Interrater reliability for incomplete and dependent data

ten Hove, D.

Publication date
2023

Chapter 7

Interrater Reliability for Interdependent Social Network Data

Abstract

We propose interrater reliability coefficients for observational interdependent social network data, which are dyadic data from a network of interacting subjects that are observed by external raters. Using the social relations model, dyadic scores of subjects' behaviors during these interactions can be decomposed into actor, partner, and relationship effects. These effects constitute different facets of theoretical interest about which researchers formulate research questions. Based on generalizability theory, we extended the social relations model with rater effects, resulting in a model that decomposes the data into effects of actors, partners, relationships, raters, and their interactions. We used the variances of these effects to define intraclass correlation coefficients (ICCs) that express to which degree the actor, partner, and relationship effects can be generalized over raters. We proposed Markov chain Monte Carlo estimation of a Bayesian hierarchical linear model to estimate the ICCs, and tested their bias and coverage in a simulation study. The method is illustrated using data on social mimicry.

Keywords: Bayesian hierarchical linear modeling, generalizability theory, interrater reliability, social network data, social relations model.

Chapter 7 is under review as: **Ten Hove, D.**, Jorgensen, T. D., & Van der Ark, L. A. (*Under review*). Interrater reliability for interdependent social network data: A generalizability theory approach.

7.1 Introduction

In observational research, raters rate subjects to gather information about subjects' attributes (e.g., their behavior or appearances). For example, Majdandžić et al. (2016) used raters to assess parents' (the subjects) challenging parenting behavior (the attribute). In statistical analyses, these ratings are used to answer research questions about the associations between different attributes or about the effect of predictor variables (e.g., interventions) on these attributes. It is important that the observed differences in the ratings reflect differences across the subjects' attributes rather than differences across raters and their perspectives. The degree to which the ratings are independent of raters can be estimated with interrater reliability (IRR) coefficients (e.g., Gwet, 2014; Hallgren, 2012; Zhao et al., 2013). If the IRR is low, the ratings depend on the raters. Because low IRR results in biased estimates or loss of power in statistical analyses about subjects' attributes (cf. Lord & Novick, 1968, p. 72), the rating procedure should be improved. IRR is thus vital for social and behavioural research, because low IRR may lead to faulty conclusions.

People commonly behave (e.g., smiling, bullying, mimicking) differently depending on the different persons with whom they interact, who in turn may elicit different behaviors from them (e.g., Card & Hodges, 2010; Coie et al., 1999; Salazar Kämpf et al., 2018; Simpkins & Parke, 2002). Variables that measure such behaviors are known as dyadic variables (Kenny et al., 2006). Dyads consist of two subjects, and the behaviors of subjects are dependent on the subject with whom they interact. Interdependent social network data involves dyadic data from a network of subjects. Each individual belongs to multiple dyads (e.g., a group of students consists of multiple pairs of peers) within such a network. Consequently, dyads are nested within both subjects, and measures of subjects are dependent across the dyads to which they belong.

In observational social-network research, raters rate subjects' attributes during interactions with several other subjects (e.g., Salazar Kämpf et al., 2018; Hughes et al., 2021; Huang et al., 2017), and the raters are not part of these interactions. Table 7.1 provides an example of an observational social network design, in which four students (Anna, Brooke, Charlie, and Deborah) interact with each other. These students form a social network, and each combination of subjects is a dyad (e.g., Anne and Brooke, Anna and Charlie, Anna and Deborah, etc.). The behavior of the students during their interactions with the other students is rated twice by each of three raters who are not part of the network. Let Y_{ijk} denote the number of times that subject i mimicked subject j (where $i \neq j$), as rated by rater k. Both y_{AB1} and y_{BA1} indicate a score on the dyadic variable mimicry of the dyad Brooke and Anna, as rated by Rater 1. The score y_{AB1} indicates Anna's mimicry of Brooke, as rated by Rater 1, and the score y_{BA1} indicates Brooke's mimicry of Anna as rated by Rater 1.

In social relations research, researchers could be interested in different aspects of dyadic interactions. The social relations model (SRM; Kenny, 1996; Kenny & La Voie,

Table 7.1: Example of an observational social network design with four subjects and three raters. y_{ijk} denotes the score on attribute Y of person i while interacting with person j (where $i \neq j$), rated by rater k.

Rater	Subject	Subject			
		Anna	Brooke	Charlie	Deborah
1	Anna	-	y_{AB1}	y_{AC1}	y_{AD1}
	Brooke	y_{BA1}	-	y_{BC1}	y_{BD1}
	Charlie	y_{CA1}	y_{CB1}	-	y_{CD1}
	Deborah	y_{DA1}	y_{DB1}	y_{DC1}	-
2	Anna	-	y_{AB2}	y_{AC2}	y_{AD2}
	Brooke	y_{BA2}	-	y_{BC2}	y_{BD2}
	Charlie	y_{CA2}	y_{CB2}	-	y_{CD2}
	Deborah	y_{DA2}	y_{DB2}	y_{DC2}	-
3	Anna	-	y_{AB3}	y_{AC3}	y_{AD3}
	Brooke	y_{BA3}	-	y_{BC3}	y_{BD3}
	Charlie	y_{CA3}	y_{CB3}	-	y_{CD3}
	Deborah	y_{DA3}	y_{DB3}	y_{DC3}	-

1984) decomposes dyadic variables in different facets of theoretical interest about which research questions could be formulated: actor effects, partner effects, and relationship effects. During the interaction between Anna and Brooke, the actor effect of Anna would involve how often Anna mimics her conversation partners on average (i.e., how imitative is Anna?). The partner effect of Brooke would involve how much Brooke is on average mimicked by the people she is having a conversation with (i.e., how imitable is Brooke?). Given how much Anna generally mimics her conversation partners (her actor effect), and how often Brooke is generally mimicked by her conversation partners (her partner effect), the relationship effect would then be how much more (or less) Anna mimics Brooke. Similarly, Brooke's mimicry of Anna can be decomposed into the actor effect of Brooke, the partner effect of Anna, and a relationship effect.[1] Dyadic variables have complex dependency structures. Actor and partner effects of individuals are often dependent, as are the relationship effects within a dyad (Kenny, 1996; Kenny & La Voie, 1984). For example, a subject who, on average, mimics much may also elicit more mimicry on average, and if Anna mimics Brooke more than is expected based on Anna's actor effect and Brooke's partner effect, Brooke may also mimic Anna more than is expected based on Brooke's actor effect and Anna's partner effect.

