation to the observed variance, by decomposing observations into a grand mean and the main and interaction effects of all identified facets. Considering a standard observational study, with subjects and raters as the facets, this results in the decomposed effects in Equation 4. The variance of multiple observations can then be decomposed into the variance of each of the effects in the observations, as in Equation 5. These variance components can be classified as *universe-score variance* (σ^2_τ; Equation 3) and *error variance* (σ^2_ϵ; Equation 3). A *decision study* investigates how the error-variance in a measurement can be reduced by changing the research design. This typically involves the investigation of various sample sizes for the facets of generalization. These effects can be investigated using so-called *generalizability coefficients* and *indices of dependability.* Both types of coefficients are defined as the proportion of universe-score variance to itself plus the error variance (Equation 3), but they differ in the definition of the error term—just like ICCs for IRR (Table 1). Generalizability coefficients include *relative error* and indices of dependability include *absolute error* as the error term.

Universe-Score Variance

According to CTT, the true score is a subject's expected observed score over infinite repeated measurements (Lord & Novick, 1968, p. 30). The true score cannot be measured directly. Therefore, the true score is typically estimated as a subject's average score over repeated observations. Universe score τ is a generalization of true score T from CTT. Whereas T is the expected value over a single facet (i.e., replications), τ can be the expected value over any number of facets (e.g., replications, raters). Let subjects be the facet of differentiation. A subject's expected score over the universe of admissible observations—here, over all possible raters—is this subject's *universe score*, τ_s, that is

$$\tau_s = \mu + \mu_s. \tag{10}$$

Because μ has no variance, the universe-score variance is the variance in all subjects' universe scores, that is,

$$\sigma^2_\tau = \sigma^2_s. \tag{11}$$

The universe-score variance is thus identical to the subject variance, σ^2_s. Note though, that in the situation of fixed raters, σ^2_s may be biased, which we discuss in section Random or Fixed Rater Facets.

Absolute or Relative Error Variance

GT distinguishes absolute error and relative error. The absolute error in the observation of subject s by rater r ($\epsilon.abs_{sr}$) is the observed score (y_{sr}) minus the universe score (τ_s); that is, $\epsilon.abs_{sr} = y_{sr} - \tau_s$. Substituting y_{sr} by $\mu + \mu_s + \mu_r + \mu_{sr}$ (Equation 4), and substituting τ_s by $\mu + \mu_s$ (Equation 10) yields $\epsilon.abs_{sr} = y_{sr} - \tau_s = \mu + \mu_s + \mu_r + \mu_{sr} - \mu - \mu_s = \mu_r + \mu_{sr}$. The variance of the absolute error then equals

$$\sigma^2_{\epsilon.abs} = \sigma^2_r + \sigma^2_{sr}. \tag{12}$$

Absolute error is of interest when measurements are given absolute interpretations (Brennan, 2001a, p. 13). Examples include diagnostic tests or school exams.

The relative error in the observation of subject s by rater r ($\epsilon.rel_{sr}$) is the observed deviation score ($y^*_{sr} = \ y_{sr} - \mu_r$) minus the universe score (τ_s); that is, $\epsilon.rel_{sr} = \ y^*_{sr} - \tau_s$. Observed deviation scores are the difference between a subject's universe score and the average observed score (the latter of which is the grand mean). In a complete two-way design, all subjects are assessed by the same rater(s), so rater effects do not contribute to the deviation score. Again, substituting y_{sr} by $\mu + \mu_s + \mu_r + \mu_{sr}$ (Equation 4) produces $y^*_{sr} = \mu + \mu_s + \mu_{sr}$, and substituting τ_s by $\mu + \mu_s$ (Equation 10) results in

$\epsilon.rel_{sr} = \ y^*_{sr} - \tau = \mu + \mu_s + \mu_{sr} - \mu - \mu_s = \mu_{sr}$. The variance of the relative error then equals

$$\sigma^2_{\epsilon.rel} = \sigma^2_{sr}. \tag{13}$$

Relative error is the GT analogue of the measurement error in CTT, and is appropriate when measurements are used for relative comparisons, such as correlational studies, regressions models, and comparing groups (Brennan, 2001a, p. 13).

Error Variance in Average Ratings

When subjects are assessed by multiple raters, decisions about subjects' attributes are often based on the average of these ratings. Similarly, average ratings of subjects are often used as subjects' scores in subsequent statistical analyses. The error variance is reduced proportionally to the number of raters over which the subjects' observed scores are averaged (Brennan, 2001a, p. 31). This is similar to dividing the sample variance by N to obtain the variance of the sample mean (or its square-root: $SE_{mean} = SD/\sqrt{N}$): Sample means vary less than individual ratings, proportional to the number of ratings that are averaged. We use ε rather than ϵ (e.g., Equation 12) to indicate that the interest is the error variance in average ratings per subject. Absolute error variance in average ratings is then defined as,

$$\sigma^2_{\varepsilon.abs} = \frac{\sigma^2_r + \sigma^2_{sr}}{k}, \tag{14}$$

and relative error variance in average ratings is defined as,

$$\sigma^2_{\varepsilon.rel} = \sigma^2_{sr}/k. \tag{15}$$

Equations 14 and 15 readily generalize to single ratings where $k = 1$ and disappears from the equations. For simplicity, we further ignore Equations 12 and 13, and refer to Equations 14 and 15 to discuss the error variance in both single and average ratings.

Error Variance in Incomplete and Unbalanced Designs

Equations 14 and 15 express the error variance in a complete two-way design (Figure 1, Design b) in which each subject is assessed by the same set of raters. In an *incomplete two-way design* (Figure 1, Design c), raters are crossed with subjects, but assess different subsets of subjects. A one-way design (Figure 1, Design a) is a special case of an incomplete two-way design, without any overlap of raters across subjects. The definition of absolute or relative error variance in incomplete designs is more complicated than for complete two-way designs for two reasons. First, the design may be *unbalanced*, meaning that the number of raters may vary across subjects. Therefore, the definition of error variance should allow for differing numbers of raters per subject. Second, the relative differences across subjects' observed scores are not only dependent on the subject effects and the subject-by-rater interaction effects, but also influenced by the rater effects. One subject may be assessed by raters who, on average, provide relatively high scores, whereas another subject may be assessed by raters who, on average, provide relatively low scores. Therefore, the relative error variance includes (a portion of) rater variance. First, we discuss the definition of absolute and relative error variance in incomplete—potentially unbalanced—two-way designs. Second, we discuss how this generalizes to the special case of—potentially unbalanced—one-way designs.

