




Interrater Reliability for Interdependent Social Network Data: A Generalizability Theory Approach

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

^aDepartment of Educational and Family Studies, Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; ^bResearch Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands

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 variance of dyadic observational data into effects of actors, partners, relationships, raters, and their statistical interactions. We used the variances of these effects to define intraclass correlation coefficients (ICCs) that indicate the extent the actor, partner, and relationship effects can be generalized across external 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

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 heavily on the raters. Low IRR can result in biased estimates or loss of power in statistical analyses of subjects' attributes (cf. Lord & Novick, 1968, p. 72),

potentially leading to faulty conclusions. Thus, studying the IRR is vital for social and behavioral research. When the IRR is too low, researchers should improve the IRR, for example, by improving rating procedures. In this paper, we show that for interdependent social network data (to be explained shortly), the current measures of IRR are suboptimal, so researchers cannot assess the IRR. We propose and test more elaborated and improved IRR measures for this type of data, and compare our proposed IRR measures to existing (interrater) reliability measures for interdependent social network data (Bonito & Kenny, 2010; Kenny et al., 1994; Malloy & Kenny, 1986).

People often behave differently (e.g., smiling, bullying, mimicking) depending on whom they interact with, and those individuals may, in turn, 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 interdependent behaviors are known as dyadic variables (Kenny

Table 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 $i, j \in A, B, C, D$ and $i \neq j$), rated by rater k ($k = 1, 2, 3$).

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}	–

et al., 2006). Dyads consist of two subjects: An actor displaying the behavior and a partner reacting on the behavior. 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 (e.g., Table 1, top panel). Consequently, all dyadic observations including actor i are nested within actor i , and all dyadic observations including partner j are nested within partner j .

In observational social-network research, raters rate subjects' attributes during interactions with several other subjects (e.g., Huang et al., 2017; Hughes et al., 2021; Salazar Kämpf et al., 2018), and the raters are not part of these interactions. Table 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., Anna and Brooke, Anna and Charlie, Anna and Deborah, etc.). The behavior of the students during their interactions with the other students is rated twice—as an actor and as a partner—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, 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, Anna's actor effect would reflect how often Anna mimics her conversation partners on average (i.e., how imitative is Anna?). Brooke's partner effect would reflect how often she is mimicked by her conversation partners on average (i.e., how imitable is Brooke?). Given how often Anna generally mimics her conversation partners (her actor effect), and how often Brooke is generally mimicked (her partner effect), the relationship effect would reflect 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 mimics others frequently may also elicit more mimicry on average. 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 (i.e., estimated actor, partner and relationship effects, and their variances), may differ across external raters, hence fail to generalize. Therefore, the IRR of the actor, partner, and relationship effects should be investigated to inspect the degree to which these effects depend on external raters. A low IRR may inform researchers that raters require more training in using the observation instrument, or that more raters should assess each dyadic interaction. IRR methods have been proposed that can handle multilevel data, in which the reliability is estimated for the different hierarchical levels in the data (Ten Hove et al., 2022), or that correct for dependencies in two-level nested data (Ten Hove et al., 2022; Vanbelle, 2017; Yang & Zhou, 2014). 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

¹There 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.

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 SEs may be biased because traditional IRR coefficients ignore the complex dependency structures in the data.

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 the 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 the 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 statistical 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.

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., 2024c). 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; e.g., Vangeneugden et al., 2005), such as subjects, or sources of nuisance variability (termed facets of generalization; e.g., 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.

Interrater reliability for independent data

In independent observational data, independent subjects (i)² are rated by independent raters (k). 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 observations Y_{ik} can be decomposed into a grand mean (M), a mean for each of the facets (S_i , R_k), and the statistical interaction terms between the aforementioned facets (SR_{ik}). The highest-order statistical interaction term between the facets (here the two-way interaction SR_{ik}) is confounded with error:

$$Y_{ik} = M + S_i + R_k + SR_{ik}. \quad (1)$$

Observations Y_{ik} then involve the following orthogonal variances components (Shavelson et al., 1989): σ_S^2 for the main subject variance component, σ_R^2 for the main rater variance component, and σ_{SR}^2 for the variance component representing the statistical interactions between raters and subjects, which is confounded with any other source of error variance:

$$\sigma_Y^2 = \sigma_S^2 + \sigma_R^2 + \sigma_{SR}^2. \quad (2)$$

The variance decomposition in Equation 2 can be applied to several definitions of the intraclass correlation coefficient (ICC), all of which correspond to different definitions of the IRR. These ICCs are identical to *Generalizability coefficients* and *indices of dependability* in GT (Ten Hove et al., 2024c). 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 (McGraw & Wong, 1996; Shrout & Fleiss, 1979). ICCs of

²In this subsection, we do not distinguish between actor and partner effects of subjects. We therefore use the subscript i to indicate subjects, rather than actors.

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, this applies to 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, this applies to correlational studies using the social relations model. Also, ICCs have been defined for fixed as well as random raters. When raters in an observation study are considered fixed, they are viewed as the entire population of possible raters. When raters are considered random, they are viewed as a random sample of potential raters that could have been trained using the rating protocol at hand. Typically, raters are considered random. Ten Hove et al. (2024c) argued that the fixed-rater assumption is rarely, if ever, justified for IRR, so here 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 the subject variance 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 all subjects are assessed by the same raters (cf. norm-referenced reliability; Winer, 2013).³ 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

$$\text{ICC}(C, K) = \frac{\sigma_S^2}{\sigma_S^2 + \frac{\sigma_{SR}^2}{K}}, \quad (3)$$

where C indicates consistency. For single ratings, $K = 1$ and $\frac{\sigma_{SR}^2}{K}$ reduces to σ_{SR}^2 .

³For 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 (Brennan, 2001; Putka et al., 2008; Ten Hove et al., 2024c).