The results of the SRM may differ across raters. The IRR of the actor, partner, and relationship effects should thus be investigated to inspect to which degree these effects depend on raters. IRR methods have been proposed that can handle multilevel data, in

[1]There are two types of dyadic variables in interdependent social network data: Those that differ between dyads within a network, but are stable within dyads such as whether Joyce and Lisa shook hands, and those that may be different for actor and partner, such as how often Lisa mimicked Joyce and vice versa. In this paper, interdependent social network data refers to the latter: Dyadic variables that can differ both between and within dyads of a social network.

which the reliability is estimated for the different hierarchical levels in the data (Ten Hove et al., 2022a), or that correct for dependencies in two-level nested data (Yang & Zhou, 2014; Vanbelle, 2017; Ten Hove et al., 2022a). However, IRR coefficients for the more complex dependencies in interdependent social network data are currently unavailable. Applying existing IRR coefficients to interdependent social network data is a conflated approach, as they do not provide practically useful information on the IRR for all facets of theoretical interest. Also, we expect that point estimates or *SE*s may be biased because traditional IRR coefficients ignore the complex dependency structures in the data.

In this paper, we use Generalizability theory (GT; Cronbach et al., 1963; Shavelson et al., 1989) to develop IRR coefficients for interdependent social network data. First, we discuss the GT approach for estimating IRR of independent rater data and its limitations for interdependent social network data. Second, we discuss the SRM approach for variance decomposition of interdependent social network data and its implications for estimating IRR, and combine GT and the SRM to propose a rater-extended SRM (RESRM). The RESRM decomposes the variance in interdependent social network data into actor, partner, relationship, and rater-specific components, plus their interactions. Third, we propose RESRM-based IRR coefficients for each facet of theoretical interest (i.e., the actor, partner, and relationship components) in interdependent social network data separately, and for their integrated score. In addition, we propose an estimation procedure to obtain these IRR estimates from data. Fourth, we compare the conflated and RESRM methods for defining IRR using example data on social mimicry of Salazar Kämpf et al. (2018). Fifth, we inspect the bias and coverage rates of the proposed IRR coefficients in a simulation study. We end with a discussion of our findings and directions for future research.

7.1.1 Generalizability Theory

GT (Cronbach et al., 1963; Shavelson et al., 1989) is an extension of classical test theory (CTT) that can be used to estimate the IRR (e.g., Ten Hove et al., 2022c). CTT is used to estimate reliability of (composite) observed scores when observed scores consist of a single true-score component and a single error component. GT allows to estimate reliability for multiple facets of interest. Within GT, a single observation (e.g., the assessment of a person's degree of social mimicry) is considered to be sampled from a universe of admissible observations. The specific conditions under which an observation is made are called facets; in observational social network data typical facets are actors, partners, raters, and occasions. These facets can be divided into sources of theoretical interest (termed facets of differentiation; Vangeneugden et al., 2005), such as subjects, or sources of nuisance variability (termed facets of generalization; Vangeneugden et al., 2005), such as raters or measurement occasions. Reliability is then defined as the degree to which observations of the facets of differentiation can be generalized over the facets of generalization. Reliability is expressed with generalizability coefficients or indices of dependability. Within a single

study, multiple facets of generalization can be present. Hence, a single generalizability coefficient can account for multiple facets of generalization simultaneously (e.g., both multiple raters and multiple occasions), or generalizability coefficients can be defined for the separate facets of generalization. IRR refers to degree to which the observations of facets of interest can be generalized over raters.

7.1.2 Interrater Reliability for Independent Data

In independent observational data, independent subjects (S) are rated by independent raters (R). The universe of admissible observations consists of subjects and raters, where subjects are the facet of differentiation and raters are the facet of generalization. Considering a fully crossed (i.e., two-way) design, in which each subject is rated by each rater, multiple measurements Y_{SR} can be decomposed into a grand mean (μ), a mean for each of the facets (μ_S, μ_R), and the interaction terms between the aforementioned facets (μ_{SR}). The highest-order interaction term between the facets (here the two-way interaction μ_{SR}) is confounded with error:

$$Y_{SR} = \mu + \mu_S + \mu_R + \mu_{SR}. \tag{7.1}$$

A number of measurements Y_{SR} then involves the following orthogonal variances components (Shavelson et al., 1989): σ^2_S for the main subject component, σ^2_R for the main rater component, and σ^2_{SR} for the component representing the interaction between raters and subjects, which is confounded with any other source of error variance:

$$\sigma^2_Y = \sigma^2_S + \sigma^2_R + \sigma^2_{SR}. \tag{7.2}$$

The variance decomposition in Equation 7.2 can be used for several definitions of IRR, which are all intraclass correlation coefficients (ICCs; cq. *Generalizability coefficients* and *indices of dependability* in GT terminology; Ten Hove et al., 2022c). The IRR literature distinguishes between ICCs of interrater agreement and ICCs of interrater consistency, which can both be defined for single as well as averaged ratings (Shrout & Fleiss, 1979; McGraw & Wong, 1996). ICCs of interrater agreement express the degree to which the observed subject scores can be generalized over raters, and are of interest when subjects' scores are interpreted absolutely. For example, scores on educational tests that are used to decide whether students pass a test. Interrater agreement is thus useful when the absolute scores of subjects are used to make decisions about individual subjects; a practice we consider unlikely for the actor, partner, and relationship effects from the SRM. We therefore further ignore ICCs for interrater agreement. ICCs of interrater consistency express the degree to which the observed differences across subjects can be generalized over raters, and are of interest when subjects' scores are interpreted relatively to each other. For example, in correlational studies using the social relations model. Also, ICCs

have been defined for fixed as well as random raters. Ten Hove et al. (2022c) showed that the fixed-rater assumption is rarely, if ever, justified for IRR, so we focus only on random raters.

ICCs for interrater consistency are defined as the proportion of subject variance (i.e., the facet of interest; σ_S^2) over itself plus the variance in the subject-by-rater interaction effects, which are confounded with random error (σ_{SR}^2). The ICCs of interrater consistency do not include the relative standings of raters across ratings in the denominator (i.e., σ_R^2), because main-rater effects do not influence the observed differences across subjects when these subjects are all assessed by the same raters (cf. norm-referenced reliability; Winer, 2013).[2] Hence, the denominator only includes the variance components that are associated with rank ordering the facets of differentiation (here subjects). The rater-related variance component in the denominator of the ICC is divided by the number of raters over which subjects' scores are averaged (K), resulting in

$$\text{ICC}(C, K) = \frac{\sigma_S^2}{\sigma_S^2 + \frac{\sigma_{SR}^2}{K}}, \tag{7.3}$$

where C indicates consistency. For single ratings, $K = 1$ and disappears from the equation.

7.1.3 The Social Relations Model

The SRM models dyadic data as nested within both actors and partners, and actor and partner effects of individuals are allowed to correlate. The SRM can therefore be conceived as a cross-classified two-level model with a bivariate outcome variable that allows both positive and negative correlated actor and partner effects (Snijders & Kenny, 1999). The dyad-level observation Y_{ij} is partitioned into a grand mean M and three components, which are all deviations from this grand mean (similar to a GT decomposition; cf. Malloy & Kenny, 1986): A_i is the actor effect of person i, P_j is the partner effect of person j, and E_{ij} is the relationship effect when person i is the actor and person j is the partner; that is,

$$Y_{ij} = M + A_i + P_j + E_{ij}. \tag{7.4}$$

If multiple observations of Y_{ij} are available (e.g., multiple ratings by several raters) the relationship effect (E_{ij}) can be distinguished from error; if not, the relationship effect is confounded with random error.