Incomplete Two-Way Designs

The differing numbers of raters across subjects in an incomplete design can be accounted for by substituting k in Equations 14 and 15 by $\hat{k}$, the harmonic mean number of raters per subject (Brennan, 2001a, p. 229). Let k_s $(s = 1, \ldots, S)$ be the number of raters that assessed subject s, then

$$\hat{k} = \left(\frac{k_1^{-1} + k_2^{-1} + \ldots + k_S^{-1}}{S} \right)^{-1}. \tag{16}$$

As is the case for the absolute error variance in complete designs (Equation 14), the relative error variance for incomplete designs includes both the rater variance, σ_r^2, and the subject-by-rater interaction variance, σ^2_{sr}. However, the relative error variance in incomplete designs includes only a proportion of the rater variance, which size depends on the proportion of non-overlapping raters across subjects (Brennan, 2001a, p. 236). Let k_s and $k_{s'}$ be the number of raters who rated subject s and subject s', where $s \neq s'$, and let $k_{s,s'}$ be the number of raters that subject s and s' share. Then, q, the proportion of non-overlap between raters across subjects equals

$$q = \frac{1}{\hat{k}} - \frac{\sum_s \sum_{s'} \frac{k_{s,s'}}{k_s k_{s'}}}{S(S-1)} \tag{17}$$

(Brennan, 2001a; Putka et al., 2008). Following Brennan (2001a, pp. 229–236), in case of incomplete two-way designs, the estimated variance of the absolute error ($\widehat{\sigma}^2_{\varepsilon.abs}$) and relative error ($\widehat{\sigma}^2_{\varepsilon.rel}$) equal

$$\widehat{\sigma}^2_{\varepsilon.abs} = \sigma^2_r/\hat{k} + \sigma^2_{sr}/\hat{k}, \tag{18}$$

and

$$\widehat{\sigma}^2_{\varepsilon.rel} = q\sigma^2_r + \sigma^2_{sr}/\hat{k}, \tag{19}$$

respectively. Note that, in a complete two-way design, $\frac{\sum_s \sum_{s'} \frac{k_{s,s'}}{k_s k_{s'}}}{S(S-1)} = \frac{1}{\hat{k}}$, resulting in $q = \frac{1}{\hat{k}} - \frac{\sum_s \sum_{s'} \frac{k_{s,s'}}{k_s k_{s'}}}{S(S-1)} = \frac{1}{\hat{k}} - \frac{1}{\hat{k}} = 0$, and also $\hat{k} = k$. Hence, for $q = 0$ and $\hat{k} = k$, Equation 18 and Equation 19 reduce to Equation 14 and Equation 15, respectively. This explains why σ^2_r does not contribute to the relative error in a two-way design (Equation 15).

One-Way Designs

Distinguishing between absolute and relative error variance requires differentiation between σ^2_r and σ^2_{sr}, which is not possible for one-way designs; σ^2_r and σ^2_{sr} are confounded in $\sigma^2_{r:s}$ (Equation 7). One-way designs are a special case of incomplete two-way designs, without any overlap across raters ($k_{s,s'} = 0$ for any pair of subjects), resulting in

$$q = \frac{1}{\hat{k}} - \frac{\sum_s \sum_{s'} \frac{0}{k_s k_{s'}}}{S(S-1)} = \frac{1}{\hat{k}}. \tag{20}$$

In the unbalanced two-way design, the absolute-error variance was defined as $\sigma^2_{\epsilon.abs} = \frac{\sigma^2_r}{\hat{k}} + \frac{\sigma^2_{sr}}{\hat{k}}$. As in the one-way design $q = \frac{1}{\hat{k}}$ (Equation 20), the absolute-error variance can be written as $\sigma^2_{\epsilon.abs} = q\sigma^2_r + \frac{\sigma^2_{sr}}{\hat{k}}$. Finally, as the right-hand side of this equation equals the relative-error variance (Equation 19), it is shown that in the one-way design $\sigma^2_{\epsilon.abs} = \sigma^2_{\epsilon.rel}$. In a one-way design, absolute error variance and relative error variance are thus identical. For balanced one-way designs, $\hat{k} = k$, and therefore $\sigma^2_{\varepsilon.abs}$ reduces to $\sigma^2_{r:s}/k$.

Random or Fixed Rater Facets

A rater facet is *random* if the raters in the observational study are a random sample of all possible raters. A rater facet is *fixed* if the observational study includes all possible raters and no variation in the composition of the group of raters in possible (e.g., Brennan, 2001a, p. 14).

Treating raters as fixed or random has implications for the numerator of the ICC (see Table 1). When raters are treated as random, the universe-score variance equals the subject variance (Equation 11). If raters are fixed, the following problem occurs. Let μ_{s^*r} and μ_{s^+r} denote the subject-by-interaction effects of subjects s^* and s^+, respectively. All effects should be uncorrelated in the population. However, as $\sum_r \mu_{sr} = 0$ by definition, this sum constraint causes a spurious negative correlation between μ_{s^*r} and μ_{s^+r} (e.g., Bartko, 1974; Shrout & Fleiss, 1979). The covariance between μ_{s^*r} and μ_{s^+r} has an expected bias of $-1/(k-1)$. Using these biased covariances for estimating the main-subject variance (σ^2_s), results in biased estimate of σ^2_s, denoted ς^2_s. For designs where σ^2_{sr} can be estimated separately from random error, Shavelson et al. (1989; also see McGraw and Wong, 1996) proposed to use $\hat{\sigma}^2_s = \hat{\varsigma}^2_s - \hat{\sigma}^2_{sr}/(k-1)$ to obtained an unbiased estimate of σ^2_s. For designs where σ^2_{sr} cannot be estimated separately, the bias reduction can be applied only if one can safely assume that $\sigma^2_{sr} = 0$. In addition to the negativity bias, correlations and covariances that are spurious due to a sum constraint have several other disadvantages which makes them unattractive for practical purposes (see, e.g., Aitchison, 1986/2003, pp. 52–58).

