Interrater reliability for incomplete and dependent data

ten Hove, D.

Publication date
2023

Chapter 1

General Introduction

1.1 Introduction

The central question of this dissertation is how to estimate the interrater reliability (IRR) from incomplete and dependent observational data. This topic is of importance for observational measurement of attributes in social and behavioral practice and research. Examples of such attributes are reading fluency, teaching skills, playfulness, parenting behavior, and academic skills. Because these attributes are not directly observable, practitioners and scientists use methods such as self-report questionnaires or observations to obtain information about these attributes. Observational measurement refers to obtaining information about attributes by means of external raters.

In practice, scores based on observations are used to make decisions about subjects. For example, academic teachers (the raters) rate students' (the subjects) theses to gain information about their academic skills (the attribute). These ratings are used to decide whether individual students qualify for a master's degree. As another example, child protection officers (the raters) evaluate the recidivism risk (the attribute) of juvenile delinquents (the subjects), to provide judges with a sentencing recommendation (c.f. Van der Put et al., 2011). For fair assessment and sound decision making, it is important that these ratings reflect subjects' attribute levels, instead of differences across raters and their perspectives.

In research, scores based on observations are used to investigate the relation between different attributes, or to investigate the effect of external criteria on subjects' attributes. For example, Faber et al. (2018) used raters to measure teachers' (the subjects) differentiated instruction (the attribute), to investigate whether such instruction could predict the mathematical achievement of students (external criterion). Also, Majdandžić et al. (2016) used raters to assess parents' (the subjects) challenging parenting behavior (the attribute), and investigated if this behavior differed between fathers and mothers (external criterion). In such research settings, it is important that the observed differences across subjects are not caused by differences across raters and their perspectives. Otherwise, the conclusions of the statistical analyses that are used to answer the research question are prone to error.

1.2 Interrater Reliability

Ideally, the variation in ratings of subjects' attributes thus originates from differences among subjects, and as little as possible from the differences among the raters that provide these ratings. However, ratings may be affected by several rater-related factors; for example, some raters may be more strict than others, the observation protocol may be unclear, or raters may be affected by all sorts of cognitive biases (Tversky & Kahneman, 1974). Such differences across raters and their perspectives do not provide information about the attributes of subjects, but they do influence the observed attribute scores of subjects.

The degree to which observed attribute scores are independent of raters can be inspected with IRR coefficients. Different conceptualizations of IRR exist (for an overview, see e.g., Gwet, 2014; Hallgren, 2012; Zhao et al., 2013), which can roughly be divided into those that conceptualize IRR in terms of counting (dis)agreements (e.g., Cohen, 1960, 1968; Fleiss, 1971; Krippendorff, 1970) and those that conceptualize IRR in terms of variance components (e.g., Shrout & Fleiss, 1979). The most commonly reported coefficient is Cohen (1960)'s kappa, but for continuous measures the intraclass correlation coefficient (ICC; see e.g., McGraw & Wong, 1996; Shrout & Fleiss, 1979) is often used.

Before ratings are used in further analyses or decision making the IRR of ratings should be investigated (e.g., AERA et al., 2018). Using ratings with low IRR may result in incorrect decisions in diagnostic settings, biased estimates, or loss of power in substantive analyses (cf. Lord & Novick, 1968, p. 72). However, since IRR is not uniquely defined, we have an abundance of coefficients that may all measure something else. The effect on power and precision cannot be uniquely determined either. IRR is thus of great importance for social and behavioural research, but it is of utmost importance that we define a single conceptualization of IRR.

1.3 Incomplete and Nested Observational Designs

The standard observational design, for which traditional IRR coefficients are defined, is a complete (two-way) design (Figure 1.1, panel a). In such a design, multiple independent subjects (s) are each rated by multiple independent raters (r), and these raters are the same for each subject. However, in observational research in the social and behavioural sciences, the raters are often not the same for each subject, resulting in incomplete data (Figure 1.1, panel b). Also, the rated subjects are often nested within clusters (Figure 1.1, panel c) or relationships (Figure 1.1, panel d), resulting in dependent data.

