Interrater reliability for incomplete and dependent data

ten Hove, D.

Publication date
2023

Chapter 5

Estimating Interrater Reliability from Incomplete Data

Abstract

We compared different methods to estimate interrater reliability (IRR) by means of intraclass correlation coefficients (ICCs) from incomplete data. ICCs are flexible IRR estimators that are based on the variance decomposition of observations, that can estimate both agreement and consistency across both single and averaged ratings. ICCs are typically estimated using mean-squares from an ANOVA model, the computation of which is not straight-forward for incomplete designs in which raters partly vary across subjects. Therefore, we used simulated data to compare the computational accuracy (bias of point estimates, bias of variability estimates, root mean squared error, and coverage rates) and computational feasibility (convergence rates and estimation time) of three estimation methods for incomplete designs: Markov chain Monte Carlo estimation of Bayesian hierarchical linear models, maximum likelihood estimation of random-effects models, and maximum likelihood estimation of common-factor models. For continuous responses, maximum likelihood estimation of random-effects models with Monte-Carlo confidence intervals is preferred based on all criteria. For binary responses, Markov chain Monte Carlo estimation is preferred based on computational accuracy, but the estimation time is long. This paper is accompanied by software code that enables researchers to use these methods, and includes a tutorial that applies the software on empirical data.

Chapter 5 is under review as: **Ten Hove, D.**, Jorgensen, T. D., & Van der Ark, L. A. (*Under review*). How to estimate ICCs for interrater reliability from incomplete designs: A simulation study comparing MLE and MCMC estimation.

5.1 Introduction

Observational studies use raters to obtain information about an attribute of subjects. For example, Yuen et al. (2020) used physician faculty (the raters) to assess the communication skills (the attribute) of clinicians in training (the subjects). Ideally, the variation in these ratings originates from the differences among subjects' attributes, but as little as possible from the differences among raters and their perceptions. Interrater reliability (IRR) provides information about the ability to differentiate among subjects based on ratings. IRR bounds the precision and validity of these ratings (cf. Lord & Novick, 1968, p. 72). Using ratings with low IRR may result in biased estimates and loss of power in subsequent statistical analyses, and incorrect decisions in diagnostic settings or assessments.

Arguably, the intraclass correlation coefficients (ICCs) are the most flexible estimators of IRR (Ten Hove et al., 2022c). ICCs are defined based on the variance decomposition of observations, and can be used to estimate both agreement and consistency across both single and averaged ratings (Shrout & Fleiss, 1979; McGraw & Wong, 1996). ICCs can handle observations by more than two raters (Shrout & Fleiss, 1979; McGraw & Wong, 1996), and flexible estimation methods allow the estimation of ICCs from both continuous and discrete data (cf. Vispoel et al., 2019; Ark, 2015; Jorgensen, 2021). ICCs are similar to generalizability coefficients (Cronbach et al., 1963), and thus closely related to the classical definition of reliability (Ten Hove et al., 2022c). This is useful because the implications of reliability for statistical issues such as attenuation of correlation and measurement precision have been thoroughly investigated and communicated (e.g., Lord & Novick, 1968, p. 69).

Traditionally, ICCs are estimated using mean-squares estimation in a random-effects ANOVA model. This approach is implemented in most standard software packages for social and behavioral research (e.g., the R package `irr`, Gamer et al. (2012); and the RELIABILITY command in SPSS, IBM Corp. (2021)). This estimation method accommodates one-way designs in which all subjects have a unique set of raters (Figure 5.1, panel c) and complete two-way designs in which all subjects have the same set of raters (Figure 5.1, panel a, Cronbach et al., 1963; Webb et al., 2006). As in any study, observational studies may contain missing ratings, resulting in an incomplete two-way design in which there is some overlap in raters across subjects (e.g., Figure 5.1, panel b). These missing ratings may be unforeseen (nonresponse, attrition) but they often are planned to make the study feasible in terms of costs, response burden, or logistics. Examples of such incomplete two-way observational designs include Zee et al. (2020); Majdandžić et al. (2021); Yuen et al. (2020).

Over the past few years, researchers have proposed a variety of alternative estimation methods for ICCs (or generalizability coefficients), utilizing Markov chain Monte Carlo (MCMC) estimation of hierarchical linear models (LoPilato et al., 2015; Ten Hove et al., 2022a, 2020), maximum likelihood estimation of random effects models (Marcoulides,

(a) Complete Two-Way Design

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}

(b) Incomplete Two-Way Design

Subject	Rater		
	1	2	3
1	y_{11}	y_{12}	–
2	y_{21}	y_{22}	–
3	y_{31}	y_{32}	–
4	y_{41}	–	y_{43}
5	y_{51}	–	y_{53}
6	y_{61}	–	y_{63}
7	–	y_{72}	y_{73}
8	–	y_{82}	y_{83}
9	–	y_{92}	y_{93}

(c) 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}

Figure 5.1: Example of Three Observational Designs.

1990; Ten Hove et al., 2022a), and maximum likelihood estimation of common-factor models (Vispoel et al., 2018a, 2019; Jorgensen, 2021). Each of these methods can handle incomplete observational designs. These methods all assume that the unobserved ratings are missing random (MAR; R. J. A. Little & Rubin, 2002, chapter 15), although even the more restrictive missing-completely-at-random (MCAR) assumption is satisfied by planned missingness designs. However, the computational accuracy and feasibility of these methods yet needs to be investigated and compared, especially for typical conditions in observational studies such as incomplete designs and small samples of subjects and raters.

In this paper, we investigated which of the proposed estimation methods performs best in terms of computational accuracy and feasibility when estimating ICCs for IRR from incomplete observational designs. First, we describe the various definitions of ICCs for IRR. Second, we describe the three novel approaches to estimate ICCs from continuous data: MCMC estimation of hierarchical linear models (MCMC-HL), maximum likelihood estimation of random-effects models (MLE-RE), and maximum likelihood estimation of common-factor models (MLE-CF), and possible extensions to discrete data. Third, we describe a simulation study in which we compared the three novel methods in various conditions that differed in sample sizes of raters and subjects, proportions of missing observations, and magnitude of variance components. We focused on estimating ICCs from continuous responses, but also payed some attention to discrete data by including conditions with binary responses. Fourth, we provide an empirical illustration and soft-

ware tutorial. In this tutorial, we compare a conflated ANOVA approach and the best performing method on data from Yuen et al. (2020), who evaluated the communication skills of clinicians in training using a planned incomplete two-way design. We conclude with a discussion of our results, in which we suggest directions for future research. On the Open Science Framework (OSF), we provide all software code to replicate our analyses, and software that allows researchers to apply the proposed methods to their own observational data.

5.2 Definition of ICCs

5.2.1 Decomposition of Observations

The ICC literature defined ICCs for two-way (i.e., crossed) designs and one-way (i.e., nested) designs. In a two-way design, S subjects are each rated once by R raters in a crossed design. If this design is complete, all subjects are rated by all raters (Figure 5.1, panel a). If this design is incomplete, the raters partly vary across subjects (Figure 5.1, panel b). 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 attribute y. 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 raters rate each subject only once, μ_{sr} also includes random error (Cronbach et al., 1963). It is assumed that these effects are uncorrelated and that y_{sr} can be decomposed as

$$y_{sr} = \mu + \mu_s + \mu_r + \mu_{sr}. \tag{5.1}$$

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

$$\sigma^2_{y_{sr}} = \sigma^2_s + \sigma^2_r + \sigma^2_{sr}. \tag{5.2}$$

A one-way design is a special case of an incomplete two-way design, in which each subject has a unique set of raters (Figure 5.1, panel c; Ten Hove et al., 2022c). Because there is no overlap of raters across subjects, μ_r and μ_{sr} are indistinguishable. Let $\mu_{r:s}$ indicate the effect of rater r that is nested (:) within subject s, which is confounded with interaction variance and random-error variance. $y_{s:r}$ can therefore be decomposed as

$$y_{r:s} = \mu + \mu_s + \mu_{r:s}. \tag{5.3}$$

The variance in multiple observations $y_{r:s}$ can then be decomposed as

$$\sigma^2_{y_{r:s}} = \sigma^2_s + \sigma^2_{r:s}. \tag{5.4}$$

We focus on estimating ICCs from data with an incomplete two-way design, for which estimation methods are currently lacking. For more information about ICCs for one-way designs, see Shrout and Fleiss (1979), McGraw and Wong (1996), and Ten Hove et al. (2022c).