We argue that the ICC is not an appropriate statistic to investigate the IRR for fixed raters for three reasons. First, the raters in the study are seldom the entire population of potential raters, which is the first prerequisite for selecting a fixed-rater ICC. Raters are typically a sample from a larger pool of potential raters (e.g., other [trainable] researchers, research assistants, trainers, or teachers). Second, few observational studies have the funding or time to let all raters assess all subjects, which is the second prerequisite for selecting a fixed-rater ICC. Third, ICCs allow a generalization from the sample to the population. If the raters in the study are the entire population of raters (a prerequisite of using fixed-raters ICCs), it is unclear what the generalization means.

We discuss four situations in which rater effects have been considered fixed. (1) *Studies not pertaining to IRR.* Molenaar et al. (2021) modelled rater differences but were not interested in the variability of these effects for an IRR study. In this context, treating rater effects either as random or fixed has no influence on the parameters of interest; however, treating raters effects as fixed may simplify the estimation procedure. (2) *Convenience samples.* Several researchers (e.g., Kivisalu et al., 2016) estimated ICCs for fixed raters because the raters in their study are a convenience sample. We argue that rater effects in convenience samples should not be treated as fixed because convenience samples are also samples from a wider population of potential raters. For example, if a study containing a convenience sample of raters were to be replicated, it is unlikely that the same convenience sample of raters would be used to rate the new sample of subjects. Although rater effects in convenience samples should not be treated as fixed, we stress that convenience samples are not simple random samples either, which can prevent generalizing to the population of interest—the same limitations faced when using convenience samples of subjects (e.g., sampling from a subject pool of undergraduate psychology students). Hence, we advise to avoid using convenience samples for IRR (or any human-subjects) studies. (3) *Raters are irreplaceable.* One may be tempted to treat rater effects as fixed because the raters are truly irreplaceable, such as two parents rating their own child.

However, when parents assess only their own child, there is no intention to generalize these observations to a wider population of interchangable raters, which is the purpose of IRR. (4) *Raters are rarely replaced.* If the composition of a sample of raters is unlikely to change, or if change in team composition goes slow, such as the judges of the Supreme Court of the United States (who may act as raters assessing a legal case), we argue that the rater effects should be treated as random as the team composition does change, and other candidate judges are available.

Implications of GT on ICC Selection

Using the insights from GT, we derive a general definition of the ICC for IRR that enables updating the guidelines for selecting an ICC. For selecting the ICC, a researcher has to make four decisions that are visualized in a flowchart (Figure 2).

General Definition of the ICC

Under the assumption of randomly selected raters, the numerator of the ICC, σ^2_τ, equals the subject variance σ^2_s. For any facet of differentiation, an ICC for IRR can therefore be defined by expressing the proportion of subject variance that is independent of rater effects.

The description of the absolute and relative error variance in Equations 14 and 15, explains the difference between ICCs of interrater agreement and ICCs of interrater consistency (Table 1), respectively. ICCs of interrater agreement include absolute error variance in the denominator, that is,

$$\text{ICC(A, } k) = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\varepsilon} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{\varepsilon.abs}}. \tag{21}$$

An ICC of interrater agreement assesses the IRR of subjects' absolute scores, and expresses the degree to which the observed scores are dependent on raters. ICCs of interrater consistency include relative error variance in the denominator, that is

$$\text{ICC(C,}k) = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\varepsilon} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{\varepsilon.rel}}. \tag{22}$$

An ICC of interrater consistency assesses the IRR of the observed differences between subjects, and expresses the degree to which observed differences between subjects can be generalized across raters. The ICC of interrater consistency is similar to the CTT definition of reliability, because both can be interpreted as the correlation between two independent observations of the same subjects by two different raters, or two groups of raters in the case of ICCs for average ratings (e.g. Lord & Novick, 1968; Shrout & Fleiss, 1979).

The description of the error in single and average ratings of Equations 14 and 15 explains why ICCs for average ratings divide the error variance in the denominator by the number of raters per subject, and why ICCs for single ratings do not. Average ratings vary less than single ratings, and the rater-related error reduces proportional to the number of ratings over which subjects' scores are averaged.

The description of the error in incomplete designs of Equations 18 and 19 explains how ICCs can accommodate incomplete designs. By means of $\hat{k}$, ICCs can accommodate differing numbers of raters per subject, and ICCs of interrater consistency include a portion of main-rater variance as error to account for non-overlapping raters. Hence, ICCs of interrater consistency gradually change into ICCs of interrater agreement with increasing non-overlap of raters, and the two ICCs are identical in the specific case of a one-way design. Table 2 shows the definitions of error variances for the discussed observational designs.

Step-Wise Procedure to Selecting an ICC

From the previous sections, we identified four steps that researchers need to take when selecting the appropriate definition of an ICC. Because for one-way and two-way designs with random raters, the universe-score variance always equals σ_s^2, all four steps concern the definition of the error variance in the ICCs.

For each step, we also discuss further considerations such as implications of the choices concerning ICC estimation and the magnitude of the ICC. Figure 2 provides a flowchart to guide researchers through these steps.

Step 1: Is the Observational Design Crossed or Nested?

The observational design is *crossed* if each rater assessed multiple subjects and each subject is assessed by at least two raters. The observational design is *nested* if each rater assesses a single subject and each subject is assessed by at least two raters.

Further Considerations

- Whether the design is crossed or nested pertains only to the observations that are used to estimate the IRR and does not pertain to the way an observational instrument is used in practice or in subsequent analyses (see, e.g., Example 1 and Example 3, below).

- Estimating the IRR requires at least two raters per subject. Otherwise, the rater-related variance components cannot be distinguished from subject variance.

- For crossed observational designs, rater variance and subject-by-rater variance can be distinguished (Equation 5). Hence, ICCs for interrater agreement and ICCs for interrater consistency can be distinguished. For nested observational designs, rater variance and subject-by-rater variance cannot be distinguished (Equation 7). Hence, ICCs for interrater agreement and ICCs for interrater consistency cannot be distinguished either.

Step 2: Are Ratings used for Absolute or Relative Inferences?