Let $\mathbf{Y}_{\{ij\}}$ denote a vector containing a dyad's scores Y_{ij} and Y_{ji}. The SRM decomposes

[2] For incomplete observational design, in which the raters partly differ across subjects, a portion of the variance in main-rater effects should be added to the denominator of the ICCs of interrater consistency. This portion is based on the proportion of non-overlapping raters across subjects (Putka et al., 2008; Ten Hove et al., 2022c; Brennan, 2001a)

the dyadic scores $\mathbf{Y}_{\{ij\}}$, as

$$\mathbf{Y}_{\{ij\}} = \begin{bmatrix} Y_{ij} \\ Y_{ji} \end{bmatrix} = \begin{bmatrix} M \\ M \end{bmatrix} + \begin{bmatrix} A_i \\ A_j \end{bmatrix} + \begin{bmatrix} P_j \\ P_i \end{bmatrix} + \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix}. \tag{7.5}$$

The actor and partner effects in the SRM are assumed to be bivariate normally distributed with means of zero and variances σ_A^2 and σ_P^2. Moreover, it is assumed that both the actor and partner effects are mutually uncorrelated between individuals, but that a correlation, ρ_{AP}, exists between the actor effects A_i and partner effects P_i of each individual (named *generalized reciprocity*), resulting in

$$\begin{bmatrix} A_i \\ P_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_A^2 & \\ \rho_{AP}\sigma_A\sigma_P & \sigma_P^2 \end{bmatrix} \right). \tag{7.6}$$

The relationship effects are also assumed to be bivariate normally distributed with a mean of zero and variance σ_E^2. Also, a reciprocity or mutuality effect is expected between observations within a dyad. This reciprocity is modeled by the correlation between E_{ij} and E_{ji}, which is called the *dyadic reciprocity* (Kenny & La Voie, 1984, p. 157). It follows that

$$\begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_E^2 & \\ \rho_E\sigma_E^2 & \sigma_E^2 \end{bmatrix} \right). \tag{7.7}$$

The equality of the variance of E_{ij} and E_{ji} in Equation 7.7 could be relaxed if actors and partners have specific roles within a network (e.g., children as actors and teachers as partners). Without such roles to distinguish subjects i and j, it is reasonable to assume that $\sigma_{E_{ij}}^2 = \sigma_{E_{ji}}^2 = \sigma_E^2$.

Because all effects are mutually uncorrelated across actors, partners, and dyads (Snijders & Kenny, 1999, p. 474), the SRM decomposes the total variance of Y_{ij} into the following orthogonal variance components:

$$\sigma_Y^2 = \sigma_A^2 + \sigma_P^2 + \sigma_E^2 \tag{7.8}$$

7.1.4 A Rater Extended Social Relations Model

When external raters are used to measure dyadic variables, IRR coefficients should express the generalizability of actor, partner, and relationship effects across these external raters. Using GT, a measurement with actors, partners, and relationships as facets of interest, and raters as facets of nuisance, can be decomposed into effects for each of the facets, and their interactions.[3] The variance in each of these main- and interaction effects could then be used in generalizability coefficients. However, when decomposing the variation in the

[3]Note that the relationship effects are already interaction effects between actor and partner effects: actor × partner = relationship = E.

dyadic variables, the dependencies across subjects' actor and partner effects, and across the relationship effects within dyads should be accounted for. Therefore, we did not use a traditional GT-decomposition, but we used the GT rationale to extend the SRM with rater effects.[4]

The rater extended social relations model (RESRM) is a generalization of the SRM in Equation 7.6, which includes rater deviations around each SRM effect. The variability in M across raters is represented as a random intercept with grand-mean M, and a rater-specific deviation μ_K for each rater k. The variability in A_i across raters is represented as a random slope with mean A_i for each actor i, and a rater-specific deviation α_{ik} around A_i for each rater k. The variability in P_j across raters is represented as a random slope with mean P_j for each partner j, and a rater specific deviation π_{jk} around P_j for each rater k. Lastly, the variability in E_{ij} across raters is represented as a random slope with mean E_{ij} for each dyad ij, and a rater specific deviation ϵ_{ijk} around E_{ij} for each rater k. The resulting model is a rater-extended social relations model (RESRM):

$$Y_{ijk} = M + \mu_k + A_i + \alpha_{ik} + P_j + \pi_{jk} + E_{ij} + \epsilon_{ijk}. \tag{7.9}$$

Because the RESRM models repeated measures (i.e., multiple ratings) of the same dyadic interaction, E_{ij} can be disentangled from random error. In turn, ϵ_{ijk}, which represents the rater deviations from the relationship effect, is confounded with random error. Note that there are now K pairs of scores per dyad because each rater measures the bivariate outcome per dyad; that is,

$$\mathbf{Y}_{\{ij\}k} = \begin{bmatrix} Y_{ijk} \\ Y_{jik} \end{bmatrix} = \begin{bmatrix} M \\ M \end{bmatrix} + \begin{bmatrix} \mu_k \\ \mu_k \end{bmatrix} + \begin{bmatrix} A_i \\ A_j \end{bmatrix} + \begin{bmatrix} \alpha_{ik} \\ \alpha_{jk} \end{bmatrix} + \begin{bmatrix} P_j \\ P_i \end{bmatrix} + \begin{bmatrix} \pi_{jk} \\ \pi_{ik} \end{bmatrix} + \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix} + \begin{bmatrix} \epsilon_{ijk} \\ \epsilon_{jik} \end{bmatrix}. \tag{7.10}$$

The distributions of the actor, partner and relationship effects in the RESRM (equations 7.9 and 7.10) are provided by equations 7.6 and 7.7. The additional random intercepts for each rater are assumed to be normally distributed with a mean of zero and variance σ^2_μ; that is,

$$\mu_k \sim N(0, \sigma^2_\mu). \tag{7.11}$$

The rater deviations from the actor and partner effects are assumed to be bivariate normally distributed with means of zero and variances σ^2_α and σ^2_π, and to be mutually uncorrelated across raters. The same raters' deviations α_{ik} and π_{ik} from A_i and P_i may be correlated, $\rho_{\alpha\pi}$:

$$\begin{bmatrix} \alpha_{ik} \\ \pi_{ik} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_\alpha & \\ \rho_{\alpha\pi}\sigma_\alpha\sigma_\pi & \sigma^2_\pi \end{bmatrix} \right). \tag{7.12}$$

The rater deviations from the relationship effects, ϵ_{ijk} and ϵ_{jik}, are assumed to be bivariate

