Interrater reliability for incomplete and dependent data

ten Hove, D.

Publication date
2023

Chapter 8

General Discussion

8.1 Introduction

The central question of this dissertation was how to estimate the interrater reliability (IRR) from incomplete and dependent observational data. IRR expresses the degree to which subjects' attribute scores are independent of raters. In a standard observational design (i.e., a complete two-way design), multiple independent subjects are each rated by multiple independent raters, and these raters are the same for each subject. Most traditional IRR coefficients are defined for such a complete two-way design. However, in social and behavioural observational research, the raters are often not the same for each subject, resulting in incomplete data. Also, the rated subjects are often nested within clusters or relationships, resulting in dependent data. Dependent data have various facets of theoretical interest about which researchers may formulate research questions. For example, multilevel data contain subject and cluster-level components, and interdependent social network data contain actor, partner, and relationship components. Prior to the developments in this dissertation, none of the conceptualizations of IRR could be readily used for both incomplete and dependent data, resulting in less or non-informative IRR coefficients. Also, this dissertation provides the badly needed IRR estimation methods that can accommodate missing observations and dependence structures in dependent data.

In this dissertation, I used the framework of Generalizability theory (GT; Cronbach et al., 1963) to provide definitions and estimation methods for IRR of incomplete and dependent data. **Chapter 1** introduced the topics IRR, incomplete data, and multilevel data, and discussed the need for new methods. **Chapter 2** illustrated and discussed the issue of an abundance of IRR coefficients that follow from different conceptualizations of IRR. **Chapter 3** explained why the intraclass correlation coefficients (ICCs) are probably the best candidate for defining IRR, and used GT to extend the different definitions of ICCs to incomplete observational designs and to provide updated guidelines on when to use which ICC definition. ICCs are traditionally estimated with ANOVA-based methods, which are not straightforward for incomplete or dependent data. Markov chain Monte Carlo (MCMC) estimation of hierarchical linear models can handle such data, but requires the definition of hyperprior distributions. **Chapter 4** investigated the effect of different hyperprior distributions on ICC estimates under different conditions that are common to observational studies. **Chapter 5** described a simulation study that compared the MCMC approach to estimate ICCs from incomplete data with two maximum likelihood based approaches. **Chapter 6** generalized the definitions and MCMC-based estimation method of ICCs for IRR to multilevel data, and evaluated the IRR estimates in a simulation study. **Chapter 7** generalized the definitions and MCMC-based estimation methods of ICCs for IRR to the more complex case of interdependent social network data, and evaluated the IRR estimates in a simulation study.

8.2 Main Findings

The main findings per chapter were as follows. **Chapter 2** showed that different IRR coefficients yield a variety of IRR estimates when applied to the same data, which is due both to differences in their specific formulas and to differences in their underlying conceptualizations of IRR. I argued that researchers should select IRR coefficients based on the underlying conceptualization of IRR and its usefulness in a research setting. For example, it is known that reliability provides information regarding measurement precision, attenuation of correlations, and inflated power of statistical tests. However, the implications of different conceptualizations of IRR for subsequent analyses are unclear.

Chapter 3 used GT to explain the choices that need to be made when selecting an ICC for IRR. By extending current guidelines to incomplete two-way designs, I provided guidelines for selecting an ICC for all possible two-way observational designs. I explained that ICCs of interrater agreement are useful when ratings are used for absolute decision making, and that ICCs for interrater consistency are useful when ratings are used for relative comparisons, such as correlational studies, regressions models, and comparing groups. I challenged conventional wisdom about the 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, a one-way observational design—in which each subject is rated by a unique set of raters—is a specific case of an incomplete two-way observational design, for which the ICC of interrater agreement and interrater consistency are identical. I discussed a four-step approach to guide researchers in the process of selecting an ICC, which I visualized in a flowchart, and I used three empirical examples from clinical and developmental domains to guide researchers through this flowchart.

Chapter 4 showed that under various conditions, half-t hyperprior distributions for the random-effect *SD*s (from which ICCs are derived) have a slight advantage over uniform hyperprior distributions when estimating ICCs for IRR with MCMC of hierarchical models (MCMC-HL). However, the number of raters used to estimate IRR had a larger effect on the performance of the MCMC estimates than the choice of hyperprior distributions. I therefore recommended researchers to use half-t hyperprior distributions for the random-effect *SD*s (combined with MAP point estimates and percentile-based BCIs), and most importantly, at least three raters to estimate the IRR using an MCMC approach.