5.2.2 Variation in ICC Definitions

The variance components in Equation 5.2 are used to define several ICCs for IRR (e.g., Bartko, 1966; Shrout & Fleiss, 1979; McGraw & Wong, 1996; Putka et al., 2008; Ten Hove et al., 2022c). Each of these ICCs expresses the proportion of subject variance relative to the subject variance plus some error variance. The basis of any of these ICCs is the ICC of interrater agreement and the ICC of interrater consistency of single ratings in a complete two-way design. ICCs of interrater agreement for complete two-way designs are defined as the proportion of subject variance relative to the subject variance plus both rater-related variance components:

$$\text{ICC(A,1)} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_r + \sigma^2_{sr}}. \tag{5.5}$$

ICCs of interrater agreement express the degree to which subjects' absolute scores can be generalized over raters and are of interest when ratings are given absolute interpretations. Examples of absolute interpretations are individual assessments using educational test, diagnostic instrument, or job assessment to determine the subjects' level on an attribute.

ICCs of interrater consistency for complete two-way designs do not include the variance of the main-rater effects and are defined as the proportion of subject variance relative to the subject variance plus the interaction variance between raters and subjects, which is confounded with random error:

$$\text{ICC(C,1)} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{sr}}. \tag{5.6}$$

ICCs of interrater consistency express the degree to which observed differences across subjects can be generalized over raters and are of interest when ratings are given relative interpretations, as is the case when estimating correlations or regression coefficients.

Variations on the ICCs in Equations 5.5 and 5.6 differ in how the rater-related (i.e., error) variance components are scaled. ICCs for average ratings from a complete design divide the rater-related components by the total number of raters (k). ICCs for incomplete observational designs divide the rater-related components by the harmonic-mean number of raters per subject ($\hat{k}$) instead of k, and add a portion of main-rater variance (σ^2_r) to

the denominator of the ICC for interrater consistency, the size of which (q) depends on the proportion of non-overlapping raters across subjects. Table 5.1 shows an overview of all ICCs for two-way designs.

This paper focuses on the estimation of ICCs from incomplete data rather than the definition of ICCs for incomplete designs. We therefore refer to Ten Hove et al. (2022c) for a more detailed explanation on the different definitions of ICCs and for guidance in choosing between them. We want to emphasize, though, that the definition of an ICC for IRR depends on how ratings are used in practice. The observational design in practice may differ from the design that is used to estimate the ICCs. For example, in educational practice, students may obtain only a single thesis grade from a single supervisor. However, to estimate the variance components to derive the ICCs, the data should contain at least two ratings per student. A reliability study could then be conducted in which a sample of students' theses is graded by multiple raters each. The ICC($A, 1$) estimated with that sample would describe the reliability of single-supervisor grades obtained in practice. Thus, the design of such a reliability study does not need to match how observations are actually made in practice. From both complete and incomplete two-way data, a researcher can *derive* all ICCs as defined in Table 5.1, as long as the variance components in Equation 5.2 can be estimated. The following section describes three novel estimation methods that enable estimation of these variance components and ICCs from (in)complete data.

Table 5.1: Intraclass Correlation Coefficients (ICC) for Two-Way Designs

Type	Design	Single ratings		Average ratings	
Agreement	Complete	ICC($A, 1$)	$= \frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2+\sigma_{sr}^2}$ [a,b]	ICC(A, k)	$= \frac{\sigma_s^2}{\sigma_s^2+(\sigma_r^2+\sigma_{sr}^2)/k}$ [a,b]
	Incomplete	ICC($A, 1$)	$= \frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2+\sigma_{sr}^2}$ [a,b]	ICC($A, \hat{k}$)	$= \frac{\sigma_s^2}{\sigma_s^2+\frac{\sigma_r^2+\sigma_{sr}^2}{\hat{k}}}$ [d]
Consistency	Complete	ICC($C, 1$)	$= \frac{\sigma_s^2}{\sigma_s^2+\sigma_{sr}^2}$ [b]	ICC(C, k)	$= \frac{\sigma_s^2}{\sigma_s^2+\sigma_{sr}^2/k}$ [b]
	Incomplete	ICC($Q, 1$)	$= \frac{\sigma_s^2}{\sigma_s^2+q\sigma_r^2+\sigma_{sr}^2}$ [d]	ICC($Q, \hat{k}$)	$= \frac{\sigma_s^2}{\sigma_s^2+q\sigma_r^2+\sigma_{sr}^2/\hat{k}}$ [c]

Note. ICCs as defined by [a] Shrout and Fleiss (1979); [b] McGraw and Wong (1996); [c] Putka et al. (2008); [d] Ten Hove et al. (2022c).

5.3 Novel Estimation Methods for ICCs

5.3.1 MCMC Estimation of Hierarchical Linear Models: MCMC-HL

LoPilato et al. (2015) proposed using MCMC-HL to estimate generalizability coefficients (equivalent to an ICC of interrater consistency, Equation 5.6, when the only facet of generalization is raters). In this model, observations are cross-classified within both

subjects and raters, that are specified as random effects. With MCMC algorithms, all unknown quantities in the two-way model in Equations 5.1 and 5.2 can be estimated simultaneously by sampling from a probability distribution. The method readily provides measures of uncertainty in terms of posterior *SD*s and Bayesian credible intervals (BCIs), the Bayesian counterparts of *SE*s and CIs in frequentist statistics, respectively (Gelman & Rubin, 1992).

MCMC has the advantage that it does not assume normal sampling distributions of variance components, nor does it depend on asymptotic theory (Gelman et al., 2013). However, when random effects are treated as parameters that are estimated—as is the case when estimating ICCs—the corresponding variance components (i.e., those in Equation 5.2) are hyperparameters that require a hyperprior distribution. The specification of these hyperprior distributions can vary from informative, by incorporating prior beliefs about the parameter of interest, to uninformative, which allows the posterior to be influenced overwhelmingly by the data. When estimating ICCs under IRR-specific conditions, weakly informative half-t hyperprior distributions for the variance components yield desirable results in terms of computational accuracy (i.e., minimal bias of point estimates and posterior *SD*s; Ten Hove et al., 2020).

5.3.2 Maximum Likelihood Estimation of Random Effects Models: MLE-RE

Marcoulides (1990; cf. Jiang, 2018) proposed MLE-RE to estimate generalizability coefficients. As in the MCMC-HL approach, observations are cross-classified within both subjects and raters, so raters and subjects have random effects. When using MLE, the delta method (e.g. Oehlert, 1992) is often used to obtain *SE*s and Wald-based CIs of (functions of) random effect parameters. Alternatively, re-sampling techniques such as nonparametric bootstrapping (which can be computationally intensive), or Monte-Carlo CIs (a parametric bootstrap technique; Preacher and Selig, 2012) can be used to obtain CIs. Monte-Carlo CIs may be more robust than delta-method CIs for they only assume a normal sampling distribution for the estimated parameters, and thus not for the ICCs that are functions of those parameters. Using Monte-Carlo techniques, the CIs of the ICCs are based on percentiles of the empirical sampling distribution, which need not be normal (Preacher & Selig, 2012).