Incomplete observational data may occur due to all sorts of missing-data processes. Missing ratings may be unforeseen (nonresponse, attrition) but they often are planned to make an observational study feasible in terms of costs, rater burden, or logistics. The raters for each subject are then sampled from a larger pool of independent raters in a planned incomplete observational design, and the raters partly vary across subjects. Let $y_{s.r}$ denote a rating of subject s by rater r. Figure 1.1, panel b, provides an example of such a planned incomplete design in which subject 1 is not observed by Rater 3, as is the case for subjects 4 and 7, whereas, subjects 2, 5, and 8 are not observed by rater 2, and subjects 3, 6, and 9 are not observed by rater 3. Examples of such planned incomplete observational designs in the literature include Zee et al. (2020) and Yuen et al. (2020).

Dependent observational data results from, among other designs, a multilevel design (Figure 1.1, panel c) or a social network design (Figure 1.1, panel d). In observational multilevel designs, raters rate subjects (e.g., students) that are nested within clusters (e.g., teachers or schools; Faber et al., 2018; Helms-Lorenz et al., 2016; Harmsen et al., 2018). Let $y_{s.c.r}$ denote a rating of subject s who is nested within cluster c, and rated by

(a) Standard Observational Design

Subject	Rater		
	1	2	3
1	$y_{1.1}$	$y_{1.2}$	$y_{1.3}$
2	$y_{2.1}$	$y_{2.2}$	$y_{2.3}$
3	$y_{3.1}$	$y_{3.2}$	$y_{3.3}$
4	$y_{4.1}$	$y_{4.2}$	$y_{4.3}$
5	$y_{5.1}$	$y_{5.2}$	$y_{5.3}$
6	$y_{6.1}$	$y_{6.2}$	$y_{6.3}$
7	$y_{7.1}$	$y_{7.2}$	$y_{7.3}$
8	$y_{8.1}$	$y_{8.2}$	$y_{8.3}$
9	$y_{9.1}$	$y_{9.2}$	$y_{9.3}$

(b) Incomplete Observational Design

Subject	Rater		
	1	2	3
1	$y_{1.1}$	$y_{1.2}$	?
2	$y_{2.1}$	?	$y_{2.3}$
3	?	$y_{3.2}$	$y_{3.3}$
4	$y_{4.1}$	$y_{4.2}$	?
5	$y_{5.1}$	?	$y_{5.3}$
6	?	$y_{6.2}$	$y_{6.3}$
7	$y_{7.1}$	$y_{7.2}$	?
8	$y_{8.1}$	?	$y_{8.3}$
9	?	$y_{9.2}$	$y_{9.3}$

(c) Observational Multilevel Design

Subject	Rater	Cluster	
		1	2
1	1	$y_{1.1.1}$	-
	2	$y_{1.1.2}$	-
	3	$y_{1.1.3}$	-
2	1	$y_{2.1.1}$	-
	2	$y_{2.1.2}$	-
	3	$y_{2.1.3}$	-
3	1	-	$y_{3.2.1}$
	2	-	$y_{3.2.2}$
	3	-	$y_{3.2.3}$
4	1	-	$y_{4.2.1}$
	2	-	$y_{4.2.2}$
	3	-	$y_{4.2.3}$

(d) Observational Social Network Design

Rater	Subject	Subject		
		1	2	3
1	1	-	$y_{1.2.1}$	$y_{1.3.1}$
	2	$y_{2.1.1}$	-	$y_{2.3.1}$
	3	$y_{3.1.1}$	$y_{3.2.1}$	-
2	1	-	$y_{1.2.2}$	$y_{1.3.2}$
	2	$y_{2.1.2}$	-	$y_{2.3.2}$
	3	$y_{3.1.2}$	$y_{3.2.2}$	-
3	1	-	$y_{1.2.3}$	$y_{1.3.3}$
	2	$y_{2.1.3}$	-	$y_{2.3.3}$
	3	$y_{3.1.3}$	$y_{3.2.3}$	-

Figure 1.1: Examples of Four Observational Designs: (a) a standard observational design, in which each subject is assessed by the same raters; (b) an incomplete observational design in which the raters partly vary across subjects; (c) a multilevel observational design in which raters rate subjects that are nested within clusters; and (d) an observational social network designs, raters rate interactions between subjects. ? = Observation is missing; − = Observation does not exist.

rater r. Figure 1.1, panel c, provides an example of such a multilevel design in which, for example $y_{1.1.3}$ denotes the rating of subject 1 who is nested in cluster 1, and rated by rater 3. Subjects within clusters are often more alike than subjects between clusters, so the ratings of subjects in the same clusters are dependent. Subject-level data from multilevel settings are often modelled at both the subject level and at an aggregated cluster level. This cluster-level aggregate can even have its own qualitative interpretation. For example, researchers rated the tension towards a teacher as displayed in students' drawings, and classes of students were taught by different teachers (clusters; e.g., Goble et al., 2019; Zee et al., 2020; Chen et al., 2022). Inferences could then be made about individual students'

tension towards their teacher (the subject level) or about a classroom's average tension towards individual teachers (the cluster level). Multilevel data obtained by raters thus has multiple facets of theoretical interest, of which a researcher should investigate the IRR.