[4]Note that Kenny (1994) suggested the possibility of such an extension for items or occasions.

normally distributed with means of zero and common variance σ_ϵ^2. Also, we assume that a correlation, ρ_ϵ, exists between individual raters' deviations ϵ_{ijk} and ϵ_{jik} from E_{ijk} and E_{jik}, that is,

$$\begin{bmatrix} \epsilon_{ijk} \\ \epsilon_{jik} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\epsilon^2 & \\ \rho_\epsilon \sigma_\epsilon^2 & \sigma_\epsilon^2 \end{bmatrix} \right). \tag{7.13}$$

Because, as for the SRM, all effects are mutually uncorrelated across actors, partners, dyads (Snijders & Kenny, 1999, p. 474), and raters, the RESRM decomposes the total variance of Y_{ijk} into the following orthogonal variance components:

$$\sigma_Y^2 = \sigma_\mu^2 + \sigma_A^2 + \sigma_\alpha^2 + \sigma_P^2 + \sigma_\pi^2 + \sigma_E^2 + \sigma_\epsilon^2. \tag{7.14}$$

The RESRM thus provides the variance components that are associated with all potential facets of differentiation (i.e., actors, partners, and relationships) and the facet of generalization (i.e., raters), plus interaction effects between these facets, while taking the possible dependencies in the data into account.

7.1.5 Interrater Reliability for Interdependent Social Network Data

We used the variances in Equation 7.14 to define ICCs for interrater consistency of actor, partner, and relationship components, using the same rationale as in Equation 7.3. The numerator of each ICC of interrater consistency includes the variance components representing the facet of differentiation, that is, σ_A^2 for actor effects, σ_P^2 for partner effects, and σ_E^2 for relationship effects. The denominator includes the variance components of the facet of differentiation plus the variance of the interaction effect between this facet of differentiation and the raters. That is, $\sigma_A^2 + \sigma_\alpha^2$ for actor effects as facet of differentiation, $\sigma_P^2 + \sigma_\pi^2$ for partner effects as facet of differentiation, and $\sigma_E^2 + \sigma_\epsilon^2$ for relationship effects as facet of differentiation. For example, an ICC of interrater consistency of the actor effects is

$$\mathrm{ICC}_A(C, K) = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_\alpha^2}{K}}. \tag{7.15}$$

The ICC of interrater consistency for the integrated score (i.e., the combination of actor, partner, and relationship effects) is defined as a fraction with the sum of the variance of the facets of differentiation (i.e., $\sigma_A^2 + \sigma_P^2 + \sigma_E^2$) as the numerator, and this sum plus the variances in the interaction effects between the facets of differentiation and the raters (i.e., $\sigma_\alpha^2 + \sigma_\pi^2 + \sigma_\epsilon^2$) in the denominator. Table 7.2 provides the ICCs for all possible facets of differentiation in the SRM.

Table 7.2: Interrater Reliability Coefficients for Interdependent Social Network Data

Facet of Interest	Single rating $(C, 1)$	Average ratings (C, K)
Actor effect (ICC_A)	$\frac{\sigma_A^2}{\sigma_A^2+\sigma_\alpha^2}$	$\frac{\sigma_A^2}{\sigma_A^2+\frac{\sigma_\alpha^2}{K}}$
Partner effect (ICC_P)	$\frac{\sigma_P^2}{\sigma_P^2+\sigma_\pi^2}$	$\frac{\sigma_P^2}{\sigma_P^2+\frac{\sigma_\pi^2}{K}}$
Relationship effect (ICC_E)	$\frac{\sigma_E^2}{\sigma_E^2+\sigma_\epsilon^2}$	$\frac{\sigma_E^2}{\sigma_E^2+\frac{\sigma_\epsilon^2}{K}}$
Integrated score (ICC_Y)	$\frac{\sigma_A^2+\sigma_P^2+\sigma_E^2}{\sigma_A^2+\sigma_P^2+\sigma_E^2+\sigma_\alpha^2+\sigma_\pi^2+\sigma_\epsilon^2}$	$\frac{\sigma_A^2+\sigma_P^2+\sigma_E^2}{\sigma_A^2+\sigma_P^2+\sigma_E^2+\frac{\sigma_\alpha^2+\sigma_\pi^2+\sigma_\epsilon^2}{K}}$

7.1.6 Estimating ICCs with the RESRM

ICCs as well as the SRM are traditionally estimated using random-effects ANOVA models. In the SRM, correlations are then allowed between an individual's actor and partner effects and within dyads (Warner et al., 1979). Others proposed to estimate the SRM as a multilevel model with cross-classified random effects, using maximum likelihood estimation (Snijders & Kenny, 1999). This procedure is available in some software (e.g., MLwiN; Rasbash et al., 2017), but many multilevel modelling software packages do not allow correlations between cross classified random effects. For a more detailed description of this multilevel approach, we refer to Snijders and Kenny (1999). More recently, a structural equation modelling approach (e.g., Nestler et al., 2020) and a fully Bayesian approach (Hoff, 2005; Lüdtke et al., 2013) were proposed to estimate the SRM. Both estimation methods have also been proposed to estimate IRR and generalizability coefficients (e.g., Vispoel et al., 2018a, 2019; Jorgensen, 2021; Ten Hove et al., 2020, 2022a, Under Reviewa). We will focus on the Bayesian approach because under conditions that are expected to occur frequently in observational social-network research, this approached outperformed the IRR estimates obtained using the structural equation modelling approach (Ten Hove et al., Under Reviewa). These conditions involve small samples (e.g., of raters) and variance components close to the boundary of zero.

To estimate the parameters of the RESRM, we implemented the Bayesian approach as proposed by Hoff (2005) and more recently described by Lüdtke et al. (2013). Let ϑ be the vector of all model parameters. The posterior distribution $P(\vartheta|Y)$ of the model parameters given the data is estimated as proportional to the product of the prior probability distribution $P(\boldsymbol{\vartheta})$ and the likelihood of the data, conditional on the parameters $P(Y|\boldsymbol{\vartheta})$. By using non-informative priors for the estimation of $P(\boldsymbol{\vartheta}|Y)$ by means of Markov chain Monte Carlo (MCMC) estimations, the estimated posterior distribution is overwhelmingly influenced by the observed data. The model estimates are then similar to estimates that would be obtained with MLE.