5.3.3 Maximum Likelihood Estimation of Common-Factor Models: MLE-CF

Marcoulides (1996) proposed to use maximum likelihood estimation of a structural equation model (a common-factor model) to estimate generalizability coefficients (also, see Vispoel et al., 2018a, 2019). In this structural equation model, raters are specified as indicators of a latent variable, with constraints on the error-variance components and

factor loadings to model the interchangeability of raters. Jorgensen (2021) showed how additional constraints on the intercepts of the indicators (i.e., the rater effects) can be used to also estimate indices of dependability—the generalizability theory counterpart of ICCs for interrater agreement—and how *SE*s and CIs of the generalizability coefficients and indices of dependability can be obtained using the same methods as for MLE-RE.

5.3.4 Discrete Responses

Traditionally, generalizability theory handles discrete data by treating it as continuous (Bock et al., 2002), which makes the three described methods readily applicable to discrete responses. However, all three methods have extensions that are particularly developed for discrete responses and do not assume continuous ratings. For MCMC-HL and MLE-RE, a generalized hierarchical model with a logit or probit link function can be used to estimate the variance components and derive the ICCs (Agresti, 2010, chapter 13). A latent continuous variable is then assumed to underlie the discrete responses, and the residual variance (σ^2_{sr}) is fixed to $\pi^2/3$ (for a logit link) or 1 (for a probit link) for identification. The choice of scaling does not affect the ICC estimates because the other variance components are estimated relative to this fixed residual variance. MCMC-HL readily provides measures of uncertainty for discrete responses, just as for continuous responses (i.e., based on the posterior distribution). MLE-RE cannot provide measures of uncertainty for the ICCs, because robust asymptotic covariance matrices have yet to be derived for two-way designs (Wang & Merkle, 2018). MLE-CF can estimate the ICCs from discrete data using a latent-response metric (Jorgensen, 2021; Vispoel et al., 2019) in an item factor analysis approach (Kamata & Bauer, 2008). This approach is equivalent to specifying a probit link as in the MLE-RE and MCMC-HL approach (for details see Jorgensen, 2021).

5.3.5 Pros and Cons of the Novel Estimation Methods for ICCs

The three recently proposed methods to estimate ICCs (or generalizability coefficients) differ in their strengths and weaknesses. First, compared to MCMC-HL and MLE-RE, we expect MLE-CF to provide estimates that have less accuracy and worse coverage rates of ICCs for interrater agreement. Because MLE-CF treats rater effects as fixed effects rather than random effects, MLE-CF is expected to provide less accurate estimates and worse coverage rates of the rater variance. Second, for small samples, Bayesian estimation methods, such as MCMC-HL, are expected to yield greater accuracy and better coverage rates than maximum-likelihood estimation methods, such as MLE-RE and MLE-CF (e.g. Kruschke et al., 2012; Rupp et al., 2004), especially if the sampling distribution of the variance components is non-normal or if the variance components are close to zero. Ten Hove et al. (2020) showed that for small samples and using a half-t prior distribution, MCMC-HL provided satisfactory results with respect to accuracy of the

estimated variance components. However, for the specific case of two raters, MCMC-HL cannot estimate rater variance. Because prior specification becomes more important in Bayesian estimation as samples become smaller (Smid et al., 2019), MCMC-HL provides worse estimates in terms of accuracy when other priors than the half-t prior are selected (Ten Hove et al., 2020). Third, MCMC-HL is a resampling method and typically requires more computation time than the maximum likelihood estimation methods MLE-RE and MLE-CF. For sparser wide-format data, MLE-CF requires more computation time than MLE-RE. For MLE-CF, the computation time increases exponentially as the number of columns in the data (here the raters) increase. Fourth, whereas MCMC-HL and MLE-RE are expected to converge well, Jorgensen (2021) showed that for binary responses MLE-CF had convergence issues when applied to incomplete data.

5.4 Simulation Study

Using simulated data, we compared the ICCs estimated by MCMC-HL, MLE-RE, and MLE-CF on computational accuracy and feasibility under different levels of the sample sizes, magnitude of variance components, and measurement level of the attribute. We chose realistic design factors, inspired by empirical research of Zee et al. (2020); Majdandžić et al. (2021); Yuen et al. (2020). We evaluated the computational accuracy in terms of the bias of point estimates, bias of variability estimates, the root mean squared error, and coverage rates. We evaluated the computational feasibility in terms of convergence rates and estimation time. Our focus was on the performance of the estimators for continuous data because ICCs and their traditional estimation methods were originally proposed for continuous data. However, we were also interested to see how the novel MCMC and maximum likelihood approaches performed when applied to binary data, because many applied studies use observation instruments with binary response options (Yuen et al., 2020). We programmed the full simulation study in the R software environment (R Core Team, 2019) and performed all simulations on the Lisa Cluster of the Dutch national e-infrastructure (Surfsara, n.d.). All software used for this simulation study is publicly available on the OSF: https://osf.io/j5b8t/.

5.4.1 Methods

Data Generation

We generated normally distributed data from Equation 5.1 using the parameters in Equation 5.2. We fixed the population mean to $\mu = 0$, and that of the subject variance to $\sigma_s^2 = 2$. We varied the other variance components, as well as the size of the rater and subject pools, the number of raters per subject, and the measurement level of the attribute.

Independent Variables

The size of the rater pool (R) had three levels: 3, 5, and 10; the number of raters per subject (R_s) had two levels: 2 and 3; the number of subjects (S) had two levels: 30 and 200; the magnitude of rater variance (σ_r^2) had two levels: $\frac{1}{2}$ and 1; the magnitude of interaction + measurement-error variance (σ_{sr}^2) had two levels: 1 and 2. The measurement level of the attribute had two levels: continuous and binary. We created a binary attribute by splitting the generated continuous data using the population grand mean ($\mu = 0$) as threshold. The simulation design was fully crossed, resulting in 3 (R) $\times$ 2 (R_s) $\times$ 2 (S) $\times$ 2 (σ_r^2) $\times$ 2 (σ_{sr}^2) $\times$ 2 (measurement level) = 96 conditions, for each of which we simulated 1,000 data sets, yielding 96,000 unique data sets. The population ICCs for single ratings per condition varied from ICC($A, 1$) = 0.40-0.57 and ICC($C, 1$) = 0.50-0.67. These are the lowest possible ICCs. ICCs for average ratings are higher by definition because the rater-related terms are divided by the number of raters per subject. The proportion of missing observations ($1 - \frac{R_s}{R}$) varied from 0.00-0.80 (Table 5.2).

Table 5.2: Observational Designs in the Simulation Study

R	R_s	Missing Observations (%)
3	2	33
3	3	0
5	2	60
5	3	40
10	2	80
10	3	70

Estimation

MCMC-HL. We used the R package `brms` (Bürkner, 2017) to estimate the hierarchical linear model (HLM) with MCMC, and the `rstan` package (Stan Development Team, 2018) to obtain the posterior draws of the variance parameters from which we derived the ICCs. We used the default hyperpriors of the `brms`() function, which are similar to those proposed by Ten Hove et al. (2020) for IRR-specific conditions. We used three independent chains of 1,000 iterations, half of which served as burn-in iterations. The other half were saved in each chain, resulting in a sample of 1,500 iterations to estimate each ICC. We used the potential scale reduction factor $0.90 < \hat{R} > 1.10$ for each of the ICCs as criterion for convergence (Gelman et al., 2013). We used posterior modes as point estimators, and percentile-based BCIs (e.g., Ten Hove et al., 2020). For the binary responses, we used the same approach but we used a logit link function to map the discrete item responses to a continuous latent scale.

MLE-RE. We used the R package `lme4` (Bates et al., 2015) to estimate the hierarchical linear model (HLM) with MLE-RE. We used the `merDeriv` package (Wang & Merkle,

2018) to obtain a robust asymptotic covariance matrix of the variance components, which served as input for the `deltaMethod`() function of the `car` package (Fox & Weisberg, 2019) that we used to compute *SE*s and delta-method CIs for the ICCs. The robust asymptotic covariance matrix of the variance components also served as input for the `monteCarloCI`() function of the `semTools` package (Jorgensen et al., 2021) that we used to compute Monte-Carlo CIs for the ICCs. For the binary responses, we used the same approach but we used a logit link function to map the discrete item responses to a continuous latent scale. However, as explained above in the *Discrete Responses* subsection, we could not compute *SE*s for the ICCs from binary responses because a robust asymptotic covariance matrix is not available for cross-classified (i.e., two-way) models (Wang & Merkle, 2018).