In observational social network designs, raters rate interactions between subjects (e.g., Salazar Kämpf et al., 2018; Hughes et al., 2021; Huang et al., 2017). In such interactions, subjects have interchangeable actor roles (showing behaviors) and partner roles (eliciting behaviors), in interactions with various subjects in a social network. For example, Salazar Kämpf et al. (2018) used raters to measure subjects' social mimicry in a social network design. Let $y_{s.s'.r}$ denote a rating of subject s in interaction with another subject s', and rated by rater r. Figure 1.1, panel d, provides an example of such a social network design in which, for example, $y_{1.2.2}$ denotes the rating of subject 1's mimicry of subject 2, as rated by rater 2. The observed social mimicry scores could be decomposed into actor effects (the degree to which subjects, on average, mimic the other subjects), partner effects (the degree to which subjects, on average, elicit mimicry from the other subjects), and relationship effects (the degree to which subjects deviate from the—based on actor and partner effects—expected mimicry in interaction with a specific subject). Also, interactions are nested in subjects. For example, Subject 1's social mimicry is rated in interaction with both Subject 2 ($Y_{1,2,r}$) and Subject 3 ($y_{1,3,r}$). The actor and partner scores of individual subjects are therefore dependent across interactions (generalized reciprocity; Kenny, 1996; Kenny & La Voie, 1984). The interactions are also nested in relationships. For example, the raters observe the interaction between subjects 1 and 2 to rate both subject 1's mimicry of subject 2's behavior ($Y_{1,2,r}$), and subject 2's mimicry of subject 1's behavior ($Y_{1,2,r}$). The relationship effects within actor-partner dyads are therefore also dependent (dyadic reciprocity; Kenny, 1996; Kenny & La Voie, 1984). The interdependent actor, partner, and relationship effects imply multiple interdependent facets of theoretical interest, for which a researcher should investigate the IRR, while accounting for data dependencies.

1.4 IRR for Incomplete and Dependent Data

The complex data structures resulting from incomplete and nested observational designs have implications for the definition and estimation of the IRR. For incomplete data, only few IRR can be estimated (exceptions include Krippendorff, 1980; Putka et al., 2008). Most estimation procedures require complete observations, and use list-wise deletion to remove rows with missing data (e.g., all rows in Figure 1.1, panel b). Also, it is unclear how the IRR of ratings is affected by designs in which the raters vary across subjects.

For dependent data, the different facets of interest imply different data components of which a researcher should investigate the IRR (for a similar discussion concerning test reliability, see, Geldhof et al., 2014). To estimate the IRR of dependent data, IRR should be thus be conceptualized for each of the potential facets of interest. For multilevel data,

only one IRR coefficient has been defined that can handle the dependency structures in the data (an extension of Cohen's kappa; Vanbelle et al., 2012). For interdependent social network data, no IRR coefficients have been proposed yet.

None of the conceptualizations of IRR discussed in Section 1.2 can be readily used for both incomplete and dependent data. Ignoring the distinct facets of theoretical interest results in less informative IRR estimates. To investigate the degree to which the observed subject and cluster scores in multilevel data, or the actor, partner, and relationship effects in interdependent social network data are independent of raters, IRR should be estimated for each of these facets of interest separately. Also, estimation methods are needed to accommodate missing observations and dependence structures in dependent data. Ideally, a single conceptualization of IRR provides information about implications of the (lack of) IRR of different facets of theoretical interest for practice (i.e., incorrect decisions) and research (e.g., attenuation of correlation, power), and should be estimable for both incomplete and dependent data.