Chapter 5 showed that, for continuous data, ICCs can best be estimated with maximum likelihood estimation of random-effects models (MLE-RE). This method yielded more accurate ICC estimates and was practically more feasible compared to maximum likelihood estimation of common-factor models (MLE-CF) and MCMC-HL, two methods that were also recently proposed to estimate generalizability coefficients or ICCs. The MLE-RE point and variability estimates (i.e., *SE*s) had the least bias, and the method converged in a few seconds for almost all data sets. I recommended to complement this method with Monte-Carlo CIs, which had almost nominal coverage rates. For binary data,

MLE-RE was not available because *SE*s for the ICCs cannot be obtained. MCMC-HL yielded more accurate ICC point and uncertainty estimates than MLE-CF, and was most practically feasible. The method converged for almost all data sets, but the method costs much estimation time when the proportion of missing observations is substantial.

Chapter 6 used GT to define separate ICCs for subject- and cluster-level components in multilevel data. I proposed MCMC estimation of hierarchical linear models to estimate these subject- and cluster-level ICCs, and evaluated its estimates in a simulation study. The simulation study indicated that the proposed method is useful for typical social and behavioral research because it performs adequately even with modest sample sizes and missing data. Although selecting two raters is common practice, the simulation revealed that this is not the best design choice when estimating ICCs, because it resulted in overestimated *SE*s of the ICCs. Assigning more than two raters to each subjects may be costly for researchers, but a follow-up simulation study showed that a planned missing data design—in which as few as two raters per subject are randomly sampled from a larger rater pool—could yield accurate variability estimates. I illustrated the proposed estimation method using data on student–teacher relationships, and discussed the potential differences between conflated ICCs—that ignore cluster-related effects—and subject-level or cluster-level ICCs.

Chapter 7 used GT to extend the social relations model for interdependent social network data with rater effects, resulting in the rater-extended social relations model. This model decomposes the variance in observational interdependent social network data into the variance components stemming from differences across actors, partners, raters, and their interactions. The variances in each of these effects were used to define IRR coefficients for actor, partner, and relationship effects separately. I proposed a Bayesian hierarchical linear model to estimate these IRR coefficients, and illustrated the method using data on social mimicry. I tested the bias and coverage of the ICCs under favorable and less favorable conditions in a simulation study. Unfortunately, the method did not yield unbiased ICCs with nominal coverage rates in conditions with both small subgroups of interacting subjects and a small number of raters. However, on average, and especially in conditions with sufficient raters and subjects (i.e., 10 raters and 10 subjects), ICCs were mostly unbiased and had good coverage rates.

8.3 Implications for Observational Research

8.3.1 The IRR *of What*

The results of this dissertation have several implications for observational research. One of the take home messages is that it is of vital importance to consider *what* one wants to know the IRR of. The IRR provides researchers with information about the degree to which rater effects attenuate regression coefficients or affect measurement precision. The IRR should thus be investigated of the variables that are used in statistical analyses or

used for decision making. Said in GT terminology: When defining the IRR, researchers should first determine their facet of theoretical interest (i.e., the facet of differentiation).

In chapters 6 and 7, this *what* concerned the separate components in dependent data, instead of the observed attribute scores. For multilevel data (Chapter 6), these components were the subject- or cluster effects in a multilevel analysis. IRR was therefore defined for subject- and cluster effects separately, instead of for the observed scores. For interdependent social network data (Chapter 7), this *what* concerned the separate actor-, partner-, and relationship effects in a social relations model, and the IRR was therefore defined for each of these effect separately, instead of for the observed dyadic scores. However, if it is undesirable to decompose dependent data into several components of interest, the IRR of an integrated score should be investigated. For example when children are selected for additional tutoring or an advanced program based on an observed score, instead of based on their deviation from a cluster mean (e.g., their relative standing in the school).

The take-home message that researchers should consider of *what* they should estimate the IRR also holds for non-nested data and for data with other sources of dependency than the clusters and relationships that I considered in my dissertation. For example, IRR could be defined for items that measure an attribute, but also for a composite score of several items. If such a composite score is used to make decisions about subjects or to answer research questions about an attribute, the IRR of the composite scores is of interest. However, if the IRR of the composite scores is low, it may be interesting to inspect the item-specific IRR, to investigate for which items the rating procedure should be improved to yield higher IRR. Also, a type of dependent data I did not consider in my dissertation is longitudinal data. The variation in longitudinal observations can be decomposed into variation in subjects' average scores over time, the time-specific deviations from these averages, and rater-related effects, similar to the multilevel data decomposition in Chapter 6. For longitudinal data, researchers should thus also consider the level of theoretical interest when defining and estimating IRR coefficients.