MLE-CF. We used the R package `lavaan` (Rosseel, 2012) to estimate the ICCs with MLE-CF, using the constraints as proposed by Jorgensen (2021) to identify the model and estimate the error components. We specified the ICCs as user-defined parameters in the `lavaan` model so that the program readily provided delta-method CIs for the ICCs. We computed Monte-Carlo CIs using the R package `semTools` (Jorgensen et al., 2021). For the binary responses, we used the same approach but we used diagonally weighted least-squares (DLWS) estimation with the theta parameterization (i.e., residual variances fixed to 1) to map the discrete item responses to a continuous latent scale (equivalent to a probit link).

Dependent Variables

We computed the relative bias of the point estimates, the relative bias of the variability estimates, the root mean squared error, and the 95% BCI and CI coverage rates of the estimated ICCs across conditions to assess the computational accuracy of the different methods. We investigated these characteristics for the ICC(A,1) and ICC(C,1) only. The other ICCs definitions in Table 5.1 scale the error variance by means of k or $\hat{k}$ and q, hence only proportionally influence the computational accuracy of the ICCs. We computed the convergence rates and estimation time of each method to investigate the computational feasibility of the different methods.

Bias of Point Estimates. Let $\bar{\theta}$ denote the average ICC as estimated with either MCMC-HL, MLE-RE, or MLE-CF across replications in a condition, and let θ denote the population parameter in that condition. Relative bias was computed as $\frac{\bar{\theta}-\theta}{\theta}$, and we interpreted relative bias > .05 as minor bias and relative bias > .10 as substantial bias.

Bias of Variability Estimates Let $\overline{SE}$ denote the average posterior *SD* of the ICC (for MCMC-HL) or the average *SE* of the ICC (for MLE-RE and MLE-CF). Let SD_θ denote the *SD* of the point estimates of the ICC in a condition. Bias of variability estimates of the ICCs was computed as *SE* bias = $\frac{\overline{SE}}{SD_\theta} - 1$. Preferably, this value equals zero, indicating accurate estimates of variability. We interpreted absolute *SE* bias ¿ 0.10

as minor bias of variability estimates, and absolute SE bias > .20 as substantial bias of variability estimates.

Root Mean Squared Error (RMSE). Let $\hat{\theta}_j$ denote the estimated ICC in converged replication j of a condition, let θ denote the population parameter in that condition, and let J be the number of converged replications. Then, $RMSE = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta)^2}{J}}$. Both the bias and variance contribute to the RMSE, that represents the bias-variance trade-off and is ideally low.

BCI and CI Coverage Rates. For the 95% BCI and CI coverage rates of the ICCs, we considered a coverage rate < 90% to be insufficient.

Convergence and Estimation Time. We computed the percentage converging models and the average estimation time of converging models across conditions for each estimation method.

Preferred Characteristics

To select a preferred method to estimate ICCs for IRR, we considered the bias of point estimates and bias of variability estimates as the most important criteria. If the least biased point estimates and the least biased variability estimates were not produced by the same method, we preferred the method with the lowest RMSE. If multiple methods had similar results with respect to bias in point estimates and bias of variability estimates, we used estimation time and convergence to select the most practical feasible method. We do not prefer methods that fail to provide measures of uncertainty.

5.4.2 Results

An analysis of variance revealed that magnitude of error variance (i.e., σ_r^2 and σ_{sr}^2) had little influence on the estimation of ICCs compared to the effect of sample sizes of raters and subjects (i.e., K, k_s and N). We therefore ignored these design factors, and discuss the results for conditions with low error-variance components (i.e., $\sigma_r^2 = 0.50$ and $\sigma_{sr}^2 = 1.00$) only. We do not discuss the RMSE of the methods, because the bias of point estimates and the bias of variability estimates point to the same preferred method. Tables A1 and A2 in Appendix A describe all dependent variables across all conditions with continuous and binary responses, respectively.

Bias of Point Estimates

Figure 5.2 (upper panels) shows the relative bias in the ICCs across estimators in conditions with continuous responses. MCMC-HL underestimated the ICC$(A, 1)$ in most conditions—apart from those with many missing observations and large numbers of

subjects—but accurately estimated the ICC($C, 1$) in all conditions. MLE-RE accurately estimated the ICC($A, 1$) and ICC($C, 1$) in all conditions. MLE-CF accurately estimated the ICC($A, 1$) and ICC($C, 1$), apart from the condition with 80% missing observations in which the ICC($C, 1$) was slightly overestimated. Across all conditions with continuous responses, MLE-RE yielded the most accurate ICC estimates.

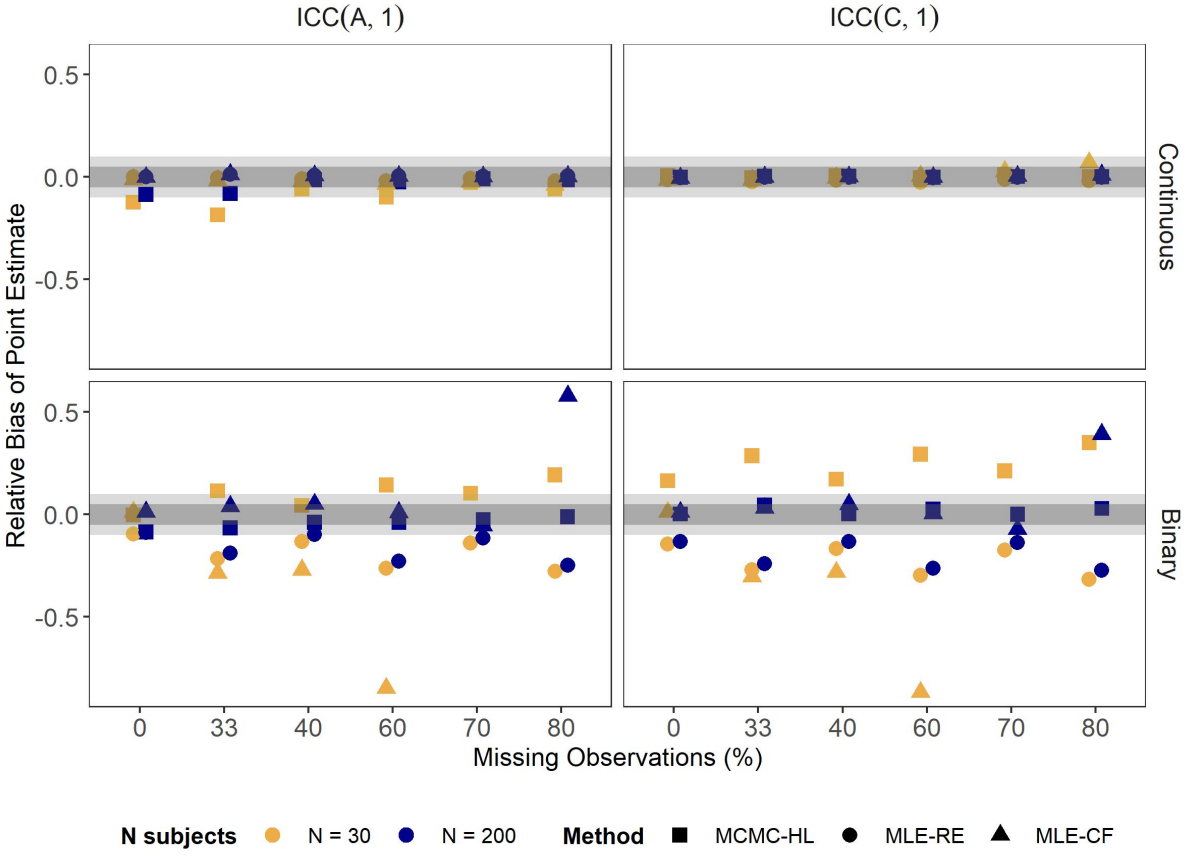

Figure 5.2: Relative bias of the ICCs across estimators in conditions with $\sigma_r^2 = \frac{1}{2}$, $\sigma_{sr}^2 = 1$, and continuous responses (upper panels) or binary responses (bottom panels). MCMC-HL = Markov chain Monte Carlo estimation of hierarchical linear models; MLE-RE = Maximum Likelihood Estimation of random-effects models; MLE-CF = Maximum Likelihood Estimation of common-factor models. The percentage missing observations on the x-axis is computed as $100 \times (1 - \frac{k_s}{K})$. White areas: substantial bias (>10%); Light-gray areas: minor bias (5–10%); Dark-gray areas: negligible bias (< 5%).