MCMC algorithms enable to estimate all unknown RESRM quantities (i.e., parameters and random effects) simultaneously. Let K be the number of raters, let N be the number

of subjects (and thus the number of actors and partners), and let $D = N\,(N-1)\,/2$ be the maximum number of dyads given N.[5] Assuming a fully crossed design in which all raters observe all dyadic interactions, and all subjects interact with all other subjects, $\boldsymbol{\vartheta}$ of the RESRM includes the following $12 + 2N + 2D + K + 2NK$ quantities: The grand mean M (Equation 7.9); variance components $\sigma_\mu^2, \sigma_A^2, \sigma_\alpha^2, \sigma_P^2, \sigma_\pi^2\ \sigma_E^2$, and σ_ϵ^2 (or their square roots;[6] Equation 7.14); correlations ρ_{AP} (Equation 7.6), ρ_E (Equation 7.7), $\rho_{\alpha\pi}$ (Equation 7.12), and ρ_ϵ (Equation 7.13); subject-level random effects $[A_i P_i]$ for $i = 1, \ldots,\ N$, dyad-level random effects $[E_{ij}\ E_{ji}]$ for $i = 1, \ldots, N; j = 1, \ldots N; i \neq j$, rater-specific deviations μ_k for $k = 1, \ldots,\ K$, and rater-specific deviations from subject-level random effects $[\alpha_{ik}\pi_{ik}]$ for $i = 1, \ldots,\ N; k = 1, \ldots, K$ (Equation 7.10). If the design is not fully crossed, each rater observes a subset of dyads rather than all dyads, or subjects do not interact with all other subjects. For such a design, the model involves the same parameters but fewer random effects to estimate.

The conditional distribution of $\mathbf{Y}_{\{ij\}k}$ given $\boldsymbol{\vartheta}$ is the bivariate normal likelihood of the observed data $P(\mathbf{Y}|\boldsymbol{\vartheta})$, that is:

$$\begin{bmatrix} Y_{ijk} \\ Y_{jik} \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{Y}_{ijk} \\ \hat{Y}_{jik} \end{bmatrix}, \begin{bmatrix} \sigma_\epsilon^2 & \\ \rho_\epsilon \sigma_\epsilon^2 & \sigma_\epsilon^2 \end{bmatrix} \right). \tag{7.16}$$

This bivariate normal likelihood is specified using the vector of expected values for $\mathbf{Y}_{\{ij\}k}$

$$\hat{\mathbf{Y}}_{\{ij\}k} = \begin{bmatrix} \hat{Y}_{ijk} \\ \hat{Y}_{jik} \end{bmatrix} = \begin{bmatrix} M \\ M \end{bmatrix} + \begin{bmatrix} \mu_k \\ \mu_k \end{bmatrix} + \begin{bmatrix} A_i \\ A_j \end{bmatrix} + \begin{bmatrix} \alpha_{ik} \\ \alpha_{jk} \end{bmatrix} + \begin{bmatrix} P_j \\ P_i \end{bmatrix} + \begin{bmatrix} \pi_{jk} \\ \pi_{ik} \end{bmatrix} + \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix}. \tag{7.17}$$

The prior distributions of the random effects are defined by equations 7.6, 7.7, 7.11, 7.12, and 7.13, whose parameters (variance components, and correlations) are then hyperparameters which require their own prior distributions. These hyperprior distributions can be uninformative (or weakly informative) to allow the posterior to be influenced only by the data. More informative distributions can be specified if researchers want to incorporate stronger prior beliefs or results of previous research.

7.2 Empirical Example: Social Mimicry

We estimated the IRR of interdependent social network data on social mimicry (Salazar Kämpf et al., 2018) using both a conflated approach and the RESRM approach. The data were collected using a round-robin design (i.e., each subject interacted with all other subjects within a network; Warner et al., 1979), and raters viewed the recorded

[5] $D = N\,(N-1)\,/2$ only when all subjects interact with all other subjects. When subjects interact in small subgroups, $D < N\,(N-1)\,/2$.

[6] Depending on the software, these hyperparameters are estimated in terms of random-effect *SD*s, or random-effect variances.

interactions to rate the social mimicry of both subjects in each dyadic interaction. Salazar Kämpf et al. (2018) made this data publicly available on the Open Science Framework: https://osf.io/b4nvf/

7.2.1 Method

Participants

In total, $N = 139$ German students participated in the study of Salazar Kämpf et al. (2018). Each of these students was randomly assigned to a group of four to six members, forming 26 same-sex networks. Each student had an interaction with all other subgroup members, which resulted in a total of $D = 309$ dyadic interactions. For a more detailed description of the sampling procedure and the sample, we refer to Salazar Kämpf et al. (2018).

Measures

Each 5-minute dyadic interaction was videotaped and $K = 3$ raters rated the degree of social mimicry of each participant during each interaction, using a 6-point Likert scale ranging from 1 (*not at all*) to 6 (*very much*). Salazar Kämpf et al. (2018) calculated a conflated IRR estimate of ICC$(2, 3) = .87$.[7] A more detailed description of the measures and the rating protocol can be retrieved from the supplementary materials of Salazar Kämpf et al. (2018).

Analysis Plan

Software for Estimation. Whereas Lüdtke et al. (2013) used Gibbs sampling techniques to estimate the variance components associated with each facet, we used a No-U-Turn Sampler (NUTS), a special case of Hamiltonian Monte Carlo (HMC) that is implemented in the Stan software (Carpenter et al., 2017) and available in the R (R Core Team, 2021) package `rstan` (Stan Development Team, 2020). NUTS, and HMC in general, is faster and more efficient than Gibbs sampling, especially when models are highly parameterized or include highly correlated parameters (Monnahan et al., 2017). Whereas Gibbs sampling techniques sample elements of $\boldsymbol{\vartheta}$ (i.e., the mean, the random-effect variances *or SD*s, the random-effect correlations, and all random effects) sequentially, NUTS simultaneously samples the entire vector $\boldsymbol{\vartheta}$ by simulating it as a point in a N_{par}-dimensional space, where, in our case, $N_{par} = K + 2N + 2D + 2KN + 12 = 3 + 2 \times 139 + 2 \times 309 + 2 \times 3 \times 139 + 12 = 1745$.

The STAN program estimates the random-effect hyperparameters in terms of *SD*s instead of variances. We derived posterior distributions of all ICCs in Equation 7.2, as functions of the posterior distributions of these estimated *SD*s. A Bayesian credible

[7] The ICC$(2, 3)$ as defined by Shrout and Fleiss (1979) is identical to the ICC(C, K) in Equation 7.3

interval (BCI) provided an estimate of the precision with which an ICC was estimated. We used the modal a posteriori (MAP) estimates as point estimates of the ICCs, and we obtained 95% BCIs using percentiles (Lüdtke et al., 2013; Ten Hove et al., 2020). For comparison, we also estimated the conflated ICCs with the R software package `irr` (Gamer et al., 2012). We provide all software code that we used for this article on the Open Science Framework: https://osf.io/jd3xb/.