8.3.2 Designing Observation Studies

When designing an observational study, researchers have to make decisions that may affect the IRR. From my dissertation, several guidelines for efficiently designing IRR studies can be derived. For example, researchers who are interested in interrater consistency (because they use ratings for relative purposes; Chapter 3), should aim for a complete two-way observational design so that average differences between raters do not reduce the reliability. An ICC for interrater consistency is lower for incomplete designs than for complete designs because the q factor, which scales the contribution of the main-rater variance to the error term in the ICCs for interrater consistency, increases as overlap of raters across subjects decreases (Chapter 3). Hence, the less overlap of raters across subjects, the lower the ICCs for interrater consistency. Because incomplete designs may

be more cost-efficient and pragmatic (e.g., in terms of workload per rater) in a research setting, researchers should balance between pragmatics, costs and reliability when they use ratings for relative purposes.

Researchers who are interested in the interrater agreement (because they use ratings for absolute purposes; Chapter 3), should gather information from more than two or three raters to yield accurate IRR estimates. ICCs for interrater agreement were biased and had poor coverage rates in conditions with few raters (chapters 4, 5, and 6). These poor ICC estimates occur because data from few raters provide little information about the population variance of the main-rater effects. This variance is included in the denominator of the ICC for interrater agreement, making this ICC harder to estimate than an ICC for interrater consistency. Especially for decision making and thus when interrater agreement is of interest, the stakes may be high, and unreliable measurement is problematic. I therefore argue that for high stakes assessments, researchers should recruit a substantial number of raters, so that the interrater agreement can be estimated with sufficient certainty. Because ICCs for interrater agreement are not affected by non-overlapping raters across subjects (Chapter 3), a planned missingness design can be used if recruiting many raters per subject is costly and the budget is limited (cf. Jorgensen et al., in press).

8.3.3 Handling Missing Observations

When handling missing observations (i.e., nonresponse), all IRR estimation methods that I discussed in my disseration assume that the nonresponse is ignorable (e.g., R. J. A. Little & Rubin, 2002). In the simulation studies in chapter 5 and 6, I considered only planned missing-data designs with missing at random ratings, for which the assumption of ignorable nonresponse is plausible. If the nonresponse is due to other factors than planned missingness, the the assumption of ignorable nonresponse may not be plausible, and the nonresponse may cause biased estimates. Hence, researchers should carefully consider their missing data mechanisms before estimating the ICCs from incomplete data.

8.3.4 Interpreting and Reporting the IRR

Most researchers who use observational measurements report the IRR as is prescribed by publishing standards (e.g., AERA et al., 2018). This is a good practice because IRR is a prerequisite for qualitative and valid measurements that provide information about the targeted attributes. However, investigating the IRR should not be a goal in itself. Instead, the IRR should inform researchers about whether their rating procedures require improvements. Several researchers have proposed benchmarks to interpret the magnitude of the IRR. For example, the often-cited paper by Cicchetti (1994) suggests to interpret ICC values below .40 as a poor, values between .40 and .60 as moderate, values between .60 and .75 as good, and values between .75 and 1.00 as excellent clinical significance (cf. Landis & Koch, 1977). However, it remains unclear what this 'clinical significance'

indicates. I argue that, instead of using prespecified benchmarks, the IRR should best be interpreted in terms of its effect on the conclusions that are drawn about the attributes in correlational studies or decision making in practice.

In designs with only one facet of generalization (here, raters), generalizability coefficients—specifically, ICCs of interrater consistency—are a generalization of CTT's definition of reliability (Chapter 3), and can be interpreted as the correlation between two parallel observations of the same attribute (i.e., $\rho_{XX'}$; e.g., Vispoel et al., 2018a). ICC$(Q, \hat{k})$, the most general IRR coefficient, expresses the squared correlation between two independent averaged ratings, that are averaged over raters that partially do not overlap across subjects. This ICC generalizes to the ICC(C, k) if the raters are the same for each subject, and to the ICC$(C, 1)$ when each subject is observed by only a single rater, which is the same rater for each subject (Chapter 3). This IRR definition is very useful, because its implications for statistical issues such as attenuation of correlation and measurement precision have been thoroughly investigated (e.g., Lord & Novick, 1968, p. 69).