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,

$$Y_{ij} = M + A_i + P_j + E_{ij}. \quad (4)$$

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

$$\mathbf{Y}_{\{ij\}} = \begin{bmatrix} Y_{ij} \\ Y_{ji} \end{bmatrix} = \begin{bmatrix} M \\ M \end{bmatrix} + \begin{bmatrix} A_i \\ A_j \end{bmatrix} + \begin{bmatrix} P_j \\ P_i \end{bmatrix} + \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix}. \quad (5)$$

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; however, a within-person correlation ρ_{AP} (named *generalized reciprocity*) exists between the actor effect A_i and partner effect P_i of the same individual, resulting in a bivariate distributional assumption:

$$\begin{bmatrix} A_i \\ P_i \end{bmatrix} \sim N\left(\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_{AP} = \begin{bmatrix} \sigma_A^2 & \rho_{AP}\sigma_A\sigma_P \\ \rho_{AP}\sigma_A\sigma_P & \sigma_P^2 \end{bmatrix}\right). \quad (6)$$

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

$$\begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix} \sim N\left(\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_E = \begin{bmatrix} \sigma_E^2 & \rho_E\sigma_E^2 \\ \rho_E\sigma_E^2 & \sigma_E^2 \end{bmatrix}\right). \quad (7)$$

The equality of the variance of E_{ij} and E_{ji} in Equation 7 could be relaxed if actors and partners have

specific roles within a network (e.g., children as actors and teachers as partners). Lacking such roles to distinguish subjects i and j implies that $\sigma_{E_{ij}}^2 = \sigma_{E_{ji}}^2 = \sigma_E^2$.

Because the actor, partner, and relationship effects in each dyadic observation of Equation 4 are uncorrelated (Snijders & Kenny, 1999, p. 474), the total variance of Y_{ij} can be decomposed by the SRM into the following orthogonal variance components⁴:

$$\sigma_Y^2 = \sigma_A^2 + \sigma_P^2 + \sigma_E^2. \quad (8)$$

When treating each row in Equation 5 separately, the variance of each row of $Y_{\{ij\}}$ is also expressed by Equation 8. However, as the dyad's scores are dependent, a multivariate approach is required. The covariance matrix of the dyad's two responses Σ_Y can be decomposed as follows:

$$\Sigma_Y = \Sigma_{AP} + \Sigma_{PA} + \Sigma_E$$

$$\begin{bmatrix} \sigma_Y^2 & \sigma_{YY} \\ \sigma_{YY} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} \sigma_A^2 & \\ & \sigma_P^2 \end{bmatrix} + \begin{bmatrix} \sigma_P^2 & \\ & \sigma_A^2 \end{bmatrix} + \begin{bmatrix} \sigma_{EE} & \\ & \sigma_{EE} \end{bmatrix}, \quad (9)$$

where the person-level matrix Σ_{PA} is a rearrangement of Σ_{AP} in Equation 6, such that components are ordered [Partner, Actor] rather than [Actor, Partner]:

$$\Sigma_{PA} = \begin{bmatrix} \sigma_P^2 & \\ \rho_{AP}\sigma_A\sigma_P & \sigma_A^2 \end{bmatrix}. \quad (10)$$

Note that the variances on the diagonal of Σ_Y are both equal to σ_Y^2 , but Σ_Y also includes the covariance between the two observations within a dyad (see also Appendix A).

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 statistical interactions.⁵ The variance in each of these main- and interaction effects could then be used in generalizability coefficients. However, when decomposing the variation in the dyadic variables, the dependencies across subjects' actor and partner effects, and across the relationship effects within dyads should be accounted for.

⁴Although this decomposition has been noted in earlier publications (e.g., Jorgensen et al., 2018, p. 30; Malloy, 2018, p. 32, Eq. 2.9), we provide the derivation in Appendix A because we could not locate this in the existing SRM literature.

⁵Note that the relationship effects are already statistical interaction effects between actor and partner effects: actor \times partner = relationship = E .

Therefore, we do not propose a traditional GT-decomposition, but we use the GT rationale to extend the SRM with rater effects.

The rater-extended social relations model (RESRM) is a generalization of the SRM in Equation 5, incorporating 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 random intercept A_i now represents the average (across raters) variability for each actor i , and an additional random intercept α_{ik} captures rater-specific deviations around A_i for each rater k . Likewise, the variability in P_j across raters is represented as a random intercept 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 intercept 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 the RESRM:

$$Y_{ijk} = M + \mu_k + A_i + \alpha_{ik} + P_j + \pi_{jk} + E_{ij} + \varepsilon_{ijk}. \quad (11)$$

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 $\{ij\}$:

$$\mathbf{Y}_{\{ij\}k} = \begin{bmatrix} Y_{ijk} \\ Y_{jik} \end{bmatrix} = \begin{bmatrix} M \\ M \end{bmatrix} + \begin{bmatrix} \mu_k \\ \mu_k \end{bmatrix} + \begin{bmatrix} A_i \\ A_j \end{bmatrix} + \begin{bmatrix} \alpha_{ik} \\ \alpha_{jk} \end{bmatrix} + \begin{bmatrix} P_j \\ P_i \end{bmatrix} + \begin{bmatrix} \pi_{jk} \\ \pi_{ik} \end{bmatrix} + \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix} + \begin{bmatrix} \varepsilon_{ijk} \\ \varepsilon_{jik} \end{bmatrix}. \quad (12)$$

The distributions of the actor, partner and relationship effects in the RESRM (Equations 11 and 12) are provided by Equations 6 and 7. The additional random intercepts for each rater are assumed to be normally distributed with a mean of zero and variance σ_μ^2 ; that is,

$$\mu_k \sim N(0, \sigma_\mu^2). \quad (13)$$

The rater deviations from the actor and partner effects are assumed to be bivariate normally distributed with means of zero and variances σ_α^2 and σ_π^2 , 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; that is:

$$\begin{bmatrix} \alpha_{ik} \\ \pi_{ik} \end{bmatrix} \sim N\left(\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_{\alpha\pi} = \begin{bmatrix} \sigma_\alpha^2 & \\ \rho_{\alpha\pi}\sigma_\alpha\sigma_\pi & \sigma_\pi^2 \end{bmatrix}\right), \quad (14)$$

where $\rho_{\alpha\pi}$ is the correlation between α_{ik} and π_{ik} . The rater deviations from the relationship effects, ε_{ijk} and