Model Diagnostics. We initially used three independent chains of 1,000 iterations to estimate the model: The first 500 iterations of each chain served as burn-in iterations, whereafter we saved 500 samples from the posterior in each chain. We used traceplots to check whether the three independent chains converged on the same posterior distribution, and inspected the potential scale reduction factor ($\hat{R}$) and effective sample size (Neff), using $\hat{R} < 1.10$ and Neff > 100 as indication for adequate mixing of the independent chains and a sufficient effective sample size (Gelman & Rubin, 1992). We had to double the number of post-burn in iterations because of insufficient effective sample sizes, which resulted in a in a sample of 3 (chains) $\times$ 1,000 (post burn-in iterations) = 3,000 iterations to obtain the MAPs and BCIs of the ICCs, for which the traceplots showed adequate mixing.

Prior Distributions. We specified weakly informative prior distributions for each parameter, assuming that the standard deviations ($\sigma_A, \sigma_P, \sigma_E, \sigma_\mu, \sigma_\alpha, \sigma_\pi$, and σ_ϵ) followed a half-$t(4, 0, 1)$ distribution (Ten Hove et al., 2020), with a range of (0, 3), which is half the range of Y and therefore the largest a SD could possibly be. Moreover, we assumed that the correlations between the effects within dyads or persons (ρ_{AP}, ρ_E, $\rho_{\alpha\pi}$, and ρ_ϵ) were uniformly distributed across the range of $(-1, 1)$.

7.2.2 Results

Variance Decomposition

Table 7.3 shows all SRM parameters as estimated with maximum likelihood by Salazar Kämpf et al. (2018), and all RESRM parameters that we estimated with `stan`, including the model diagnostics. The estimated grand mean, which is the average degree of social mimicry across subjects, showed comparable estimates in the SRM of Salazar Kämpf et al. (2018) and the RESRM, as did the estimated proportions of variance that were explained by the actor, partner, and relationship components of social mimicry. The difference is in the error components. Salazar Kämpf et al. (2018) only estimated a single error component, whereas we separated this composite into four different rater-related error components: a variance component for the differences in relative standings of raters (σ^2_μ), and variance components for the rater deviations from the actor effects (σ^2_α), partner effects (σ^2_π), and relationship effects (σ^2_ϵ). The variance component attributed to rater deviations from the relationship effects, which

is confounded with measurement error, was the largest of the four rater-error variance components (24%). Substantial parts of the total variance could also be attributed to the rater deviations from the actor effects (8%), whereas only negligible portions of the total variance were explained by the rater deviations from the grand mean (< 1%) and the partner effects (< 1%) of social mimicry. The proportion of variance that was explained by the combined rater deviations from the mean, and actor, partner and relationship effect (33%), was comparable to the undifferentiated error variance of the SRM (35%).

Table 7.3: Estimated Variance Components using the SRM (by Salazar Kämpf et al. (2018) and using the RESRM.

	SRM		RESRM					
Variance	Est	% variance	MAP	2.5%	97.5%	% variance	N_{eff}	$\hat{R}$
σ^2_A	0.32	24	0.37	0.26	0.54	27	866	1.00
σ^2_P	0.07	6	0.09	0.04	0.17	7	182	1.02
σ^2_E	0.47	35	0.46	0.37	0.58	34	497	1.00
σ^2_μ	–	–	0.00	0.00	0.39	< 1	270	1.00
σ^2_α	–	–	0.11	0.06	0.15	08	161	1.01
σ^2_π	–	–	0.00	0.00	0.05	< 1	105	1.02
σ^2_ϵ	–	–	0.33	0.30	0.36	24	903	1.00
σ^2_{Error}	0.46	35	–	–	–	–	–	–
σ^2_{Total}	1.32	100	1.39	1.27	1.81	100	427	1.00

Interrater Reliability

Table 7.4 shows all IRR estimates as estimated with a conflated approach and the RESRM approach. These results show that the RESRM estimates of the ICCs for single and averaged ratings of the integrated score (i.e., the combination of all three SRM components; ICC_Y) was comparable to the conflated IRR point estimates. However, all RESRM-based IRR point estimates were outside the 95% confidence interval of the conflated IRR estimates.

The conflated IRR estimates seem to underestimate the IRR of the actor and partner components of social mimicry and overestimate the IRR of the relationship component of social mimicry. Overall, these results imply that it is not safe to assume that the reliability associated with the integrated scores (i.e., conflated IRR estimates) adequately represents the reliability associated with each component of the data.

Table 7.4: RESRM-based ICC Estimates.

	Single Ratings $(C, 1)$			Averaged Ratings (C, K)		
ICC	Est.	2.5%	97.5%	Est.	2.5%	97.5%
ICC_{Conf}	0.68	0.65	0.72	0.87	0.85	0.88
ICC_{Y}	0.68	0.63	0.72	0.86	0.84	0.89
ICC_{A}	0.79	0.68	0.87	0.92	0.86	0.95
ICC_{P}	0.98	0.65	1.00	0.99	0.79	1.00
ICC_{E}	0.59	0.52	0.64	0.81	0.77	0.84

Note. Y = Integrated scores; A = Actor effects; P = Partner effects R = Relationship effects. The ICC as reported by Salazar Kämpf et al. (2018, i.e., ICC(2,3) = .87), resembled the conflated ICC (i.e., ICC_{Conf}) for averaged ratings as estimated with the `irr` package.

7.3 Simulation Study

7.3.1 Methods

Data Generation

We conducted a simulation study to gain a first impression of the bias and coverage of the RESRM-based ICCs under favorable and less favorable conditions. We varied the research design and the population parameters of the RESRM. In each condition, we used `mvrnorm` function in the R-package `rockchalk` (Johnson, 2016) to generate bivariate normally distributed data from Equation 7.10 using the parameters in equations 7.6, 7.7, 7.11, 7.12, 7.13, and 7.14.

Independent Variables

The *design* had two levels: A *good* design with substantial and balanced sample sizes, and an *poor* design based on the empirical example.[8] The good design with substantial sample sizes resembled a situation in which a group of 10 subjects each interacted with all other subjects in the group, yielding $\frac{10\times9}{2} = 45$ dyadic interactions. All interactions were rated twice (once to rate subject i's attribute, and once to rate subject j's attribute) by 10 raters, resulting in 45 (dyadic interactions) $\times$ 10 (raters) $\times$ 2 (ratings per interaction) = 900 dyadic observations. The poor design was based on the empirical example and exactly resembled the design of Salazar Kämpf et al. (2018), in which groups of three to six subjects were each rated by three raters, yielding 309 dyadic interactions, and 309 (dyadic observations) $\times$ 3 (raters) $\times$ 2 (ratings per interaction) = 2154 dyadic observations.