In correlational studies, the significance of IRR is in its effect on the attenuation of regression coefficients. Following Spearman's (1904) attenuation formula, the correlation between the true scores of X and Y (ρ_{XY}) is the observed correlation between these variables (r_{XY}) divided by the product of the square roots of the reliability of X and Y ($r_{XX'}$ and $r_{YY'}$, respectively):

$$\rho_{XY} = \frac{r_{XY}}{\sqrt{r_{XX'} r_{YY'}}}. \tag{8.1}$$

Hence,

$$r_{XY} = \rho_{XY} \sqrt{r_{XX'} r_{YY'}}. \tag{8.2}$$

The percentage underestimation of the true correlation due to unreliability is thus a product of the squared roots of the reliability of both variables under study: $(1 - \sqrt{r_{XX'} r_{YY'}}) \times 100\%)$. Suppose the interrater consistency of X is $r_{XX'} = .60$ (and thus can be interpreted as good), the true scores of X and Y were strongly correlated (e.g., $\rho_{XY} = .80$), and Y would be perfectly reliable (i.e., $r_{YY'} = 1$), the observed correlation between X and Y would be only $r_{XY} = \rho_{XY} \sqrt{r_{XX'}} \sqrt{r_{YY'}} = .80 \times \sqrt{.60 \times 1} = .62$, which is quite an underestimate of the true correlation. Moreover, Y is often also not perfectly reliable, which attenuates the correlation between X and Y even more. This attenuation effect, and thus the IRR of variables, should be considered when conducting power analyses: The unreliability of variables lowers the power of statistical tests because the true correlation will be underestimated. Also, using Equation 8.1, researchers could use the IRR estimates to disattenuate the estimated correlations or regression coefficients between observed variables (e.g., Cho & Preacher, 2016; Vispoel et al., 2018b).

8.3.5 Generalizability *over What*

Throughout my dissertation, I used GT to define ICCs that express the degree to which (differences between) facets of interest's scores can be generalized over raters. However, observational research often also includes other nuisance facets than merely raters, and GT-based coefficients are useful to estimate more than merely *interrater* reliability. Generalizability coefficients and indices of dependability are an extension of CTT's reliability, that reframe reliability as the generalizability over any number of nuisance facets (Brennan, 2001a; Cronbach et al., 1963). The ICCs for IRR allow researchers to inspect the quality of the rating procedure in particular, for several facets of interest. However, for information about the attenuation by multiple nuisances facets, other generalizability coefficients are more useful. For example, Bijlsma et al. (2022) and Van der Scheer et al. (2019) used multiple raters to assess teachers' skills at multiple occasions using a multiple-item scale. The nuisance facets are raters, occasions, and items, and GT can be used to express the degree to which teachers' scores can be generalized over raters, items, and occasions.

8.4 Directions for Future Research

8.4.1 ICCs for Discrete Responses

Various potential avenues for future research pertain to the estimation of ICCs for IRR from discrete responses. Traditionally, GT treats discrete data as continuous (e.g., Bock et al., 2002), but in Chapter 5 I proposed and tested different estimation methods for IRR of independent (i.e., two-way) binary data, using logit and probit link functions, which have latent response variable interpretations (Agresti, 2010). The performance of the methods in Chapter 5 could be further investigated for asymmetric thresholds, and the approach could be extended to ratings that have more than two categories. Furthermore, whereas MLE-RE performed best for continuous responses, this method could not be recommended for discrete data because measures of uncertainty are currently unavailable. Such uncertainty measures would require further development of the methods of Wang and Merkle (2018), so that a robust asymptotic covariance matrix of the variance components from two-way observational designs with discrete responses can be obtained. Another avenue would be to combine the modelling approach for binary data in Chapter 5 with the multilevel and interdependent social network data approaches of chapters 6 and 7, so that the IRR can be obtained from discrete dependent (i.e., multilevel or interdependent social network) data.