Table 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)	$\frac{\sigma_A^2 + \sigma_P^2 + \sigma_E^2}{\sigma_A^2 + \sigma_P^2 + \sigma_E^2 + \sigma_\alpha^2 + \sigma_\pi^2 + \sigma_\epsilon^2}$	$\frac{\sigma_A^2 + \sigma_P^2 + \sigma_E^2}{\sigma_A^2 + \sigma_P^2 + \sigma_E^2 + \frac{\sigma_\alpha^2 + \sigma_\pi^2 + \sigma_\epsilon^2}{K}}$

ϵ_{ijk} , are assumed to be bivariate 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,

$$\begin{bmatrix} \epsilon_{ijk} \\ \epsilon_{jik} \end{bmatrix} \sim N\left(\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_\epsilon = \begin{bmatrix} \sigma_\epsilon^2 & \\ \rho_\epsilon \sigma_\epsilon^2 & \sigma_\epsilon^2 \end{bmatrix}\right). \quad (15)$$

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} (Equation 11) into the following orthogonal variance components:

$$\sigma_Y^2 = \sigma_\mu^2 + \sigma_A^2 + \sigma_\alpha^2 + \sigma_P^2 + \sigma_\pi^2 + \sigma_E^2 + \sigma_\epsilon^2. \quad (16)$$

For the bivariate expression in Equation 12, the covariance matrix between a dyad's two responses, $Y_{\{ij\}k}$, is

$$\Sigma_Y = \begin{bmatrix} \sigma_\mu^2 & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 \end{bmatrix} + \Sigma_{AP} + \Sigma_{\alpha\pi} + \Sigma_{PA} + \Sigma_{\pi\alpha} + \Sigma_E + \Sigma_\epsilon. \quad (17)$$

Matrices Σ_{AP} , $\Sigma_{\alpha\pi}$, Σ_{PA} , Σ_E and Σ_ϵ have been defined in Equations 6, 14, 10, 7, and 15, respectively. Similar to Σ_{PA} being a rearrangement of Σ_{AP} , the matrix $\Sigma_{\pi\alpha}$ is a rearrangement of $\Sigma_{\alpha\pi}$ in Equation 14:

$$\Sigma_{\pi\alpha} = \begin{bmatrix} \sigma_\pi^2 & \\ \rho_{\alpha\pi} \sigma_\alpha \sigma_\pi & \sigma_\alpha^2 \end{bmatrix}. \quad (18)$$

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 statistical interaction effects between these facets, while taking the possible dependencies in the data into account.

Interrater reliability for interdependent social network data

We used the variances in Equation 16 to define ICCs for interrater consistency of actor, partner, and relationship components, using the same rationale as for defining an IRR using the variances in Equation 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 (σ_A^2 , σ_P^2 , or σ_E^2) plus the variance of the statistical interaction effect between the 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, the ICC of interrater consistency of the actor effects is

$$ICC_A(C, K) = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_\alpha^2}{K}}. \quad (19)$$

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 statistical 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 2 provides the ICCs for all possible facets of differentiation in the SRM.

Connections to other SRM reliability estimates

We are not the first to discuss reliability estimates for the SRM. Here we briefly clarify the distinctions and similarities between the RESRM approach and other reliability coefficients for SRM research. Malloy and Kenny (1986) showed that the SRM is a special case of GT and showed that its actor, partner, and relationship components relate to the facets in a GT decomposition. When measuring interpersonal perceptions, actors can be considered as perceivers (raters) and partners are targets (subjects), so ICCs can be used to estimate how much of the total variance is reliable across these perceivers (e.g., $ICC_P = R^2 = \sigma_P^2 / \sigma_y^2$). This ICC_P can thus be considered as an estimate of interrater agreement because it represents agreement among actors of their perceptions about the same set of partners.

Kenny (1994, p. 241) proposed a GT model that extends the SRM by incorporating item- or occasion effects. Bonito and Kenny (2010) implemented this model to define reliability coefficients for estimated SRM random effects. In their work, reliability is defined as the degree to which partner (or actor) effects can be generalized over relationships (their Equations 5 and 6) or the degree to which actor and partner effects can be generalized over both relationships and measurements (i.e., items; their Equations 9 and 10). The denominator of these generalizability coefficients accounts for the dyadic reciprocity, because these coefficients estimate the degree to which actor (but also partner) effects can be generalized over relations (i.e., the source of dependence in SRM data).

The current paper proposes coefficients to investigate the degree to which the random actor, partner, and relationship effects can be generalized over external raters. These raters do not take part in the interactions themselves and are additional to the subjects (i.e., actors and partners) in the social relations study. The GT-based RESRM we proposed is similar to Bonito and Kenny's (2010) model, however, instead of including an additional item facet, we included a rater facet as measurement facet. Furthermore, our coefficients differ from those of Bonito and Kenny (2010) because we aim to investigate how well actor, partner and relationships scores can be generalized across these independent raters, rather than across actors and partners that themselves take part in the social interactions. Therefore, the ICCs we propose in Table 2 do not include relationship variance or dyadic reciprocity in the denominator. Our ICCs are thus specifically useful for validating or improving rating procedures when external raters are used to obtain dyadic data for an SRM study.