The *parameters* also had two levels: A (co)variance structure with *substantial* RESRM parameters, and a (co)variance structure with *varying* parameters. The substantial

[8] We selected the terms *good* and *poor* to ease the discussion of the simulation results. There may be better or worse conditions than those that we selected. Also, good conditions to estimate IRR coefficients may differ from good conditions for drawing inferences about individuals' attributes in SRM analyses.

population parameters we selected were: $\sigma_A = \sigma_P = \sigma_R = \sigma_\pi = \sigma_\alpha = \sigma_\epsilon = 1$ and $\rho_{AP} = \rho_E = \rho_{\alpha\pi} = \rho_\epsilon = .3$. The varying population parameters were based on the empirical example: $\sigma_A = .60, \sigma_P = .30, \sigma_E = .70, \sigma_\alpha = .30, \sigma_\pi = .10, \sigma_\epsilon = .60, \rho_{AP} = .70, \rho_E = .70, \rho_{\alpha\pi} = -.30, \rho_\epsilon = .20$. These population parameters were specified as *SD*s and correlations instead of (co)variances, because the STAN model estimates *SD*s rather than variances, from which the variances and ICCs are derived.

This simulation design yielded 2 (design) × 2 (parameters) = 4 conditions in total, for each of which we generated 1,000 datasets. The resulting population ICCs ranged from .50 to .90 for single ratings, and from .75 to .99 for averaged ratings (Table 7.5).

Table 7.5: Population ICCs for Averaged Ratings of the Integrated Score, and of the Actor-, Partner-, and Relationship Effects across Simulation Conditions

			ICC			
Parameters	Design	Ratings	*Y*	*A*	*P*	*E*
Substantial	Good	Single	.50	.50	.50	.50
		Averaged	.91	.91	.91	.91
	Poor	Single	.50	.50	.50	.50
		Averaged	.75	.75	.75	.75
Varying	Good	Single	.67	.80	.90	.58
		Averaged	.95	.98	.99	.93
	Poor	Single	.67	.80	.90	.58
		Averaged	.86	.92	.96	.80

Note. Y = Integrated scores; A = Actor effects; P = Partner effects R = Relationship effects.

Estimation

We added an automated convergence check to the estimation procedure that we used to estimated the IRR for the empirical example. If the three independent chains did not mix well according to the $\hat{R}$ criterion of $\hat{R} < 1.10$, we doubled the number of post burn-in iterations. This was repeated until the model converged, or did not converge after the limit of 8,000 post burn-in iterations was reached, in which case we discarded the replication.

Dependent Variables

Bias of Point Estimates. Let $\bar{\theta}$ denote the average ICC as estimated 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 between .05 and .10 as minor bias and relative bias > .10 as substantial bias.

BCI Coverage Rates. We computed the coverage rates as the percentage of converged replications in a condition for which the 95% BCI contained the population ICC. Agresti-

Coull intervals indicate that with 1000 replications, 95% BCI coverage < .93 or > .96 differ significantly from .95 (Agresti & Coull, 1998). We considered only BCI-coverage rates < .90 practically too low.

7.3.2 Results

The model converged for almost all replications, varying from 98% (good-design, varying-parameters condition) to 100% (good design, substantial-parameters condition). Averaged across conditions, the RESRM provides unbiased estimates with good coverage rates for most ICCs (tables 7.6 and 7.7). However, the ICCs for single ratings of the actor effects were overestimated, and their BCIs were too narrow, as was the case for these ICCs for averaged ratings. Also, the ICCs for single ratings of the relationship effects were slightly underestimated, and the coverage rates for the ICCs of relationship effects were too low for both single and averaged ratings.

Table 7.6: Relative Bias across Simulation Conditions

		Substantial Parameters		Varying Parameters	
ICC	$M(SD)$	Good Design	Poor Design	Good Design	Poor Design
$ICC_T(C,1)$	0.00 (0.01)	-0.00	-0.01	0.00	0.00
$ICC_A(C,1)$	0.11 (0.15)	-0.02	0.31	0.04	0.09
$ICC_P(C,1)$	-0.01 (0.08)	-0.10	0.09	-0.06	0.02
$ICC_E(C,1)$	-0.07 (0.12)	-0.00	-0.26	-0.00	-0.03
$ICC_T(C,k)$	0.00 (0.00)	0.00	-0.01	0.00	0.00
$ICC_A(C,k)$	0.05 (0.06)	0.01	0.14	0.00	0.03
$ICC_P(C,k)$	0.00 (0.04)	-0.01	0.04	-0.05	0.01
$ICC_E(C,k)$	-0.04 (0.07)	0.00	-0.15	-0.00	-0.01

Table 7.7: 95% BCI Coverage Rates across Simulation Conditions

		Substantial Parameters		Varying Parameters	
ICC	$M(SD)$	Good Design	Poor Design	Good Design	Poor Design
$ICC_T(C,1)$	.95 (0.02)	.95	.92*	.96*	.95
$ICC_A(C,1)$	.74* (0.25)	.94	.45*	.94	.61*
$ICC_P(C,1)$	.94 (0.04)	.95	.90*	.94	.99*
$ICC_E(C,1)$	.72* (0.44)	.96	.06*	.95	.91*
$ICC_T(C,k)$	.95 (0.02)	.95	.92*	.96*	.95
$ICC_A(C,k)$	.74* (0.25)	.94	.45*	.94	.61*
$ICC_P(C,k)$	.94 (0.04)	.95	.90*	.94	.99*
$ICC_E(C,k)$	.72* (0.44)	.96	.06*	.95	.91*

Note. * = Coverage rate outside Agresti-Coull interval.

Bias

Figure 7.1 shows the relative bias of the ICCs across conditions. Most ICCs were accurately estimated, especially in the good-design conditions. In both good-design conditions, only the ICCs for single ratings of the partner effects were slightly underestimated. In the poor-design, substantial-parameters condition, the ICCs of the actor effects were severely overestimated, and the ICCs of the relationship effects were severely underestimated. In the poor-design, varying-parameters condition, the ICCs for single ratings of the actor effects were slightly overestimated.