8.4.2 ICCs for Incomplete Observational Designs

Incomplete—hence missing—data was an important topic in my dissertation, but it is also a topic that requires future research. Chapter 3 defined ICCs for interrater consistency for incomplete observational designs, and Chapter 5 compared methods to estimate ICCs from incomplete data. Chapter 6 suggested a planned missing data design to accurately estimate the IRR from multilevel data, but did not define ICCs of interrater consistency for incomplete dependent data. It might be useful to extend the generalize the ICC($Q, \hat{k}$) in Chapter 3 to the ICCs for multilevel and interdependent social network data in chapters 6 and 7. Both q and $\hat{k}$ should then be defined for each facet of interest separately. This definition would be less straightforward than for independent data, because the proportions of non-overlapping raters may differ both within and between clusters, as may the numbers of raters per subject and cluster.

8.4.3 Software Developments for ICC Estimation

Estimation methods are only useful if researchers can actually use the methods in their own research. Chapters 4, 5, 6, and 7, which proposed or tested estimation methods for IRR, were therefore accompanied by an Open Science Framework (OSF) page that provides researchers with R functions to apply these methods. In the various chapters, I also illustrated the use of the software on various data sets, for which the software code is provided on the OSF. To make the methods even more accessible to substantive researchers—who may not have any experiences with command-line programming in R—it would be useful to have dedicated software packages that are accompanied by a graphical user interface, like a Shiny app (Chang et al., 2022), or to implement the ICCs for incomplete and dependent data in other existing software that is often used in the social- and behavioural sciences (e.g., SPSS; IBM Corp., 2021; JASP Team, 2022).

8.4.4 Interpreting the IRR

This general discussion considered the interpretation of ICCs for interrater consistency in terms of the (dis)attenuation of regression coefficients rather extensively, but the interpretation of ICCs for interrater agreement require further investigation. These ICCs expresses the reliability of subjects observed scores (Chapter 3). Although it is of importance to know whether the IRR is high or low, it is less intuitive to interpret the impact of the IRR on decision making. For individual decisions, it is not intuitive to interpret the ICC as the proportion true-score variation in the observed scores, and more meaningful measures (that can be derived from the ICC or its error component) could have more practical value. If scores are given absolute interpretation, the standard error of measurement of individual subjects' scores may be more informative than the reliability of the scores of all subjects (cf. Brennan, 2001a; Vispoel et al., 2018a, 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 global ICCs of agreement. The ICC of interrater agreement (cf. the global index of dependability in GT) express the dependability for when the scale mean is used as the cutoff for classification. Cutoffs further from the mean yield greater dependability, which can be captured by cutoff-specific ICCs (e.g., Vispoel et al., 2018a).

Another direction for future IRR research is the fully Bayesian MCMC-HL approach. Although this method was not preferred over the MLE-RE approach (Chapter 5), I still believe that further development and investigation of this method would be useful. Using the posterior distributions of the IRR coefficients, researchers could answer questions about the probability that the IRR is lower (or higher) than a specific cut-off value (cf. Pfadt et al., 2022). Such a question cannot be answered using the frequentistic MLE-based approaches.

8.4.5 IRR for Longitudinal data

Next to multilevel designs and social network designs, longitudinal research designs constitute another important source of dependent data. As we describe in Chapter 6 and Section 8.3.1, the ICCs for multilevel data may also be useful for longitudinal data. In longitudinal data, occasions constitute Level 1, and subjects constitute Level 2, comparable to the respective subjects and cluster levels in Chapter 6. When such data are obtained through external raters, ICCs for IRR could be defined for the variation within or between subjects, depending on the facet of interest in the substantial analyses or the decision making in practice. More intensive time series data, in which subjects are observed intensively, may require other IRR definitions. IRR coefficients that account for the auto-dependence in the data may be more suitable for such longitudinal data compared to the ICCs for multilevel data (cf. Bodner et al., 2021).

8.4.6 Estimating and Defining Generalizability Coefficients

ICCs for interrater consistency are generalizability coefficients, and ICCs for interrater agreement are indices of dependability (Chapter 3). The results from the simulation studies in chapters 4, 5, 6, and 7 are therefore informative for any study in which generalizability coefficients or indices of dependability are estimated. Traditionally, generalizability coefficients are estimated using ANOVA-based methods. My dissertation showed that MLE-RE and MCMC-HL are fruitful alternatives that are even applicable to incomplete and dependent continuous data. Also, MCMC-HL allows the estimation of generalizability coefficients and indices of dependability from discrete data. Recently, Vispoel et al. (2018a, 2019) and Jorgensen (2021) proposed to estimate generalizability coefficients and indices of dependability with MLE-CF, an approach I also investigated for ICC estimation in Chapter 5. The simulation results showed that MLE-CF estimates were biased, had low coverage rates, and had convergence issues for incomplete designs.