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 packages (e.g., MLwiN; Rasbash et al., 2017), but many multilevel modeling 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

modeling approach (e.g., Nestler et al., 2020) and a fully Bayesian approach (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., Jorgensen, 2021; Ten Hove et al., 2020, 2022, 2024a; Vispoel et al., 2018, 2019). We will focus on the Bayesian approach because under conditions that are expected to occur frequently in observational social-network research, this approach outperformed the IRR estimates obtained using the structural equation modeling approach (Ten Hove et al., 2024a). 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 a Bayesian approach as proposed for a similar network model by Hoff (2005) and more recently described for the SRM by Lüdtke et al. (2013). Let θ be the vector of all model parameters. The posterior distribution $P(\theta|Y)$ of the model parameters given the data is estimated as proportional to the product of the prior probability distribution $P(\theta)$ and the likelihood of the data, conditional on the parameters $P(Y|\theta)$. By using non-informative priors for estimating $P(\theta|Y)$ through Markov chain Monte Carlo (MCMC) estimation, the estimated posterior distribution is primarily influenced by the observed data, and the model estimates are similar to those obtained using MLE.

MCMC algorithms can 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 .⁶ Assuming a fully crossed design in which all raters observe all dyadic interactions, and all subjects interact with all other subjects, parameter vector θ includes the following $12 + 2N + 2D + K + 2NK$ quantities: The grand mean M (Equation 11); 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; ⁷ Equation 16); correlations ρ_{AP} (Equation 6), ρ_E (Equation 7), $\rho_{\alpha\pi}$ (Equation 14), and ρ_ϵ (Equation 15); subject-level random effects $[A_i P_i]$ for $i = 1, \dots, N$, dyad-level random effects $[E_{ij} E_{ji}]$ for $i = 1, \dots, N; j = 1, \dots, N; i \neq j$, rater-specific deviations μ_k for $k = 1, \dots, K$, and rater-specific deviations from subject-level random effects $[\alpha_{ik} \pi_{ik}]$ for $i = 1, \dots, N; k = 1, \dots, K$ (Equation 12). If the design is not fully crossed, each rater observes a subset of dyads rather than all dyads, or

⁶ $D = N(N - 1)/2$ only if all subjects interact with all other subjects. When subjects interact in small subgroups, $D < N(N - 1)/2$.

⁷Depending on the software, these hyperparameters are estimated in terms of random-effect SDs, or random-effect variances.

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 $\boldsymbol{\theta}$ is the bivariate normal likelihood of the observed data; that is,

$$\begin{bmatrix} Y_{ijk} \\ Y_{jik} \end{bmatrix} \sim N \left(\begin{bmatrix} \hat{Y}_{ijk} \\ \hat{Y}_{jik} \end{bmatrix}, \begin{bmatrix} \sigma_{\varepsilon}^2 & \rho_{\varepsilon} \sigma_{\varepsilon}^2 \\ \rho_{\varepsilon} \sigma_{\varepsilon}^2 & \sigma_{\varepsilon}^2 \end{bmatrix} \right). \quad (20)$$

Equation 20 is specified using the vector of predicted values for $\mathbf{Y}_{\{ij\}k}$, given all random intercepts:

$$\begin{aligned} \hat{\mathbf{Y}}_{\{ij\}k} &= \begin{bmatrix} \hat{Y}_{ijk} \\ \hat{Y}_{jik} \end{bmatrix} = \begin{bmatrix} M \\ M \end{bmatrix} + \begin{bmatrix} \mu_k \\ \mu_k \end{bmatrix} + \begin{bmatrix} A_i \\ A_j \end{bmatrix} \\ &+ \begin{bmatrix} \alpha_{ik} \\ \alpha_{jk} \end{bmatrix} + \begin{bmatrix} P_j \\ P_i \end{bmatrix} + \begin{bmatrix} \pi_{jk} \\ \pi_{ik} \end{bmatrix} + \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix}. \end{aligned} \quad (21)$$

The prior distributions of the random effects are defined by equations 6, 7, 13, 14, and 15. The parameters (i.e., variance components, and correlations) are hyperparameters requiring 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. In the empirical example below, we further elaborate on the technical specifications of the estimation procedure.

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 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/b4nfv/>

Method

Participants

In total, $N = 139$ German students participated in the study of Salazar Kämpf et al. (2018). Each student was randomly assigned to a group of four to six members, forming 26 same-sex networks. Each student had an interaction with all other group 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-min 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 $\text{ICC}(2, 3) = .87$.⁸ 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 Lütcke 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 $\boldsymbol{\theta}$ (i.e., the mean, the random-effect variances or SDs, the random-effect correlations, and all random effects) sequentially, NUTS simultaneously samples the entire vector $\boldsymbol{\theta}$ 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 SDs instead of variances. We derived posterior distributions of all ICCs listed in Table 2 from the posterior SD estimates. A Bayesian credible 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 (Lütcke 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 (Ten Hove et al., 2024b): <https://osf.io/9az5x>.

⁸The $\text{ICC}(2, 3)$ as defined by Shrout and Fleiss (1979) is identical to the $\text{ICC}(C, K)$ in Equation 3.

Table 3. Estimated variance components using the SRM (by Salazar Kämpf et al., 2018) and using the RESRM.

Variance	SRM		RESRM					N_{eff}	\hat{R}
	Est	% variance	MAP	2.5%	97.5%	% variance			
σ_A^2	0.32	24	0.37	0.26	0.54	27	866	1.00	
σ_P^2	0.07	6	0.09	0.04	0.17	7	182	1.02	
σ_E^2	0.47	35	0.46	0.37	0.58	34	497	1.00	
σ_μ^2	–	–	0.00	0.00	0.39	< 1	270	1.00	
σ_α^2	–	–	0.11	0.06	0.15	8	161	1.01	
σ_π^2	–	–	0.00	0.00	0.05	< 1	105	1.02	
σ_ϵ^2	–	–	0.33	0.30	0.36	24	903	1.00	
σ_{error}^2	0.46	35	–	–	–	–	–	–	
σ_{Total}	1.32	100	1.39	1.27	1.81	100	427	1.00	

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 (N_{eff}), using $\hat{R} < 1.10$ and $N_{eff} > 100$ 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 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 (see our supplementary material on the Open Science Framework).