Absolute inferences are inferences for which the absolute score of the subject is of interest. Absolute inferences apply if the ratings are compared to a fixed criterion; for example, when a specific grade or score needs to be obtained to pass a test or to qualify for treatment. *Relative inferences* are inferences for which the relative position of the subject is of interest. Relative inferences apply for most statistical analyses—including correlations, regressions, factor analysis, ANOVA, and generalized linear models—and for ranking subjects.

Further Considerations

- In practice, ICCs for absolute inferences (i.e., ICCs of interrater agreement), result in lower ICC values than ICCs for relative inferences (i.e., ICCs of interrater consistency). Only in the hypothetical case that there are no main rater effects ($\sigma_r^2 = 0$) are the two equal.

Step 3: Are Single or Average Ratings Used?

Whether ratings are single or average pertains to the situation after the IRR has been estimated; for example, when an observational instrument is used in practice or in scientific studies. *Single* ratings should be selected when in practice or in subsequent analyses, the ratings are provided by a single rater. *Average ratings* should be selected when in practice or in subsequent analyses, the ratings are provided by taking the average rating of multiple raters.

Further Considerations

- When the number of raters per subject (k or $\hat{k}$) increases, so does the ICC. A single rating ($k = 1$) thus yields the lowest possible IRR, which increases with the number of raters. Researchers should thus balance the cost of additional raters against the desired level of IRR.

- Even if in practice or in scientific studies, scores are based on single ratings, multiple ratings per subject are required to assess the IRR (see Step 1). A pilot study should then be conducted to inspect the IRR of the measures, for example, using multiple ratings of a subset of the subjects (see Examples 1 and 3).

- A minimum of 2 raters per subject is required to estimate an ICC, though at least 3 raters are required to yield accurately estimated ICCs (see, e.g., Briesch et al., 2014; Ten Hove et al., 2020).

Step 4: Are Observations Complete and/or Raters Balanced?

If the observational design is crossed (Step 1) and relative scores are of interest (Step 2), it should be determined whether the observational design is *complete* or *incomplete*. The design is *complete* when each rater assesses each subject (Figure 1, design b). The design is *incomplete* when one of more ratings are missing (i.e., the raters vary across subjects; Figure 1, design c).

When the observational design is crossed (Step 1) and the absolute scores are of interest (Step 2), or when the observational design is nested (Step 1), it should be determined whether the number of raters per subject is *balanced* or *unbalanced*. The design is *balanced* if the number of raters is equal for all subjects. The design is *unbalanced* if the number of raters varies across the subjects. Note that whether the design is balanced or unbalanced is also relevant for incomplete crossed designs. However, $\hat{k}$ in the definition of the error term for an incomplete two-way design (Equations 18 and 19) already accommodates unbalanced designs.

Further Considerations

- Incomplete designs are a pragmatic approach to distribute workload across raters and result in more accurately estimated ICCs (Ten Hove et al., 2021). When everything else remains constant, the IRR of interrater consistency decreases with decreasing overlap across raters. Researchers should thus balance the cost of additional raters or higher overlap of raters across subjects with the increase of the IRR.

- A preliminary study assessing the IRR can be helpful for observational design considerations, such as how many raters should observe each subject. Researchers can consider different hypothetical values for q and $\hat{k}$ or k, and the estimated magnitude of the variance components to see which observational design yields a sufficient IRR.

Emperical Examples

Example 1: Communication Skills of Clinicians in Training

Yuen et al. (2020) developed an instrument to assess advance care planning (ACP) communication skills (attribute) of 29 clinicians in training (subjects). To validate the instrument, a sample of 6 raters assessed clinicians' ACP communication skills, and each clinician was assessed by 2 raters. In practice, the ACP communication skills of each clinician will be assessed by a single rater.

Each rater in the validation study of Yuen et al. (2020) assessed multiple subjects. Therefore, the observational design was *crossed* (Step 1). Clinicians' *absolute* scores on the instrument are used to assess whether they have sufficient ACP communication skills (Step 2), and each clinician's ACP communication skills are determined using a *single* rater's observation (Step 3). Step 4 is redundant, because single ratings are used and clinician's absolute scores are of interest. Following the flowchart in Figure 2, the error term in the ICC should be defined as $\sigma_r^2 + \sigma_{sr}^2$. The ICC that should be selected is therefore an ICC for interrater agreement for single ratings: ICC$(A, 1)$.

Example 2: Students' Emotional Distance from their Teacher

Zee et al. (2020) studied student–teacher relationships through students' drawings. One of the attributes they measured was students' emotional distance towards their teacher, which was used to predict students' externalizing, internalizing, and pro-social behavior. To measure students' (subjects) emotional distance (attribute), a total of eight researchers (raters) coded students' drawings in which the students depicted themselves with their teacher. Each students' drawing was assessed by three of these eight researchers. The average of the three ratings per subject was used in the subsequent statistical analyses.

Each rater in the study of Zee et al. (2020) assessed multiple subjects. Therefore, the observational design was *crossed* (Step 1). The aim was to use the emotional distance measure to predict other student characteristics. The ratings were thus used in a regression model, so the researchers were interested in *relative* scores (Step 2). For all subjects, the average of three ratings was used in the subsequent statistical analyses. So, the researchers were interested in the IRR of *average ratings* (Step 3), and because the set of raters differed across subjects, the design was *incomplete* (Step 4). Following the flowchart in Figure 2, the error term in the ICC should be defined as $q\sigma^2_r + \sigma^2_{sr}/\hat{k}$. The ICC that should be selected is therefore an ICC for interrater consistency for an incomplete two-way design: ICC(Q, k). Note that although the design is incomplete, $\hat{k} = k$ because the number of raters is the same for each subject.

Example 3: Challenging Parenting Behavior

Majdandžić et al. (2021) studied the effect of parents' (subjects) severity of anxiety disorders on challenging parenting behavior (CPB; attribute) at five ages from infancy through middle childhood. CPB was operationalized as active physical and verbal behaviors to encourage the child to push their limits; for example, rough-and-tumble play, teasing, competitive games, and verbal tension-inducing sounds or verbal encouragement to do something difficult. The rating procedure was the same at each of the five ages, although different (numbers of) raters were used for each age. We therefore focus on the measurement at the age of 1 year for the description on how to select the appropriate ICC. To measure parents' CPB, a group of four raters coded parents' CPB during ten different tasks in which parents played with their child. For approximately 20% of the parents, CPB was rated by all four raters. For these parents, the researchers used the average rating in the subsequent statistical analyses. For the remaining 80% of the parents, CPB was rated by one of the four raters. For these parents, a single rating was therefore used in the subsequent statistical analyses.