In MLE-CF estimation, raters are modeled as fixed rather than random indicators of a latent variable, which may explain the poor results. I expect that these poor results will also be found if MLE-CF is used to estimate generalizability coefficients and indices of dependability.

8.5 Concluding Remarks

Based on GT, and in line with Bartko's (1966, p. 3) advice that the "use of the ICC [...] should be restricted by the underlying model which most adequately describes the experimental situation", my dissertation provides a comprehensive framework of IRR for incomplete and dependent data. In addition, it provides different insights in the definition of ICCs for IRR in general, and alternatives for the traditional ANOVA-based estimation methods for ICCs. Considering the findings of my dissertation, I argue that the ICC would be a good candidate for a standard interrater *consistency* measure.

In their call for a standard *agreement* measure for coding data, Hayes and Krippendorff (2007) argued that Krippendorff's alpha is the preferred agreement coefficient because, opposed to other agreement measures, it meets all the following criteria:

1. The coefficient measures agreement between any number of raters and subjects, without requiring fully crossed or nested designs;
2. The coefficient can be interpreted on a fixed scale bounded by perfect agreement and absence of agreement;
3. The coefficient's size does not depend on the number of categories of discrete data;
4. The coefficient's sampling distribution is estimable so that uncertainty about the population parameter can be quantified; and
5. The coefficient is applicable to (and interpretable on the same scale for) any scale of measurement: nominal, ordinal, interval, or ratio data.

Without complying with or arguing against Hayes and Krippendorff (2007), I argue that the ICC meets the listed criteria that are of importance for a *consistency* measure, because:

1. ICCs can estimate the interrater consistency from any number of raters and subjects, without requiring a fully crossed or nested design (chapters 3 and 5). Besides, ICCs can be estimated *for* ratings that are *averaged over* multiple raters in any type of observational design (Chapter 3);
2. All ICCs that I discussed in my dissertation can be interpreted on a fixed scale bounded by perfect agreement and absence of agreement. The ICCs are defined as proportions of the observed variance in ratings, and are therefore bounded between

0 and 1. If none of the variation in the ratings can be attributed to differences across subjects, the ICC = 0, indicating absence of IRR. When all variation in ratings can be attributed to differences across subjects, none can be attributed to rater-related effects, hence the ICC = 1, indicating perfect IRR;

3. The magnitude of the ICC is not dependent on the number of categories in discrete data because the logit links for MCMC-HL and MLE-RE, have latent response variable interpretations (Chapter 5);

4. All ICCs that I discussed in my dissertation are accompanied by uncertainty measures. For the maximum likelihood based methods I recommended Monte-Carlo CIs (Chapter 5), and for the Bayesian methods I recommended BCIs based on the percentiles of the posterior distributions (chapters 4, 5, 6, and 7);

5. Although ICCs were originally developed for continuous data, I showed that they can also be applied to binary or ordinal data using a logit or probit link or a latent-variable interpretation (chapters 4, 5, and 6). More research is required to determine whether ICCs can be made applicable to nominal data. However, I expect that very few observational studies use raters to gain information about purely nominal attributes of subjects for correlational studies.

Moreover, the ICCs meet three additional criteria, not listed by Hayes and Krippendorff (2007), which I consider relevant for a consistency measure as well:

6. The ICC for interrater consistency can be interpreted in terms of its effect on a regression coefficient (i.e., attenuation) because this ICC is a generalization of CTT's definition of reliability (Chapter 3). Agreement measures include effects in their denominator that do not affect the observed differences across subjects. Agreement measures would therefore overestimate the impact of rater effects on the estimated regression coefficients, correlation coefficients, or statistical power;

7. The ICCs can be estimated from dependent data, and can express the IRR for separate facets of interest in such data (chapters 6 and 7);

8. The ICCs for IRR are no stand-alone reliability measures, but are embedded in GT, and are closely linked to other types of reliability (Chapter 3). They allow comparison with other reliability measures that generalize over other facets of interest than raters, and facilitate comparison of coefficients not only across types of data (Criterion 5) but also across types of measurement error, while still being interpreted on the same scale (Criterion 2).

ICCs are thus particularly valuable, especially when ratings are used for relative purposes, such as correlational studies. I therefore highly recommend to adopt ICCs as a standard measure for IRR.