Prior distributions. We specified weakly informative prior distributions for each parameter, assuming that the standard deviations ($\sigma_A, \sigma_P, \sigma_E, \sigma_\mu, \sigma_\alpha, \sigma_\pi, \text{ and } \sigma_\epsilon$) followed a half- $t(4, 0, 1)$ distribution, with a range of (0, 3), which is half the range of Y and therefore the largest a SD could possibly be. These priors are specifically useful for studies as these, when variances are estimated from (very) small samples of subjects and raters, and possibly close to the lower-bound of zero (Ten Hove et al., 2020). Given our lack of theoretical expectations regarding the correlations between the effects within dyads or persons ($\rho_{AP}, \rho_E, \rho_{\alpha\pi}, \text{ and } \rho_\epsilon$), we assumed that these correlations were uniformly distributed across the range of $(-1, 1)$.

Results

Variance decomposition

Table 3 shows all SRM parameters as estimated with maximum likelihood by Salazar Kämpf et al. (2018), and all RESRM parameters that we estimated with

Table 4. RESRM-based ICC estimates.

ICC	Single ratings (C, 1)			Averaged ratings (C, K)		
	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; E = Relationship effects. The ICC as reported by Salazar Kämpf et al. (2018, i.e., ICC(2,3) = 0.87), resembled the conflated ICC (i.e., ICC_{Conf}) for averaged ratings as estimated with the `irr` package.

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_μ^2), and variance components for the rater deviations from the actor effects (σ_α^2), partner effects (σ_π^2), and relationship effects (σ_ϵ^2). 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%).

Interrater reliability

Table 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.

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.

Simulation study

Method

Data generation

We conducted a simulation study to gain a first impression of the bias and coverage of the RESRM-based ICC estimates 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 12 using the parameters in Equations 6, 7, 13, 14, 15, and 16.

Independent variables

The factor *design* had two levels: A *good* design with substantial and balanced sample sizes, and an *poor* design based on the empirical example.⁹ 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 $\frac{10 \times 9}{2} = 45$ 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 four 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 factor *parameters* also had two levels: A (co)variance structure with *substantial* RESRM parameters, and a (co)variance structure with *varying* parameters. The substantial population parameters we selected were: $\sigma_A = \sigma_P = \sigma_R = \sigma_\pi = \sigma_\alpha = \sigma_\epsilon = 1.00$ and $\rho_{AP} = \rho_E = \rho_{\alpha\pi} = \rho_\epsilon = .30$. That is, all components followed a standard-normal distribution, and within-person and dyadic correlations differed considerably from zero. The varying population parameters were based on the empirical example: $\sigma_A = 0.60$, $\sigma_P = 0.30$, $\sigma_E = 0.70$, $\sigma_\alpha = 0.30$, $\sigma_\pi = 0.10$, $\sigma_\epsilon = 0.60$, $\rho_{AP} =$

⁹We 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 and poor conditions to estimate IRR coefficients may differ from good and poor conditions for drawing inferences about individuals' attributes in SRM analyses.

Table 5. Population ICCs for averaged ratings of the integrated score, and of the actor-, partner-, and relationship effects across simulation conditions.

Parameters	Design	Ratings	ICC			
			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; E = Relationship effects.

.70, $\rho_E = .70$, $\rho_{\alpha\pi} = -.30$, $\rho_\epsilon = .20$. These population parameters were specified as SDs and correlations because Stan output provides standard-deviation components rather than variance components.

This simulation design yielded 2 (design) \times 2 (parameters) = 4 conditions in total, for each of which we generated 1,000 datasets. The resulting population ICCs ranged from 0.50 to 0.90 for single ratings, and from 0.75 to 0.99 for averaged ratings (Table 5).

Estimation

We used the NUTS method discussed earlier for parameter estimation and added an automated convergence check. 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 thus provides a measure of systematic over- or underestimation of the true ICCs. We interpreted relative bias between 0.05 and 0.10 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 0.95 (Agresti & Coull, 1998). We considered only BCI-coverage rates $< .90$ practically too low.

Results

The model converged for almost all replications, varying from 98% (good-design, varying-parameters

Table 6. Relative bias across simulation conditions.

ICC	<i>M(SD)</i>	Substantial parameters		Varying parameters	
		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. 95% BCI coverage rates across simulation conditions.

ICC	<i>M(SD)</i>	Substantial parameters		Varying parameters	
		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.

condition) to 100% (good design, substantial-parameters condition). Averaged across conditions, the RESRM provided unbiased estimates with good coverage rates for most ICCs (Tables 6 and 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.

Bias

Figure 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 substantially overestimated, and the ICCs of the relationship effects were substantially underestimated. This bias can be explained by the combination of few raters, small groups of interacting subjects, and highly correlated rater effects in this condition, which produces an underestimation of the variance of the rater-