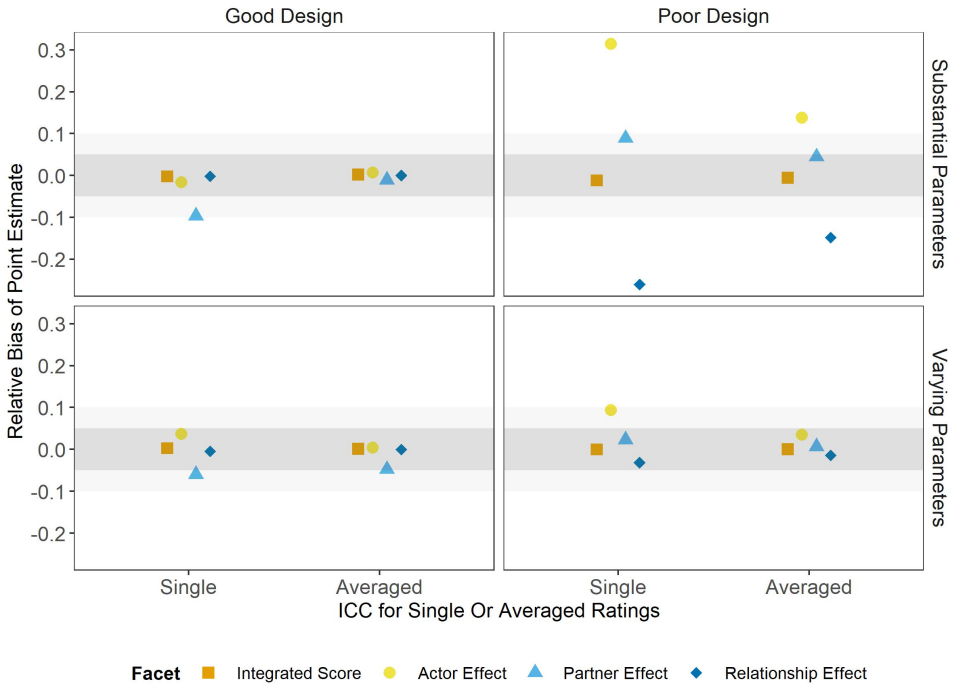

Figure 7.1: Relative bias of the ICCs across conditions. White areas: substantial bias (>10%); Light-gray areas: minor bias (5–10%); Dark-gray areas: negligible bias (< 5%).

Coverage

Figure 7.2 shows the 95% BCI coverage rates of the ICCs across conditions. Most ICCs had near-nominal coverage rates, especially in the good-design conditions. In the poor-design conditions, and particularly in the poor-design, substantial-parameters condition, coverage rates of the ICCs of actor and relationship effects were too low.

7.4 Discussion

In this study, we proposed, illustrated, and tested a rater extended social relations model to estimate the IRR of interdependent social network data. Using an empirical example on

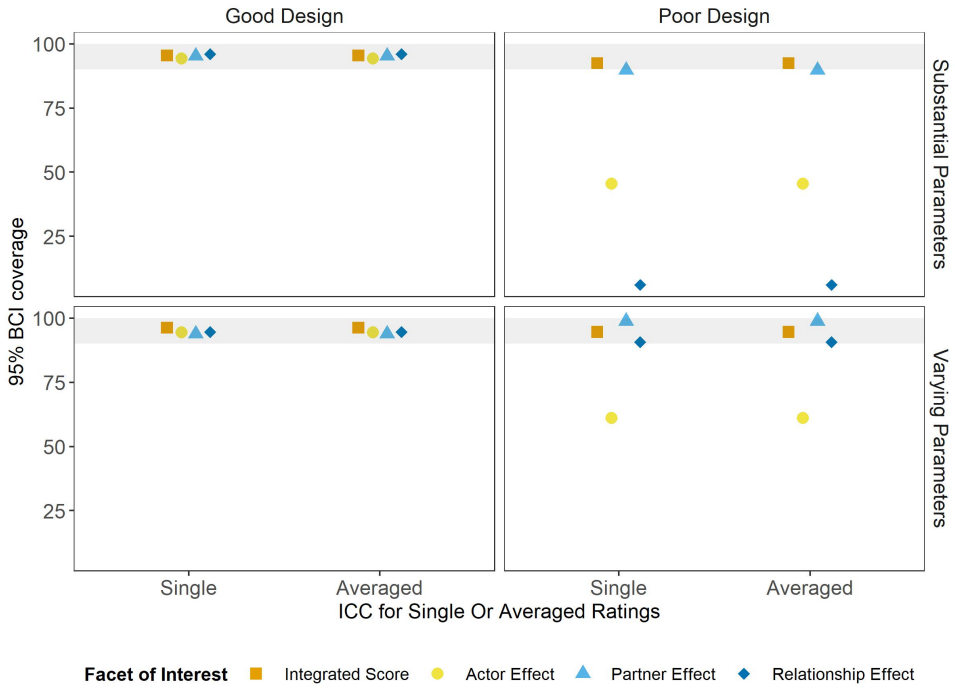

Figure 7.2: Coverage of the ICCs across conditions. White areas: Practically too low coverage rates $< 90\%$

.

social mimicry, we showed that the different components of interdependent social network data may have their own IRR and that it is thus unsafe to assume that a conflated estimate adequately represents the IRR associated with each of these facets of interest. We tested the properties of the proposed estimator in a simulation study, which indicated that the proposed ICCs were mostly unbiased and generally had good coverage rates.

Unfortunately, the method did not yield unbiased ICCs with nominal coverage rates for all conditions. It appeared that in conditions with both small subgroups of interacting subjects and a few number of raters, the RESRM could not accurately estimate all ICCs. However, the simulation conditions provide a first impression of the performance estimator under favorable and unfavorable conditions. Follow-up research is needed to test the properties of the proposed estimation method in more conditions, to disentangle the effects of various design factors, such as the number of raters and the type of social network design, on the bias and coverage of the ICCs. Because researchers often use Likert-type scales, or dichotomous variables in network studies (i.e., mainly to indicate whether a relation exists), other useful follow-up research includes developing an RESRM that handles discrete data.

The good design, for which the method performed well, represents conditions with many raters. Such conditions may seem non-pragmatic, because using many ratings is typically time consuming and expensive. However, although the good-design conditions had more raters per subject ($K = 10$) than than the poor-design conditions ($K = 3$),

the good-design conditions were more efficient as they required 900 ratings whereas the poor-design conditions required 2154 ratings. The good design conditions used fewer participants than may be desirable for an SRM study, but such a design is useful to inspect the quality of the rating procedure. If researchers want to be able to accurately estimate the ICCs, but also want sufficient information about subjects and dyads to have sufficient power for their SRM analyses, we suggest to use a validation study to estimate the IRR. In such a validation study, the quality of the rating procedure can be investigated using a subsample of all subjects, with many raters per subject. Using the variance components estimated with the RESRM based on this subsample, IRR coefficients could be defined for each desired design, thus also for a design with more subjects but single ratings. If the IRR for single ratings is sufficient, the remaining subjects in an SRM study could then be observed by a single rater, and the SRM could be fitted to these single ratings.

We believe that the RESRM approach is a promising conceptual and analytical tool for evaluating the IRR of dyad-level predictors in social relations research. In SRM research, observed dyadic variables are decomposed into their actor, partner, and relationship components, each of which might be of interest as predictors or outcomes in a statistical model. For example, Salazar Kämpf et al. (2018) modeled the effects of (and on) each SRM component of social mimicry separately. The conflated IRR estimates appeared to underestimate the IRR of the actor and partner effects, and to overestimate the IRR of relationship effects. Conflated or integrated ICCs are not useful in SRM research, because such ICCs do not consider the IRR for each SRM component separately. The RESRM provides a basis for estimating the IRR for all SRM components separately, hence yields more informative IRR estimates than a conflated GT approach. However, in social network research, the integrated (i.e., observed) score may be used as a predictor of network ties, for example in the p_2 or j_2 models (Van Duijn et al., 2004; Zijlstra, 2017). In such situations, the ICCs for an integrated score would be useful.