Figure 5.2 (bottom panels) shows the relative bias in the ICCs across estimators in conditions with binary responses. MCMC-HL accurately estimated the ICC($A, 1$) and ICC($C, 1$) in most conditions with a large number of subjects and slightly overestimated these in conditions with few subjects. MLE-RE substantially underestimated the ICC($A, 1$) and ICC($C, 1$) in almost all conditions. MLE-CF severely underestimated estimated the ICC($A, 1$) and ICC($C, 1$) in conditions with an incomplete design and a small sample of subjects. In conditions with a large sample of subjects, the method

accurately estimated the ICCs, apart from the condition with the highest percentage missing observations, for which it severely overestimated the ICCs. Averaged across all binary conditions, MCMC-HL yielded the most accurate ICC estimates.

Bias of Variability Estimates

Figure 5.3 (upper panels) shows the relative bias of variability estimates of the ICCs across estimators in conditions with continuous responses. MCMC-HL's posterior *SD*s accurately estimated the variability of the ICC($\boldsymbol{A}, 1$) in most conditions with few subjects and of the ICC($\boldsymbol{C}, 1$) in all conditions. However, MCMC-HL's posterior *SD*s overestimated the variability of the ICC($\boldsymbol{A}, 1$) in conditions with many subjects. MLE-RE's *SE*s accurately estimated the variability of the ICC($\boldsymbol{A}, 1$) and ICC($\boldsymbol{C}, 1$) in all conditions. MLE-CF's *SE*s underestimated the variability of ICC($\boldsymbol{A}, 1$), especially in conditions with many subjects, but accurately estimated the variability of the ICC($\boldsymbol{C}, 1$) apart from the condition with the highest percentage missing observations and few subjects. Across conditions with continuous responses, MLE-RE yielded the most accurate variability estimates of the ICCs.

Figure 5.3 (bottom panels) shows the relative bias of variability estimates of the ICCs across estimators in conditions with binary responses. MCMC-HL's posterior *SD*s overestimated the variability of ICC($\boldsymbol{A}, 1$) in most conditions with a substantial number of subjects and underestimated the variability of the ICC($\boldsymbol{A}, 1$) and ICC($\boldsymbol{C}, 1$) in most conditions with few subjects. This method yielded accurate ICC($\boldsymbol{C}, 1$) variability estimates in conditions with a substantial number of subjects. MLE-RE did not yield *SE*s for the ICCs from binary data. MLE-CF's *SE*s underestimated the variability of the ICC($\boldsymbol{A}, 1$) and ICC($\boldsymbol{C}, 1$) in almost all conditions. Averaged over all conditions with binary responses, MCMC-HL yielded more accurate variability estimates of the ICCs than MLE-CF.

95% BCI and CI Coverage Rates

Figure 5.4 (upper panels) shows the 95% BCI and CI coverage rates of the ICCs across estimators in conditions with continuous responses. MCMC-HL and MLE-RE (with delta-method as well as Monte-Carlo CIs) yielded acceptable (but not perfect) BCI and CI coverage rates of the ICC($\boldsymbol{A}, 1$), in nearly all conditions with continuous responses and a substantial number of subjects, but not in conditions with high percentages of missing observations. MLE-CF yielded extremely low 95% CI coverage rates of both delta-method CIs and Monte-Carlo CIs. MCMC-HL and both maximum likelihood based methods with Monte-Carlo CIs yielded acceptable (but not perfect) BCI and CI coverage rates of the ICC($\boldsymbol{C}, 1$), in nearly all conditions. Only in the condition with the highest percentage of missing observations, the 95% BCI and CI coverage rates of the ICC($\boldsymbol{C}, 1$) were too low, as was the case for the both maximum likelihood based methods with delta-method CIs. Across conditions with continuous responses, MLE-RE with Monte-Carlo CIs yielded coverage rates closest to the nominal level of 95%.

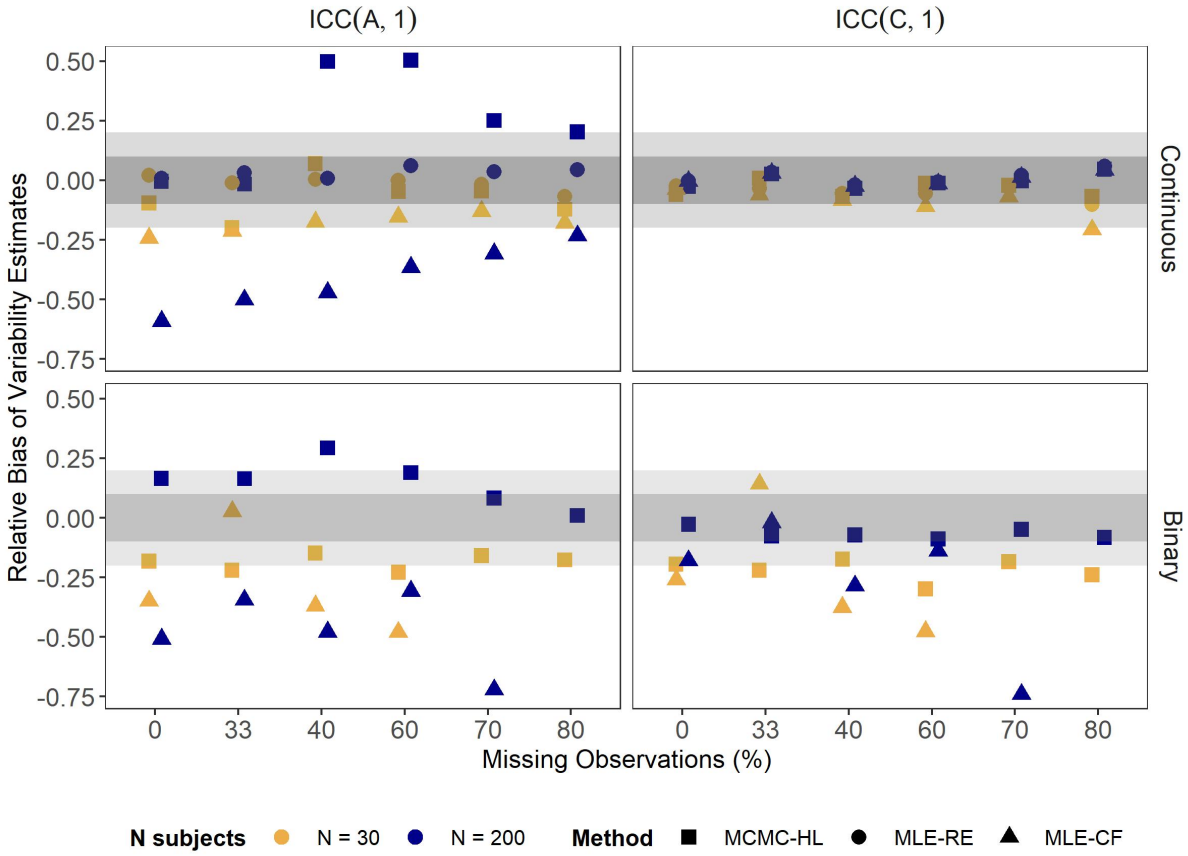

Figure 5.3: Relative bias of variability estimates of the ICCs across estimators in conditions with $\sigma_r^2 = \frac{1}{2}$, $\sigma_{sr}^2 = 1$, and continuous responses (upper panels) or binary responses (bottom panels). MCMC-HL = Markov chain Monte Carlo estimation of hierarchical linear models; MLE-RE = Maximum Likelihood Estimation of random-effects models; MLE-CF = Maximum Likelihood Estimation of common-factor models. The percentage missing observations on the x-axis is computed as $100 \times (1 - \frac{k_s}{K})$. White areas: substantial bias (>20%); Light-gray areas: minor bias (10–20%); Dark-gray areas: negligible bias (< 10%).