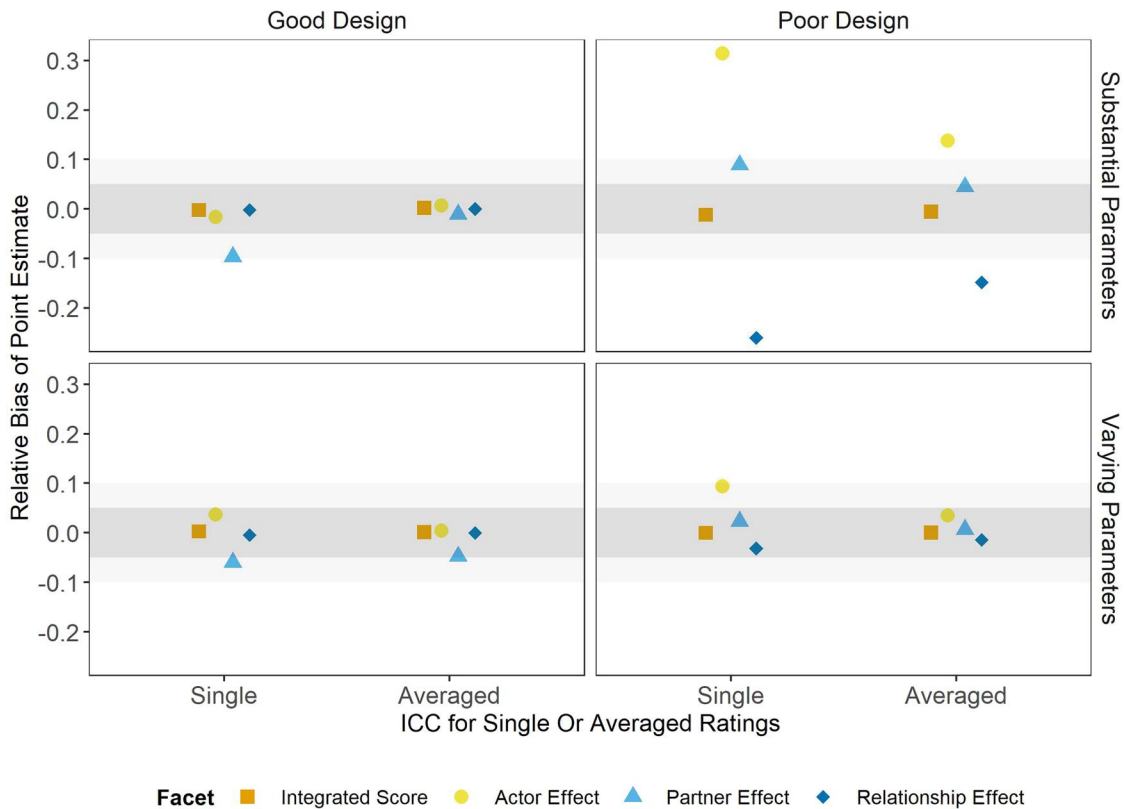


Figure 1. Relative bias of the ICCs across conditions. White areas: substantial bias (>10%); Light-gray areas: minor bias (5–10%); Dark-gray areas: negligible bias (< 5%).



Figure 2. Coverage of the ICCs across conditions. White areas: Practically too low coverage rates < 90%.

deviation from the actor effects ($s_x^2 = 0.71$), and overestimation of the rater-deviation from the relationship effects ($s_e^2 = 1.30$). Using these biased variance-component estimates in the formulae for the ICCs (Table 2) produced biased ICCs. In the poor-design, varying-parameters condition, the ICCs for single ratings of the actor effects were slightly overestimated.

Coverage

Figure 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.

Discussion

We proposed, illustrated, and tested an RESRM to estimate the IRR of interdependent social network data. Using an empirical example on 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 if the research design includes sufficient raters and a substantial number of interacting subjects.

The simulation conditions provide a first impression of the performance estimator under favorable and unfavorable conditions. The study showed considerable differences between the bias and coverage of the ICC estimates for the good and poor design conditions. In conditions with both small subgroups of interacting subjects and a few number of raters, the RESRM could not accurately estimate all ICCs, especially if the magnitude of variances and dependence of observations was substantial. This was to be expected, because little information generally leads to biased point or SE estimates of variances. The bias in the poor design, substantial parameters condition indicated that the RESRM-based ICCs cannot be trusted for designs with few raters and small subgroups of subjects. We therefore advise against drawing conclusions based on the IRR estimates for the empirical example and would advise using the RESRM-based ICCs for studies with similar designs. 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, the type of social network design and the degree of dependence in the social network, 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 and one substantially large group of subjects. 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 the poor-design conditions ($K = 3$), the good-design conditions were more efficient as they required 900 ratings in total whereas the poor-design conditions required 2154 ratings. The good design conditions used fewer participants than may be desirable for an SRM study, but the design is useful to inspect the quality of the rating procedure. If researchers develop rating procedures for SRM research we suggest to use a validation study to inspect the quality of the rating procedure in terms of 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.

Traditional IRR coefficients are not useful in SRM research, because such coefficients do not consider the IRR for each SRM component separately. 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. The RESRM-based IRR coefficients can inform researchers in improving rating procedures for dyadic variables, by identifying which components are most prone to rater effects. We therefore believe that the RESRM approach is a promising conceptual and analytical tool for evaluating the IRR of dyad-level predictors in social relations research.