The observational design of Majdandžić et al. (2021) was *crossed*: Each rater assessed multiple subjects and 20% of the subjects was assessed by multiple raters (Step 1). Note that only the CPB that was rated by four raters can be used to estimate the variance components for the ICCs. The aim of the study was to predict parents' CPB, and the ratings were used in a regression model. Therefore, the researchers were interested in the *relative* scores (Step 2). For 20% of the subjects, multiple ratings were available and each subject's average rating (across the four raters) was used in the subsequent statistical analyses. Hence, *average* ratings should be selected (Step 3). However, for the remaining 80% of the subjects only a single rating was available. The design was thus *incomplete* (Step 4). Following the flowchart in Figure 2, the error term in the ICC should be defined as $q\sigma_r^2 + \sigma_{sr}^2/\hat{k}$. The ICC that should be selected is therefore an ICC for interrater consistency for an incomplete two-way design: ICC$(Q, \hat{k})$. Note that unlike Example 2, $\hat{k} \neq k$ because the number of raters varies across subjects: The design is both incomplete and unbalanced.

Discussion

We used GT to explain the choices that need to be made when selecting an ICC for IRR, and we guided researchers through these choices by means of a flowchart. By extending current guidelines to incomplete-two way designs, we provided guidelines for the most common observational designs in psychological research. We challenged conventional wisdom about these ICCs by claiming that raters should typically not be considered fixed and that when the overlap of raters across subjects decreases, an ICC of interrater consistency gradually changes into an ICC of interrater agreement. From this perspective, the unique situation of a one-way observational design is simply a specific case of an incomplete two-way observational design, for which the ICC of interrater agreement and interrater consistency are identical.

When designing an observational study, a researcher will have to make decisions that may affect the IRR of a study. As we showed in this paper, an ICC for single ratings or incomplete designs is typically lower than an ICC for average ratings and complete designs. Because single raters or incomplete designs may be more pragmatic, researchers should balance between costs and reliability. If a researcher is interested in interrater consistency, which is typically higher than agreement, it is wise to aim for a (complete) two-way observational design so that average differences between raters do not decrease the reliability. Sample size planning for such studies is a topic that requires further research. If a researcher is interested in interrater agreement, it does not matter to which degree raters overlap across the subjects: Average differences across raters fully contribute to the error term in an ICC for both complete and incomplete designs. The ICC is therefore not affected by non-overlap across raters.

Researchers have to consider the applicability of ICCs—in terms of measurement level and independence of raters—for the purpose of their study. Although ICCs were originally proposed for continuous data, ICCs can also be extended to binary or ordinal ratings, for example assuming a latent continuous distribution underlying the ordinal responses (Ark, 2015; Vispoel et al., 2019)(cf. Cho et al., 2019), or by treating the data as continuous (Robitzsch, 2020). We want to emphasize that ICCs should be estimated based on independent ratings. In practice, researcher often organize calibration sessions, in which raters discuss their ratings with each other and decide how to handle (severe) discrepancies. When calibrated, hence non-independent, ratings are used to estimate the IRR, the IRR does not reflect the correlation between the ratings of two independent raters anymore and cannot be perceived as a measure of the reliability. In situations with insufficient reliability, calibration sessions can however be informative to investigate in what situations differences across raters occur and how higher reliability can be obtained.

After an appropriate ICC has been selected, it must be estimated from the data. We provide some guidance in estimating ICCs. User-friendly software is available for estimating IRR for complete one-way and two-way designs (e.g., the R package `irr`, Gamer et al., 2012; and the RELIABILITY command in SPSS, IBM Corp., 2020). In addition, software is available for estimating variance components and generalizability coefficients (e.g., the R package gtheory; Bloch and Norman, 2012, also see Huebner and Lucht, 2019; and the R package GENOVA, Brennan, 2001b). Software for estimating IRR in unbalanced or incomplete designs is not yet available. We therefore provide software at the Open Science Framework (Ten Hove et al., 2022) that can compute all ICCs we discussed, including those for incomplete and unbalanced two-way designs. The code includes functions to estimate ICCs by means of Maximum Likelihood Estimation (MLE) using the R package lme4 (Bates et al., 2015) or Markov chain Monte Carlo (MCMC) estimation using the R package brms (Bürkner, 2017), and is accompanied by example data—mimicking the design of Example 3—and example analyses of these data using both estimation methods. For more information about the MCMC method, we refer to Ten Hove et al. (2020, 2021), and for more information about the MLE estimation method we refer to work by Jiang (2018).

Researchers investigating the IRR of an observational instrument should report all variance components, and not just the ICC alone. This is necessary because it allows future researchers using the same instrument or protocol to derive the ICCs that suit their own measurement purpose (i.e., different types of research questions or practical applications, and using different observational designs). Also, when ICCs are reported in substantive studies, it is important to describe which ICC was used, so that reviewers can verify the appropriateness of the selected ICC. Moreover, all reliability estimates should be accompanied by measures of uncertainty (i.e., confidence intervals or standard errors; AERA et al., 2018).

According to APA standards, any paper reporting about a measurement instrument should report information about the reliability of the instrument, which is therefore mostly put to practice. Besides routinely reporting ICCs and checking whether their value meets some pre-determined threshold, researchers could interpret the ICCs to gain insight about results of subsequent statistical analyses using attenuation formula's (e.g., Lord & Novick, 1968, p. 69). These formulae provide valuable information about the attenuation of correlation (and thus limited power) in the subsequent statistical analyses due to the unreliability of the measured attributes.