1.5 Aim and Outline

Using the framework of Generalizability theory (GT; Cronbach et al., 1963), this dissertation provides definitions and estimation methods of IRR for incomplete and dependent data. As a starting point, **Chapter 2** illustrates and discusses the issue of an abundance of IRR coefficients that follow from different conceptualizations of IRR. The IRR coefficients that are freely available in the `R` software package `irr` (Gamer et al., 2012) are described, alongside their properties, and all IRR coefficients are applied to the same four datasets to investigate whether the coefficients provide similar estimates of the IRR for a given data set.

One of the coefficients in the `irr` package is the intraclass correlation coefficient (ICC; Bartko, 1966; McGraw & Wong, 1996; Shrout & Fleiss, 1979), which is also implemented in other standard software for social and behavioral research (e.g., IBM Corp., 2021; JASP Team, 2022). In **Chapter 3**, I explain why the ICC is probably the best candidate for defining IRR. I investigated the various definitions of ICCs for one-way designs, in which each subject has a unique set of raters, and two-way designs, in which the set of raters is the same for each subject. Choosing between these ICCs is complicated, and IRR researchers seem to have conflicting opinions about when to use which ICCs. Using GT, the different definitions of ICCs are extended to incomplete observational designs, and updated guidelines are provided on when to use which ICC definition. The resulting guidelines are summarized in a flow-chart, which is applied to three examples from clinical and developmental domains.

The variance components that are used to define ICCs are traditionally estimated using mean-squares from an ANOVA model, which is not straightforward for incomplete and dependent observational data. One approach that can handle such data is Markov chain Monte Carlo (MCMC) estimation of hierarchical linear models (MCMC-HL), a fully

Bayesian approach. This Bayesian approach requires the choice of hyperprior distributions for the variance components that are used to define the ICCs. To select an appropriate hyperprior distribution for estimating the ICCs, **Chapter 4** investigated the effect of different hyperprior distributions on ICC estimates under different conditions that are common to observational studies, such as small variance components and small samples of raters and subjects.

Next to MCMC-HL, two maximum likelihood estimation methods have been proposed to estimate generalizability coefficients, the GT equivalent of ICCs: maximum likelihood estimation of random-effects (MLE-RE) models, and maximum likelihood estimation of common-factor models (MLE-CF). **Chapter 5** describes a simulation study that compared MCMC-HL, MLE-RE, and MLE-CF to estimate ICCs. The performance of the estimators is evaluated based on computational accuracy (bias of point estimates, bias of variability estimates, root mean squared error, and coverage rates) and computational feasibility (convergence rates and estimation time), and the design factors have a strong focus on incomplete observational designs, for which estimation methods were lacking. Ratings of subjects are often of nominal or ordinal measurement levels, which complicates estimation procedures for IRR coefficients founded on variance components. This chapter, therefore also investigates extensions of the three discussed estimation techniques for binary data. The chapter ends with a tutorial, using data on communication skills of clinicians in training. The tutorial illustrates the problems of traditional ANOVA-like approaches for incomplete data, and shows how researchers can use the preferred methods to estimate ICCs from incomplete data using software that is freely available on the Open Science Framework.

In **Chapter 6**, the ICCs for independent data (i.e., the one-way or two-way ICCs in **Chapter 3**) are generalized to ICCs for multilevel data. Using GT, IRR coefficients are defined for the subject- and cluster component in the data, and hierarchical linear models are proposed to estimate these coefficients. The quality of MCMC-based point and interval estimates are investigated in two simulation studies. The first simulation study investigated the computational accuracy of the ICC estimates in conditions with varying sample sizes of raters and subjects, and varying magnitudes of variance components. The second simulation study investigated whether a planned missing data design could improve the ICC estimates in conditions with few raters per subject. The chapter ends with an illustration of the method on data about student–teacher relationships, using software that is freely available on the Open Science Framework.

In **Chapter 7**, a definition of IRR for interdependent social network data is established, by extending the social relations model (Kenny & La Voie, 1984; Kenny, 1996) with rater effects. It defines IRR in terms of ICCs for the actor, partner, and relationship facets separately, and proposes a hierarchical model and MCMC approach to estimate these ICCs. The chapter includes an illustration of the method on data about social mimicry from Salazar Kämpf et al. (2018), using software that is freely available on the Open Science Framework. The quality of the MCMC estimates is inspected in a simulation

study.

Finally, **Chapter 8** provides a general discussion of the results in this dissertation. It includes a concluding remark about the findings in this dissertation, a discussion on implications of these findings for estimating IRR and designing observational studies, and a discussion of potential directions for future IRR developments and research.