References

- Agresti, A., & Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119–126. <https://doi.org/10.1080/00031305.1998.10480550>
- Bonito, J. A., & Kenny, D. A. (2010). The measurement of reliability of social relations components from round-robin designs. *Personal Relationships*, 17(2), 235–251. <https://doi.org/10.1111/j.1475-6811.2010.01274.x>
- Brennan, R. L. (2001). *Generalizability theory*. Springer.
- Card, N. A., & Hodges, E. V. E. (2010). It takes two to fight in school, too: A social relations model of the psychometric properties and relative variance of dyadic aggression and victimization in middle school. *Social Development (Oxford, England)*, 19(3), 447–469. <https://doi.org/10.1111/j.1467-9507.2009.00562.x>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Coie, J. D., Cillessen, A. H. N., Dodge, K. A., Hubbard, J. A., Schwartz, D., Lemerise, E. A., & Bateman, H. (1999). It takes two to fight: A test of relational factors and a method for assessing aggressive dyads. *Developmental Psychology*, 35(5), 1179–1188. <https://doi.org/10.1037/0012-1649.35.5.1179>
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16(2), 137–163. <https://doi.org/10.1111/j.2044-8317.1963.tb00206.x>
- Gamer, M., Lemon, J., Fellows, I., Singh, P. (2012). *irr: Various coefficients of interrater reliability and agreement [Computer Software]*. <https://CRAN.R-project.org/package=irr>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. LLC.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
- Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469), 286–295. <https://doi.org/10.1198/01621450400001015>
- Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn’t hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology*, 113(3), 430–452. <https://doi.org/10.1037/pspi0000097>
- Hughes, B. T., Flournoy, J. C., & Srivastava, S. (2021). Is perceived similarity more than assumed similarity? An interpersonal path to seeing similarity between self and others. *Journal of Personality and Social Psychology*, 121(1), 184–200. <https://doi.org/10.1037/pspp0000369>
- Johnson, P. E. (2016). *rockchalk: Regression estimation and presentation [Computer Software]*. <https://cran.r-project.org>
- Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych*, 3(2), 113–133. <https://doi.org/10.3390/psych3020011>
- Jorgensen, T. D., Forney, K. J., Hall, J. A., & Giles, S. M. (2018). Using modern methods for missing data analysis with the social relations model: A bridge to social network analysis. *Social Networks*, 54, 26–40. <https://doi.org/10.1016/j.socnet.2017.11.002>
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. Guilford.
- Kenny, D. A. (1996). Models of non-independence in dyadic research. *Journal of Social and Personal Relationships*, 13(2), 279–294. <https://doi.org/10.1177/0265407596132007>
- Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the big five. *Psychological Bulletin*, 116(2), 245–258. <https://doi.org/10.1037/0033-2909.116.2.245>
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *The analysis of dyadic data*. Guilford.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 142–182). Academic Press.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Lüdtke, O., Robitzsch, A., Kenny, D. A., & Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychological Methods*, 18(1), 101–119. <https://doi.org/10.1037/a0029252>
- Majdandžić, M., de Vente, W., & Bögels, S. M. (2016). Challenging parenting behavior from infancy to toddlerhood: Etiology, measurement, and differences between fathers and mothers. *Infancy*, 21(4), 423–452. <https://doi.org/10.1111/inf.12125>
- Malloy, T. E. (2018). *Social relations modeling of behavior in dyads and groups*. Academic Press. <https://doi.org/10.1016/C2016-0-02324-0>
- Malloy, T. E., & Kenny, D. A. (1986). The social relations model: An integrative method for personality research. *Journal of Personality*, 54(1), 199–225. <https://doi.org/10.1111/j.1467-6494.1986.tb00393.x>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using hamiltonian monte carlo. *Methods in Ecology and Evolution*, 8(3), 339–348. <https://doi.org/10.1111/2041-210X.12681>
- Nestler, S., Lüdtke, O., & Robitzsch, A. (2020). Maximum likelihood estimation of a social relations structural equation model. *Psychometrika*, 85(4), 870–889. <https://doi.org/10.1007/s11336-020-09728->
- Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *The Journal of Applied Psychology*, 93(5), 959–981. <https://doi.org/10.1037/0021-9010.93.5.959>
- R Core Team. (2021). *R: A language and environment for statistical computing [Computer Software]*. R Foundation for Statistical Computing.

- Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2017). *A user's guide to MLwiN, v3.00*. Centre for Multilevel Modelling, University of Bristol.
- Salazar Kämpf, M., Liebermann, H., Kerschreiter, R., Krause, S., Nestler, S., & Schmukle, S. C. (2018). Disentangling the sources of mimicry: Social relations analyses of the link between mimicry and liking. *Psychological Science*, 29(1), 131–138. <https://doi.org/10.1177/0956797617727121>
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44(6), 922–932. <https://doi.org/10.1037/0003-066X.44.6.922>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Simpkins, S. D., & Parke, R. D. (2002). Do friends and non-friends behave differently? A social relations analysis of children's behavior. *Merrill-Palmer Quarterly*, 48(3), 263–283. <https://doi.org/10.1353/mpq.2002.0014>
- Snijders, T. A. B., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, 6(4), 471–486. <https://doi.org/10.1111/j.1475-6811.1999.tb00204.x>
- Stan Development Team. (2020). RStan: The R interface to Stan. [R package version 2.21.2].
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2020). Comparing hyperprior distributions to estimate variance components for interrater reliability coefficients. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 84th annual meeting of the Psychometric Society* (pp. 79–93). Springer. https://doi.org/10.1007/978-3-030-43469-4_7
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2022). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, 27(4), 650–666. <https://doi.org/10.1037/met0000391>
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2024a). How to estimate intraclass correlation coefficients for interrater reliability from planned incomplete data. [Under review].
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2024b). Supplementary materials to “Interrater reliability for interdependent social network data: A generalizability theory approach”. <https://doi.org/10.17605/osf.io/9az5x>
- Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2024c). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*, 29(5), 967–979. <https://doi.org/10.1037/met0000516>
- Vanbelle, S. (2017). Comparing dependent kappa coefficients obtained on multilevel data. *Biometrical Journal*, 59(5), 1016–1034. <https://doi.org/10.1002/bimj.201600093>
- Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, 61(1), 295–304. <https://doi.org/10.1111/j.0006-341X.2005.031040.x>
- Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, 23(1), 1–26. <https://doi.org/10.1037/met0000107>
- Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods*, 24(2), 153–178. <https://doi.org/10.1037/met0000177>
- Warner, R. M., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, 37(10), 1742–1757. <https://doi.org/10.1037/0022-3514.37.10.1742>
- Winer, B. J. (2013). *Statistical principles in experimental design* (2nd ed.). McGraw-Hill.
- Yang, Z., & Zhou, M. (2014). Kappa statistic for clustered matched-pair data. *Statistics in Medicine*, 33(15), 2612–2633. <https://doi.org/10.1002/sim.6113>
- Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association*, 36(1), 419–480. <https://doi.org/10.1080/23808985.2013.11679142>

Appendix A

Variance decomposition

We derive the variance decomposition of Y for the SRM. The same principles apply to the RESRM, with similar results, complicated only by the inclusion of more terms (μ_k , α , π , and ε).

The bivariate vector $\mathbf{Y}_{\{ij\}}$ is decomposed as shown in Equation 5. The mean does not contribute to the variance because it is a constant, and the remaining terms all have expected values of zero. Thus, the second central moment (variance) is simply the expected value of squared random effects. For example, the variance of actor effect (A) is:

$$\text{Var}(A) = E[(A - E[A])^2] = E[A^2], \quad (\text{A.1})$$

and likewise for partner (P) and relationship (E) effects. To prevent confusion with the expectation operator $E[\cdot]$, the remainder of this appendix refers to the relationship effect E_{ij} (e.g., Equation 12) with the variable $R_{(ij)}$.