We end with some suggested directions for future research into ICCs for IRR. First, we believe that more attention is required for multifaceted designs, in which there are more sources of variation present than merely subjects and raters. The ICCs for one-way and two-way designs are well understood but more complex designs occur in practice (e.g., Majdandžić et al., 2021; Yuen et al., 2020, which included multiple items and scenarios or timepoints, also facets of generalization), for which there is little guidance beyond the GT literature. Previous work already touched the topic of multilevel designs—in which clusters are the additional facet—but this considers only the situations in which the additional facet can be treated as a random effect (Ten Hove et al., 2021). IRR for situations in which the additional facet is fixed (e.g., items in a predetermined questionnaire) still need to be developed and tested. Second, we believe that the usefulness of ICCs for interrater agreement requires further investigation. If scores are given absolute interpretation, the standard error of measurement of individual subjects' scores may be more informative than the reliability of the ordering of all subjects (cf. Brennan, 2001a; Vispoel et al., 2018, 2019). Similarly, when a specific cut-score is used to determine whether subjects, for example, pass a test or qualify for treatment, cut-score specific ICCs may be more interesting than general ICCs of agreement because they are more reliable for extreme scores than for average scores (Vispoel et al., 2018). The ICC of interrater agreement is comparable to the index of dependability in GT, which always indicates the lowest dependability, achieved using the scale mean as the threshold for classification. Thresholds or cut-scores further from the mean yield greater dependability, which can be captured by threshold-specific ICCs (c.f., Vispoel et al., 2018). Third, further research is required to find the most appropriate estimation technique for estimating ICCs for IRR from incomplete and unbalanced observational designs. The MCMC and MLE estimation methods we provide on the OSF were shown to yield similar point estimates (Ten Hove et al., 2021). The MCMC estimation readily provides credible intervals for the ICCs (see, e.g., Ten Hove et al., 2020, 2021). The MLE-estimation method is supplemented with Monte-Carlo based confidence intervals that are specifically useful for coefficients such as ICCs (i.e., functions of parameters whose sampling distributions cannot be expected to be normal; c.f. MacKinnon et al., 2004). Future research should investigate which estimation technique provides the most precise ICC estimates with the best coverage rates of credible or confidence intervals in IRR-specific situations, such as incomplete or unbalanced observational designs with small numbers of raters.

References

AERA, APA, & NCME. (2018). *Standards for educational and psychological testing*. American Educational Research Association.

Aitchison, J. A. (1986/2003). *The statistical analysis of compositional data*. Chapman & Hall/Blackburn Press.

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework* (Doctoral dissertation). https://open.library.ubc.ca

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

Bartko, J. J. (1974). Corrective note to: "The intraclass correlation coefficient as a measure of reliability." *Psychological Reports*, *34*(2), 418. https://doi.org/10.2466/pr0.1974.34.2.418

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using `lme4`. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: Amee guide no. 68. *Medical Teacher*, *34*(11), 960–992. https://doi.org/10.3109/0142159X.2012.703791

Brennan, R. L. (2001a). *Generalizability theory*. Springer.

Brennan, R. L. (2001b). *Manual for* `urGENOVA` *version 2.1*. The University of Iowa. https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Bürkner, P.-C. (2017). `brms`: An r package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Cho, S.-J., Shen, J., & Naveiras, M. (2019). Multilevel reliability measures of latent scores within an item response theory framework. *Multivariate Behavioral Research*, *54*(6), 856–881. https://doi.org/10.1080/00273171.2019.1596780

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

De Vet, H. C., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*(1), 45–49. https://doi.org/s10.1016/j.jclinepi.2017.01.013

Fürst, G. (2020). Measuring creativity with planned missing data. *The Journal of Creative Behavior*, *54*(1), 150–164. https://doi.org/10.1002/jocb.352

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). `irr`: Various coefficients of interrater reliability and agreement [Computer software]. https://CRAN.R-project.org/package=irr

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, *24*(1), Article 5. https://doi.org/10.7275/5065-gc10

IBM Corp. (2020). *IBM SPSS Statistics for Windows* (Version 27.0) [Computer software]. IBM Corp.

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R. *Methodology*, *14*, 133–142. https://doi.org/10.1027/1614-2241/A000149

Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych*, *3*(2), 113–133. https://doi.org/10.3390/psych3020011

Kivisalu, T. M., Lewey, J. H., Shaffer, T. W., & Canfield, M. L. (2016). An investigation of interrater reliability for the rorschach performance assessment system (R–PAS) in a nonpatient U.S. sample. *Journal of Personality Assessment*, *98*(4), 382–390. https://doi.org/10.1080/00223891.2015.1118380

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

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

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate behavioral research*, *39*(1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4

Majdandžić, M., de Vente, W., Möller, E. L., & Bögels, S. M. (2021). Severity of fathers' and mothers' anxiety disorders predicts their observed and self-rated parenting behavior [in preparation].

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

Molenaar, D., Uluman, M., Tavşancıl, E., & De Boeck, P. (2021). The hierarchical rater thresholds model for multiple raters and multiple items. *Open Education Studies*, *3*(1), 33–48. https://doi.org/10.1515/edu-2020-0105

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5:589965*. https://doi.org/10.3389/feduc.2020.589965

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932. 10.1037/0003-066X.44.6.922

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). On the usefulness of interrater reliability coefficients. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 82th annual meeting of the Psychometric Society, Zurich, Switzerland, 2019.* (pp. 67–75). Springer. https://doi.org/10.1007/978-3-319-77249-3_6

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2020). Comparing hyperprior distributions to estimate variance components for interrater reliability coefficients. In M. Wiberg, J. González, D. Molenaar, H. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019.* (pp. 79–93). Springer. https://doi.org/10.1007/978-3-030-43469-4_7

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2021). Interrater reliability for multilevel data: A generalizability theory approach [Advanced online publication]. *Psychological Methods*. https://doi.org/10.1037/met0000391

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2022). Supplementary materials to 'updated guidelines on selecting an ICC for interrater reliability'. https://doi.org/10.17605/OSF.IO/8J26U

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

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. https://doi.org/10.1037/met0000107

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods*, *24*(2), 153–178. https://doi.org/10.1037/met0000177

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*(1), 108–131. https://doi.org/10.1037/0021-9010.90.1.108

Yuen, J. K., Kelley, A. S., Gelfman, L. P., Lindenberger, E. E., Smith, C. B., Arnold, R. M., Calton, B., Schell, J., & Berns, S. H. (2020). Development and validation of the ACP-CAT for assessing the quality of advance care planning communication. *Journal of Pain and Symptom Management*, *59*(1), 1–8. https://doi.org/10.1016/j.jpainsymman.2019.09.001