Figure 5.4 (bottom panels) shows the 95% BCI and CI coverage rates of the ICCs across estimators in conditions with binary responses. MCMC-HL yielded nearly nominal coverage rates for the ICC$(A, 1)$ and ICC$(C, 1)$ in conditions with a substantial number of subjects, but the BCIs of the ICC$(C, 1)$ were too narrow in conditions with few subjects. MLE-RE did not yield CIs for the ICCs from binary data. MLE-CF with both delta-method and Monte-Carlo CIs yielded too low coverage rates for the ICC$(A, 1)$ and ICC$(C, 1)$ in nearly all conditions. Across conditions with binary responses, MCMC-HL yielded coverage rates closest to the nominal level of 95%.

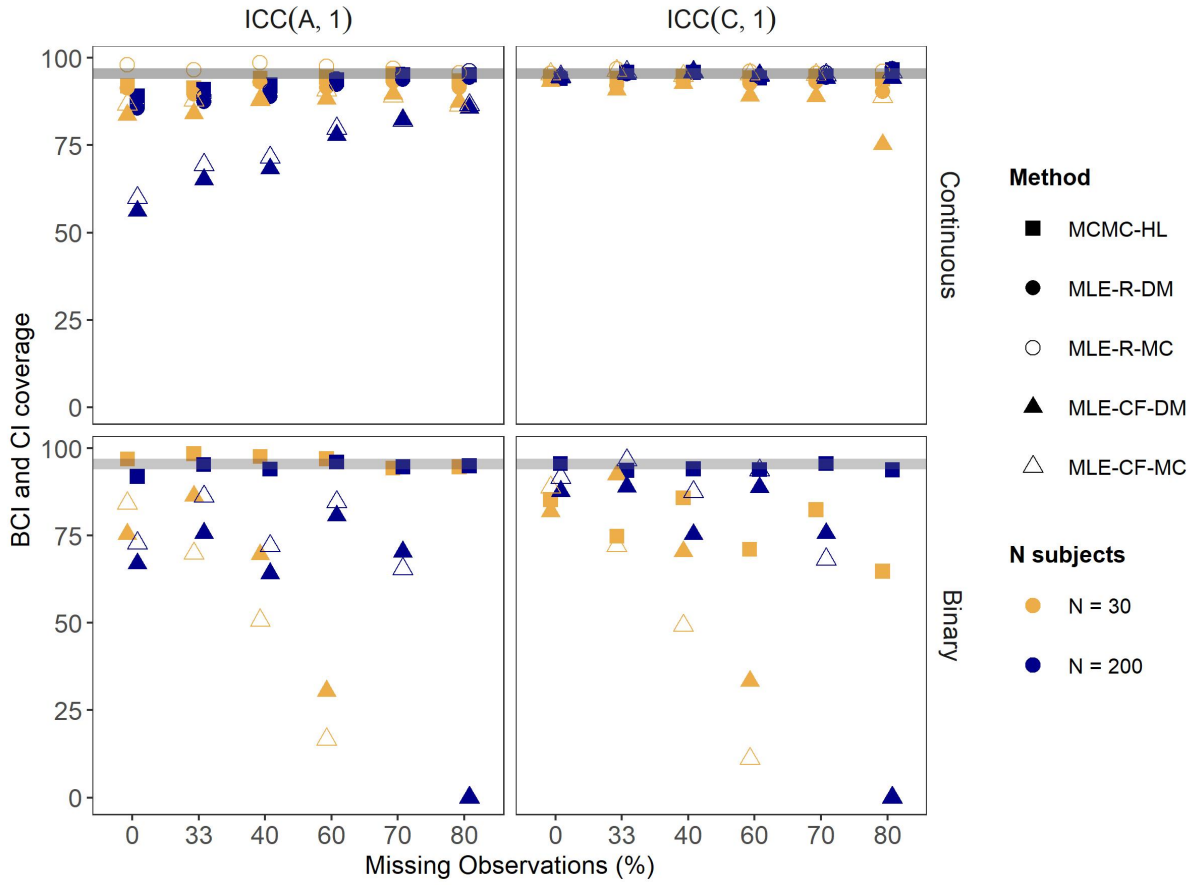

Figure 5.4: Percentage of 95% BCI and CI coverage of the ICCs across estimators in conditions with $\sigma_r^2 = \frac{1}{2}$, $\sigma_{sr}^2 = 1$, and continuous responses (upper panels) or binary responses (bottom panels). MCMC-HL = Markov chain Monte Carlo estimation of hierarchical linear models; MLE-RE = Maximum Likelihood Estimation of random-effects models; MLE-CF = Maximum Likelihood Estimation of common-factor models. DM = delta-method CIs; MC = Monte Carlo CIs. The percentage missing observations on the x-axis is computed as $100 \times (1 - \frac{k_s}{K})$.

Convergence

Figure 5.5 shows the percentage converging replications across estimators for continuous and binary responses. For continuous responses, MCMC-HL on average had the lowest convergence rates, especially in conditions with a low percentage of missing observations (i.e., those with a small rater pool). Both maximum likelihood based methods converged for almost all replications. However, MLE-CF had low converge rates in the conditions with the highest percentage of missing observations and a small number of subjects.

For binary responses, MCMC-HL converged for most (but not all) replications, with slightly lower convergence rates in conditions with a low percentage missing observations (i.e., conditions with a small rater pool). MLE-RE converged for almost all replications. MLE-CF converged for almost all replications in most conditions with few missing observations but did not convergence for most replications in conditions with high percentages

of missing observations.

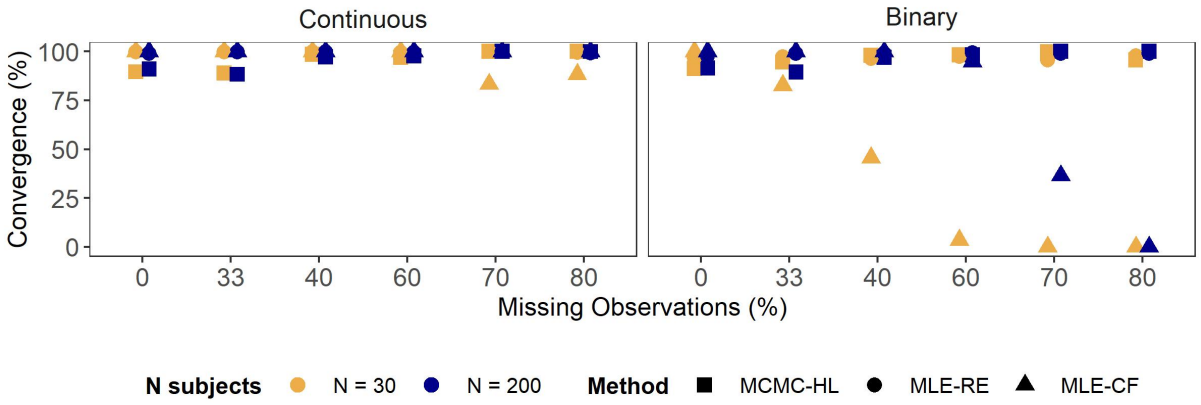

Figure 5.5: Percentage converging replications across methods in conditions with $\sigma_r^2 = \frac{1}{2}$, $\sigma_{sr}^2 = 1$, and continuous responses (left panel) or binary responses (right panel). MCMC-HL = Markov chain Monte Carlo estimation of hierarchical linear models; MLE-RE = Maximum Likelihood Estimation of random-effects models; MLE-CF = Maximum Likelihood Estimation of common-factor models. The percentage missing observations on the x-axis is computed as $100 \times (1 - \frac{k_s}{K})$.