For the univariate SRM in Equation 4, the variance of Y_{ij} is

$$\text{Var}(Y_{ij}) = E[(Y - E[Y])^2] \quad (\text{A.2})$$

$$= E[(M + A_i + P_j + R_{ij} - E[M + A_i + P_j + R_{ij}])^2] \quad (\text{A.3})$$

$$= E[(M + A_i + P_j + R_{ij} - (E[M] + E[A_i] + E[P_j] + E[R_{ij}]))^2], \quad (\text{A.4})$$

which follows from the property that the expectation of a sum (or difference) equals the sum (or difference) of expectations. The expected value of a constant is itself ($E[M] = M$), so the means cancel out M , and Equation A.4 reduces to

$$\text{Var}(Y_{ij}) = E[(A_i + P_j + R_{ij} - E[A_i] - E[P_j] - E[R_{ij}])^2] \quad (\text{A.5})$$

$$= E[(A_i + P_j + R_{ij} - 0 - 0 - 0)^2] \quad (\text{A.6})$$

$$= E[(A_i + P_j + R_{ij})^2]. \quad (\text{A.7})$$

The expected value of each random effect is zero, further simplifying the formula above before squaring the

parenthetical term:

$$\text{Var}(Y_{ij}) = E[(A_i + P_j + R_{ij})^2] \tag{A.8}$$

$$= E[A_i^2 + P_j^2 + R_{ij}^2 + 2(A_iP_j + A_iR_{ij} + P_jR_{ij})] \tag{A.9}$$

$$= E[A_i^2] + E[P_j^2] + E[R_{ij}^2] + 2 \times (E[A_iP_j] + E[A_iR_{ij}] + E[P_jR_{ij}]) \tag{A.10}$$

$$= E[A_i^2] + E[P_j^2] + E[R_{ij}^2] + 2 \times (0 + 0 + 0) \tag{A.11}$$

$$= E[A_i^2] + E[P_j^2] + E[R_{ij}^2]. \tag{A.12}$$

Because the random intercepts are uncorrelated across (person and dyad) levels, the expectation of the product of R_{ij} with A_i or P_j is zero. Likewise, the person-level random effects are independent across cases i and j , so the product A_iP_j has an expectation of zero. The simplified formula in Equation A.12 contains only the remaining expectations of squared random effects, which correspond to the variance components, as shown in Equation A.1 for the actor effect:

$$\text{Var}(Y_{ij}) = E[A_i^2] + E[P_j^2] + E[R_{ij}^2] \tag{A.13}$$

$$= \text{Var}(A) + \text{Var}(P) + \text{Var}(R) \tag{A.14}$$

$$\sigma_Y^2 = \sigma_A^2 + \sigma_P^2 + \sigma_R^2. \tag{A.15}$$

The same result can be found when deriving the variance of Y_{ji} , by swapping the i and j subscripts in Equation A.3 and proceeding with the same steps. The variance is equal for both observations in the bivariate vector $\mathbf{Y}_{\{ij\}}$, as implied by Equation 9:

$$\Sigma_{Yij} = \Sigma_{AP} + \Sigma_{PA} + \Sigma_E \tag{A.16}$$

$$= \begin{bmatrix} \sigma_A^2 & \\ \sigma_{AP} & \sigma_P^2 \end{bmatrix} + \begin{bmatrix} \sigma_P^2 & \\ \sigma_{AP} & \sigma_A^2 \end{bmatrix} + \begin{bmatrix} \sigma_R^2 & \\ \sigma_{RR} & \sigma_R^2 \end{bmatrix} \tag{A.17}$$

$$= \begin{bmatrix} \sigma_A^2 + \sigma_P^2 + \sigma_R^2 & \\ 2\sigma_{AP} + \sigma_{RR} & \sigma_A^2 + \sigma_P^2 + \sigma_R^2 \end{bmatrix} \tag{A.18}$$

$$= \begin{bmatrix} \sigma_Y^2 & \\ \sigma_{YY} & \sigma_Y^2 \end{bmatrix}. \tag{A.19}$$

Covariance decomposition

We use covariance algebra of linear combinations to derive the decomposition of the covariance σ_{YY} between the two observations in $\mathbf{Y}_{\{ij\}}$:

$$\text{Cov}(Y_{ij}, Y_{ji}) = \text{Cov}(M + A_i + P_j + R_{ij}, M + A_j + P_i + R_{ji}) \tag{A.20}$$

$$= \text{Cov}(A_i + P_j + R_{ij}, A_j + P_i + R_{ji}) \tag{A.21}$$

$$= \text{Cov}(A_i, A_j) + \text{Cov}(A_i, P_i) + \text{Cov}(A_i, R_{ji}) + \text{Cov}(P_j, A_j) + \text{Cov}(P_j, P_i) + \text{Cov}(P_j, R_{ji}) + \text{Cov}(R_{ij}, A_j) + \text{Cov}(R_{ij}, P_i) + \text{Cov}(R_{ij}, R_{ji}) \tag{A.22}$$

$$= 0 + \text{Cov}(A_i, P_i) + 0 + \text{Cov}(P_j, A_j) + 0 + 0 + 0 + 0 + \text{Cov}(R_{ij}, R_{ji}). \tag{A.23}$$

The remaining covariances in Equation A.23 correspond to the parameters that compose σ_{YY} in Equation A.18: σ_{YY} between the two observations in $\mathbf{Y}_{\{ij\}}$:

$$\text{Cov}(Y_{ij}, Y_{ji}) = \text{Cov}(A_i, P_i) + \text{Cov}(P_j, A_j) + \text{Cov}(R_{ij}, R_{ji}) \tag{A.24}$$

$$= 2 \times (\text{Cov}(A_i, P_i) + \text{Cov}(R_{ij}, R_{ji})) \tag{A.25}$$

$$\sigma_{YY} = 2 \times \sigma_{AP} + \sigma_{RR}. \tag{A.26}$$

As with the variance decomposition, the same principles yield similar results for the RESRM, complicated only by the inclusion of more terms.