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

Table 1

Intraclass Correlation Coefficients (ICC) for One- and Two-Way Designs

Design	Agreement or	Random or	ICC	
	consistency	fixed	Single ratings	Average ratings
Two-way	Agreement	Random	$\text{ICC}(A,1) = \frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2+\sigma_{sr}^2}$	$\text{ICC}(A,k) = \frac{\sigma_s^2}{\sigma_s^2+(\sigma_r^2+\sigma_{sr}^2)/k}$
		Fixed	$\text{ICC}(A,1)^* = \frac{\sigma_s^2-\sigma_{sr-e}^2/(k-1)}{\sigma_s^2+\theta_r^2+(\sigma_{sr-e}^2+\sigma_e^2)}$	$\text{ICC}(A,k)^* = \frac{\sigma_s^2-\sigma_{sr-e}^2/(k-1)}{\sigma_s^2+(\theta_r^2+(\sigma_{sr-e}^2+\sigma_e^2)/k}$
	Consistency	Random	$\text{ICC}(C,1) = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{sr}^2}$	$\text{ICC}(C,k) = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{sr}^2/k}$
		Fixed	$\text{ICC}(C,1)^* = \frac{\sigma_s^2-\sigma_{sr-e}^2/(k-1)}{\sigma_s^2+(\sigma_{sr-e}^2+\sigma_e^2)}$	$\text{ICC}(C,k)^* = \frac{\sigma_s^2-\sigma_{sr-e}^2/(k-1)}{\sigma_s^2+\sigma_{sr}^2/k}$
One-way	–	Random	$\text{ICC}(1) = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{r:s}^2}$	$\text{ICC}(k) = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{r:s}^2/k}$

Note. * = ICC for fixed raters, for details see McGraw and Wong (1996) and Shrout and Fleiss (1979).

Table 2

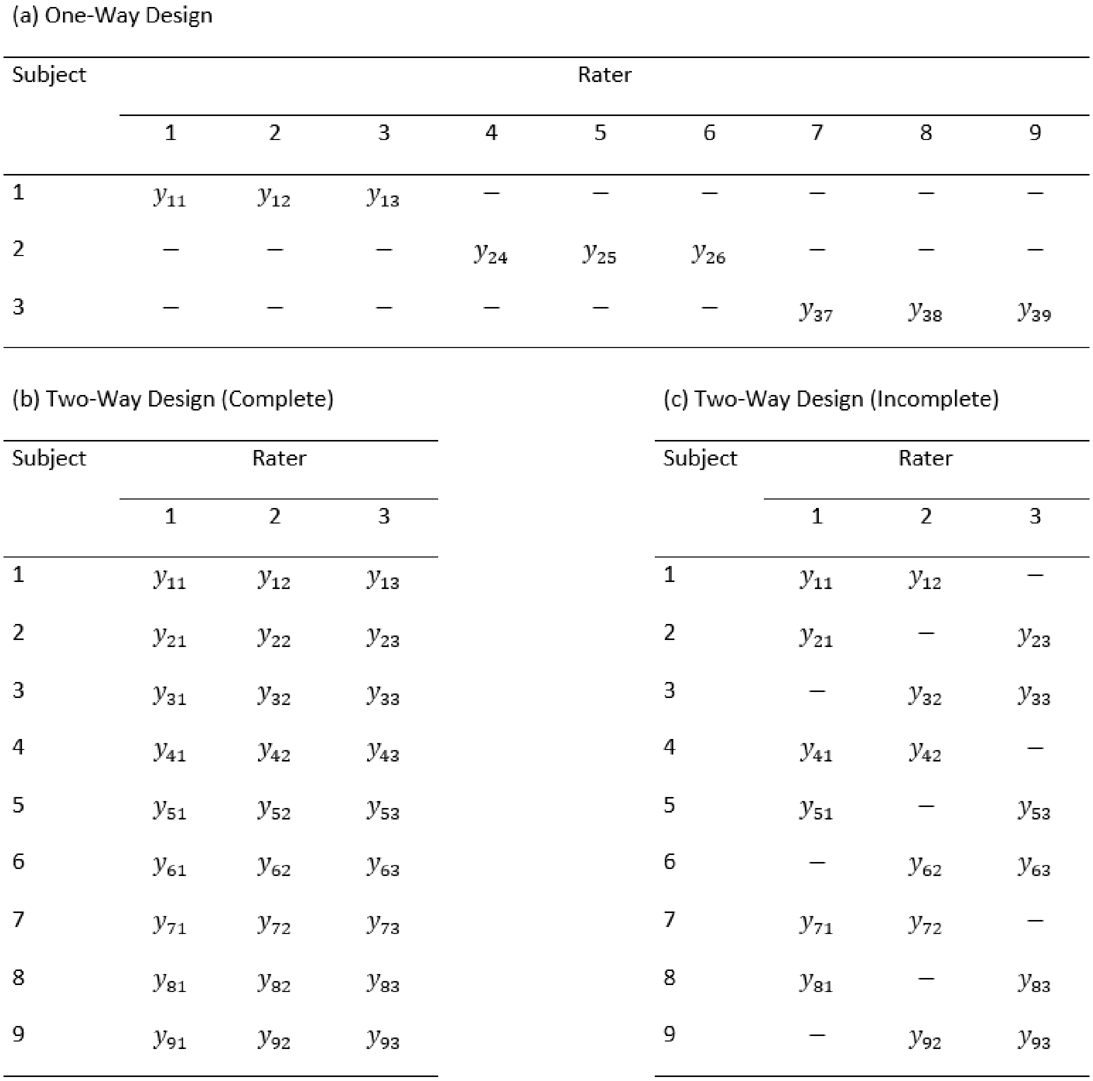
Definition of the Error Variance for Different Observational Designs