Estimation Time

Figure 5.6 shows the estimation time in log(seconds) across estimators for continuous and binary responses. For both maximum likelihood approaches, we only present results for the Monte-Carlo CIs because these were preferred based on their coverage rates, and the estimation time differed negligibly between delta-method and Monte-Carlo CIs. For both types of responses, MCMC-HL required a substantial amount of estimation time (i.e., minutes for continuous responses and seconds to minutes for binary responses). MLE-RE converged within a few seconds. MLE-CF required much estimation time, which increased from minutes to hours for continuous responses with a high proportion of missing observations.

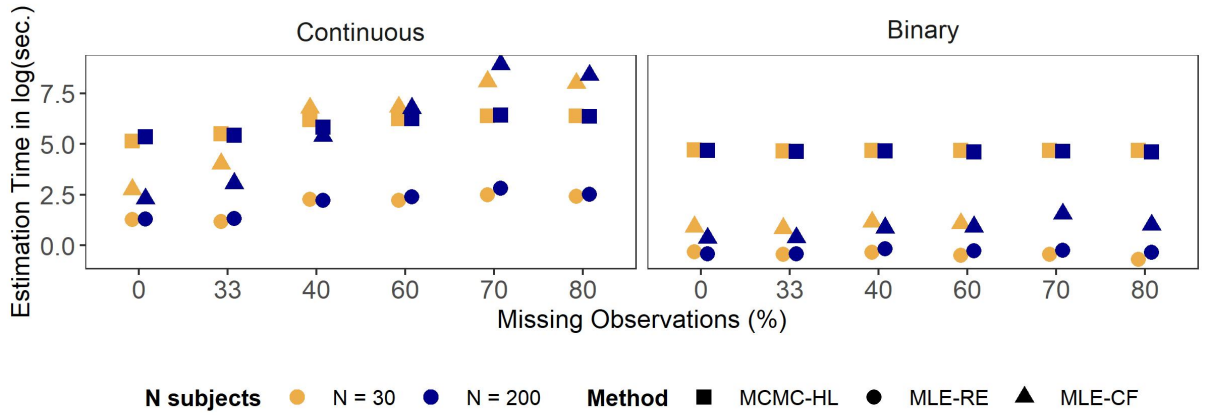

Figure 5.6: Estimation time in log(seconds) across methods in conditions with $\sigma_r^2 = \frac{1}{2}$, $\sigma_{sr}^2 = 1$, and continuous responses (left panel) or binary responses (right panel). MCMC-HL = Markov chain Monte Carlo estimation of hierarchical linear models; MLE-RE = Maximum Likelihood Estimation of random-effect models; MLE-CF = Maximum Likelihood Estimation of common-factor models. The percentage missing observations on the x-axis is computed as $100 \times (1 - \frac{k_s}{K})$.

5.5 Tutorial and Empirical Example

We used data from Yuen et al. (2020), who developed an instrument to assess the advance care planning (ACP) communication skills of clinicians in training. Both these data and the R functions and R syntax required to replicate these analyses or to estimate ICCs from different data are available on the OSF: https://osf.io/j5b8t/.

5.5.1 Sample

The ACP data are provided in *long format*. Each row of data includes one observation (in the column `score`) for a single subject as assessed by a single rater. The data thus have multiple rows per subject, the number of which depends on how many raters assessed the subject:

```
> long <- read.table("ACPdata.txt")
> head(long)
 subject rater score
1  1     2     1
2  1     1     2
3  2     3     3
4  2     1     4
5  3     4     2
6  3     5     3
> ## Inspect the Sample
> # Sample size of subjects
```

```
> length(unique(long$subject))
[1] 29
> # Rater Pool
> length(unique(long$rater))
[1] 6
# Raters per subject
> unique(colSums(table(rater = long$rater, subject = long$subject)))
[1] 2
```

The study used an incomplete two-way observational design, in which 29 subjects were each assessed by 2 raters, who partly overlapped across subjects. The rater pool consisted of 6 raters in total.

5.5.2 Estimating IRR from Incomplete Data

Inappropriately using Traditional ANOVA-Based Methods

The `icc()` function of the `irr` package requires data to be presented in *wide format*. In wide-format data, all scores for a single subject are presented in a single row, with a column for each unique rater. We transformed the long-format data as follows:

```
> wide <- reshape(long,  v.names = "score",
+                 timevar = "rater", idvar = "subject",
+                 direction = "wide")
> head(wide)
    subject score.2 score.1 score.3 score.4 score.5 score.6
1         1       1       2      NA      NA      NA      NA
3         2      NA       4       3      NA      NA      NA
5         3      NA      NA      NA       2       3      NA
7         4       2      NA      NA      NA      NA       3
9         5       5       4      NA      NA      NA      NA
11        6       1      NA      NA       2      NA      NA
```

The `icc()` function in the `irr` package uses list-wise deletion (cf. the `ICC` functionality of the `RELIABILITY` command in `SPSS`). Rows with missing observations are thus deleted. For the ACP data, that results in zero rows of data, because each row of data includes missing observations:

```
> # Identify number complete observations
> sum(complete.cases(wide))
[1] 0
```

The `icc()` function would thus delete all rows of data, so ICCs could not be estimated. To estimate ICCs using the one-way or two-way ANOVA approach, we could—inappropriately—present the data such that it can be treated as (a) one-way data, in which the raters per subject are treated as unique raters, or (b) complete two-way data, in which each subject's pair of raters are treated as if they were the same across subjects:

```
> # Tweak the long data into a 'complete' wide data frame:
> wideTweaked <- matrix(long$score, ncol = 2, byrow = TRUE,
+                       dimnames = list(paste0("S", 1:nrow(wide)), c("R1", "R2")))
    R1 R2
S1  1  2
S2  3  4
S3  2  3
S4  2  3
S5  5  4
S6  2  1
> # Compare to original data:
> head(long)
   subject rater score
 1       1     2     1
 2       1     1     2
 3       2     3     3
 4       2     1     4
 5       3     4     2
 6       3     5     3
```

Note that in `wideTweaked` data, Subject 1 still has the same ratings as in the `long` data (i.e., 1 and 2), which holds for every other subject. Also note that this inappropriate data transformation is only possible if the number of ratings is the same across subjects.

When researchers have such a tweaked data set, they could use the `icc()` functions from the `irr` package by treating these data as coming from a one-way or two-way design. A one-way model estimates one undifferentiated error term only, in which the rater variance and interaction variance are confounded (Equation 5.4). The two-way model treats each column as observations from a unique rater, to estimate rater effects (and the variance of these effects) using the column means. However, the two columns with ratings in the `wideTweaked` data do not represent the ratings from two unique raters R1 and R2, but each include ratings from various raters. R1 and R2 should thus be read as *Rating 1* and *Rating 2* instead of *Rater 1* and *Rater 2*. Therefore, the main-rater effects, μ_r, will be confounded with both the subject-by-rater interaction variance and error, μ_{sr}. The

main-rater variance, σ_r^2, is then part of the interaction variance, σ_{sr}^2 (cf. Moerbeek, 2004), similar to the one-way model. By including σ_{sr}^2 in it's denominator, the ICC of interrater consistency will incorrectly include the main-rater variance in it's denominator and will therefore underestimate the interrater consistency. Furthermore, the order of R1 and R2 scores in each row is arbitrary and could be randomly permuted, so a two-way model that treats each column as a distinct rater would result in estimates of main rater effects (and their variance) that vary across possible permutations. When using the ANOVA-based `icc()` functions, we therefore expect to find (on average) no differences between the estimated ICCs for one-way data, and the two-way ICCs for interrater consistency and interrater agreement.

```
> # ANOVA-based ICC(1) for one-way data:
> aov_oneway <- icc(wideTweaked, model = "oneway")
> # ANOVA-based ICC(C,1) and ICC(A,1) for two-way data:
> c1_aov <- icc(wideTweaked, model = "twoway",
+               type = "consistency", unit = "single")
> a1_aov <- icc(wideTweaked, model = "twoway",
+               type = "agreement", unit = "single")
> # Check estimates
> (ICCs <- c(ICC1 = aov_oneway$value,
             ICCc1_aov = c1_aov$value,
             ICCa1_aov = a1_aov$value))
    ICC1 ICCc1_aov ICCa1_aov
0.5603865 0.5566265 0.5593220
```

These results indeed show almost identical ICCs of interrater consistency and ICCs of interrater agreement, which are also similar to the one-way ICC. This informs us that the models did not find any main-rater effects, which would be observable in different column means for R1 and R2.