Absolute or	Single or		Two-way Design		One-way Design	
Relative	Average	Error variance	Complete	Incomplete	Balanced	Unbalanced
Absolute	Single	$\sigma^2_{\epsilon.abs}$	$\sigma^2_r + \sigma^2_{sr}$	$\sigma^2_r + \sigma^2_{sr}$	$\sigma^2_{r:s}$	$\sigma^2_{r:s}$
	Average	$\sigma^2_{\varepsilon.abs}$	$\frac{\sigma^2_r+\sigma^2_{sr}}{k}$	$\frac{\sigma^2_r+\sigma^2_{sr}}{\hat{k}}$	$\frac{\sigma^2_{r:s}}{k}$	$\frac{\sigma^2_{r:s}}{\hat{k}}$
Relative	Single	$\sigma^2_{\epsilon.rel}$	σ^2_{sr}	$q\sigma^2_r + \sigma^2_{sr}$	$\sigma^2_{r:s}$	$\sigma^2_{r:s}$
	Average	$\sigma^2_{\varepsilon.rel}$	σ^2_{sr}/k	$q\sigma^2_r + \sigma^2_{sr}/\hat{k}$	$\frac{\sigma^2_{r:s}}{k}$	$\frac{\sigma^2_{r:s}}{\hat{k}}$

Note. $\hat{k}$ is the harmonic-mean number of raters; $q =$ the proportion of non-overlap across raters. An incomplete two-way design is potentially unbalanced, which is accommodated by $\hat{k}$.

Figure 1

Example of Three Observational Designs.



(a) One-Way Design

Subject	Rater								
	1	2	3	4	5	6	7	8	9
1	y_{11}	y_{12}	y_{13}	–	–	–	–	–	–
2	–	–	–	y_{24}	y_{25}	y_{26}	–	–	–
3	–	–	–	–	–	–	y_{37}	y_{38}	y_{39}

(b) Two-Way Design (Complete)

Subject	Rater		
	1	2	3
1	y_{11}	y_{12}	y_{13}
2	y_{21}	y_{22}	y_{23}
3	y_{31}	y_{32}	y_{33}
4	y_{41}	y_{42}	y_{43}
5	y_{51}	y_{52}	y_{53}
6	y_{61}	y_{62}	y_{63}
7	y_{71}	y_{72}	y_{73}
8	y_{81}	y_{82}	y_{83}
9	y_{91}	y_{92}	y_{93}

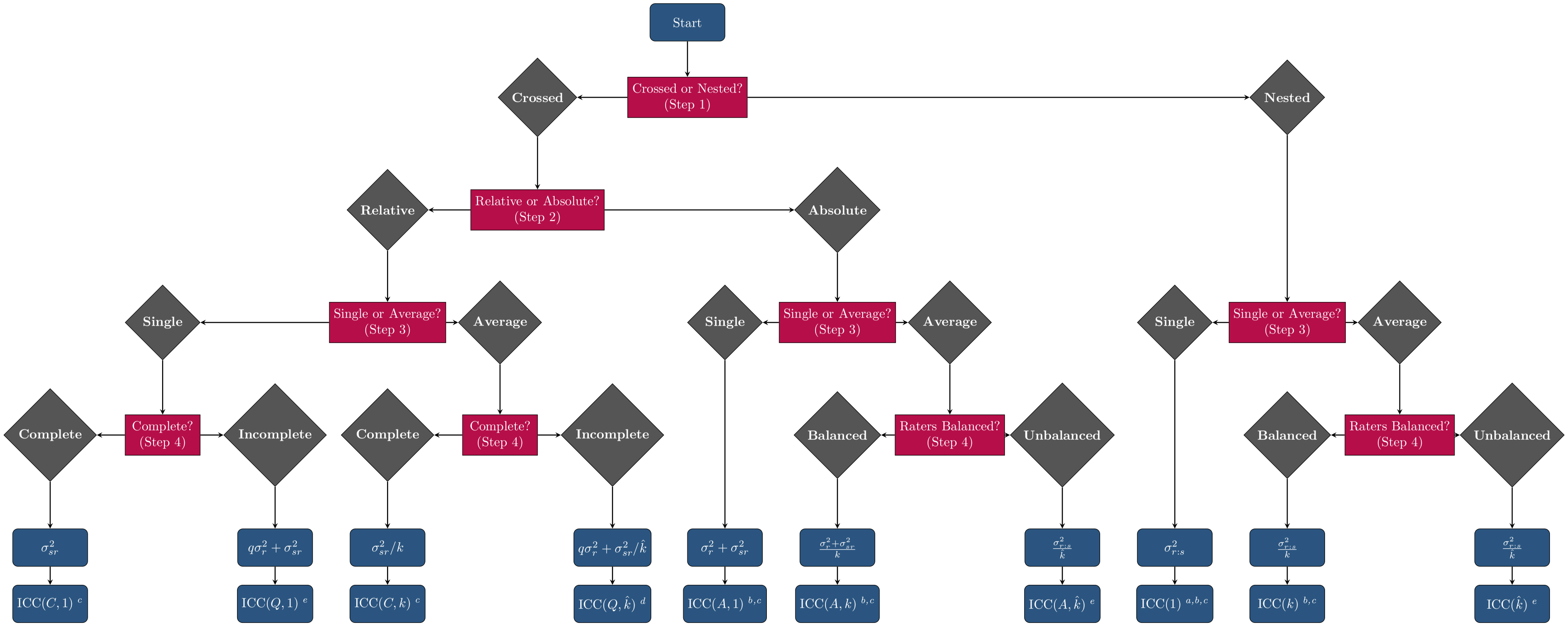
(c) Two-Way Design (Incomplete)

Subject	Rater		
	1	2	3
1	y_{11}	y_{12}	–
2	y_{21}	–	y_{23}
3	–	y_{32}	y_{33}
4	y_{41}	y_{42}	–
5	y_{51}	–	y_{53}
6	–	y_{62}	y_{63}
7	y_{71}	y_{72}	–
8	y_{81}	–	y_{83}
9	–	y_{92}	y_{93}

Note. *Note.* (a) A one-way design, where 9 raters are nested within 3 subjects; (b) A complete two-way design, where all 9 subjects were rated by all 3 raters, and (c) An incomplete two-way design, where the 3 raters are crossed with, but vary across the 9 subjects.

Figure 2

Flowchart to Select an Error Term and ICC for Interrater Reliability



Note. *Note.* ICCs as defined by [a] Bartko (1966); [b] Shrout and Fleiss (1979); [c] McGraw and Wong (1996); [d] Putka et al. (2008); [e] ICC has not been defined in the literature.