Using MLE-RE for Incomplete Data

Estimating ICCs from incomplete two-way data can be done with the `estICCs()` function we provide in the `ICCfunctions.R` document. This function requires long-format data. As input for the function, researchers should provide the column of the attribute scores of which they want to know the IRR (here `Y = "score"`), the column with subject IDs (here `subjects = "subject"`), and the column with rater IDs (here `raters = "rater"`). As estimator, researchers can choose between MCMC-HL (`estimator = "MCMC"`) and MLE-RE estimation (`estimator = "MLE"`). MLE-CF is not available because it was preferred over MCMC-HL or MLE-RE. Also, we have to indicate if Y is a continuous response (`response = "continuous"`) or a a binary response (`response = "binary"`). The ACP

`score` was continuous, for which MLE-RE was preferred. We therefore chose MLE-RE as estimator. For binary responses, we recommend using MCMC-HL as estimator.

```
> # Source functions
> source("ICCfunctions.R")
> # Compute ICCs for incomplete Data
> IRR_MLE <- estICCs(long, Y = "score", subjects = "subject", raters = "rater",
+                    estimator = "MLE", response = "continuous")
```

The `estICCs()` function output contains all ICCs and variance components that can be estimated from the data, accompanied by confidence intervals and *SE*s. Also, it provides the values of k, $\hat{k}$, and q, which were used to derive the ICCs of average ratings (i.e., the ICC(A,k), ICC$(A,\hat{k})$, ICC(C,k), and ICC$(Q,\hat{k})$), and which were computed based on the rater-subject combinations in the data:

```
> # Inspect estimates
> IRR_MLE
$ICCs
               est  ci.lower  ci.upper    ICC_se
ICCa1    0.7164607 0.4663275 0.9031543 0.09756677
ICCak    0.9381230 0.8401408 0.9825418 0.02787953
ICCakhat 0.8348117 0.6361235 0.9491487 0.06623144
ICCc1    0.8162174 0.6088185 0.9229590 0.06609010
ICCck    0.9638301 0.9036523 0.9863355 0.01535940
ICCqkhat 0.8536545 0.6726934 0.9445100 0.05613101

$variances
          est   ci.lower  ci.upper sigma_se
S_s  6.395159  2.6067348 10.174147 1.9318809
S_r  1.090926 -0.5456998  2.733138 0.8399057
S_sr 1.439958  0.6274233  2.250297 0.4146040

$raterDesign
    k  khat   Q
6.000 2.000 0.345
```

Using MLE-RE, the ICC of interrater agreement (ICC$(A,1) = 0.72$) and the ICC of interrater consistency (ICC$(A,1) = 0.82$) have different values because the random-effects model identified variance in main-rater effects (i.e., $\sigma_r^2 = 1.09$). Using MLE-RE, this variance is not confounded with the interaction variance, and the ICC$(C,1)$ is therefore

substantially higher than the ICC($A, 1$). The interrater consistency is thus higher than the interrater agreement, which we could not observe from the ANOVA-based approach.

5.6 Discussion

Our simulation study investigated the computational accuracy and feasibility of MCMC-HL, MLE-RE, or MLE-CF to estimate ICCs for IRR from incomplete observational designs. For continuous data, we showed that MLE-RE yielded the most accurate estimates and was most practically feasible. The point estimates and *SE* estimates had the least bias, and the method converged in a few seconds for almost all data sets. We recommend to complement this method with Monte-Carlo CIs that have almost nominal coverage rates. For binary data, we showed that MCMC-HL yielded the most accurate estimates and, of the two methods that provide measures of uncertainty for the ICCs, it was most practically feasible. The point estimates en posterior *SD* estimates had the least bias, and the method converged for almost all data sets. However, the method may cost minutes to obtain ICCs when the proportion of missing observations is substantial.

As with any study, our results are limited to the selected experimental conditions. However, we believe that the wide range of ICCs—based on the selected magnitude of variance components—and the sample sizes of subjects and raters that we selected reflect real situations we have encountered in practice (cf. Zee et al., 2020; Yuen et al., 2020; Majdandžić et al., 2021). Given current methodological knowledge and software availability, we found a single preferred method for continuous responses (MLE-RE) and a single preferred method for binary responses (MCMC-HL), and our conclusions varied negligibly across sample sizes of subjects and raters, or across proportions of missing observations.

The coverage rates of ICCs for interrater agreement were considerably lower for small samples of raters, than those for larger samples of raters, and were lower for interrater consistency than for interrater agreement. ICCs of interrater agreement include the main-rater variance in the denominator, which was estimated based on small sample sizes ($3-10$ raters). These small samples of raters resulted in relatively large *SE*s for the rater-variance component, but the MLE-RE approach requires the variances components to be positive. For bounded variables, CIs estimated with the delta method need not be range preserving (i.e., can include negative values) and can yield low coverage when sample sizes are small. Others have resolved this issue by transforming the bounded coefficients or parameters from the restricted sample space to the unrestricted sample space (e.g., Koopman et al., 2021; De Leeuw et al., 1990).

At this moment, estimation methods for discrete responses require further development. Avenues for further development include investigating the performance of the three selected methods for asymmetric thresholds, and ratings that have more than two categories. For binary responses, we could not recommend MLE-RE because (a) point estimates were consistently underestimated, similar to MLE-CF, and (b) measures of

uncertainty are currently unavailable for the random-effect variances and thus also for the ICCs. Such uncertainty measures require the robust asymptotic covariance matrix of the variance components from two-way observational designs with discrete responses (cf. Wang & Merkle, 2018).

MLE-CF yielded many estimation difficulties for binary responses. Models did not converge when a rater assigned the same score to all subjects they observed (i.e., when there was no within-rater variance for one of the raters). Because raters rated few subjects each in conditions with a high percentage of incomplete observations, this issue was more common for conditions with large rater pools and few raters per subject. If MLE-CF were applied to data with a-symmetric thresholds (yielding substantially unequal marginal proportions), this would be even more likely to occur because more similar responses are to be expected. Also, when discrete responses have even more ordinal categories, convergence of the MLE-CF approach is expected to deteriorate, because all categories must be observed across all raters (Jorgensen, 2021).

The fully Bayesian MCMC-HL approach provides another avenue for future IRR research and software developments. The method was not preferred over the MLE-RE approach for continuous responses, but it could provide answers to questions about the IRR that the maximum likelihood approaches cannot provide. Using the posterior distributions of the IRR coefficients, researchers could estimate the probability that the IRR is higher (or lower) than a specific cut-off value (Pfadt et al., 2022).

In sum, this paper fills a gap in the IRR literature, by exploring via simulation how ICCs for IRR are best estimated from incomplete data. Our tutorial showed how the preferred methods can be used to obtain IRR estimates, using the free software provided for researchers on the Open Science Framework.