Factors affecting efficiency of interrater reliability estimates from planned missing data designs on a fixed budget

van der Ark, L.A.; Jorgensen, T.D.; ten Hove, D.

DOI
10.1007/978-3-031-27781-8_1

Publication date
2023

Document Version
Submitted manuscript

Published in
Quantitative psychology

Factors Affecting Efficiency of Interrater Reliability Estimates from Planned Missing Data Designs on a Fixed Budget

Terrence D. Jorgensen[1] ✉[0000−0001−5111−6773], L. Andries van der Ark[1] [0000−0003−3131−7943], and Debby ten Hove[1,2] [0000−0002−1335−4452]

[1] Universiteit van Amsterdam,
Postbus 15776, 1001NG Amsterdam, the Netherlands,
T.D.Jorgensen@uva.nl
[2] Vrije Universiteit Amsterdam,
De Boelelaan 1105, 1081HV Amsterdam, the Netherlands

Abstract. Estimating interrater reliability (IRR) requires each of multiple subjects to be observed by multiple raters. Recruiting subjects and raters may be problematic: There may be few available, it may be costly to compensate subjects or to train raters, and participating in an observational study may be burdensome. Planned missing observational designs, in which raters vary across subjects, may accommodate these problems, but little guidance is available about how to optimize a planned missing observational design when estimating IRR. In this study, we used Monte Carlo simulations to optimize an observational design to estimate intraclass correlation coefficients (ICCs), which are very flexible IRR estimators that allow missing observations. We concluded that, given a fixed total number of ratings, the point and credibility estimates of ICCs can be optimized by means of (approximately) continuous measurement scales and assigning small teams of raters to subgroups of subjects. Also, less substantial differences between raters resulted in more efficient IRR estimates. These results highlight the importance of well-designed observational designs and proper training on an observational protocol to avoid substantial differences between raters.

Keywords: interrater reliability, intraclass correlation, generalizability theory, planned missing data, observational design

1 Introduction

This chapter provides evidence we gathered to plan a complex observational study for the Netherlands Ministry of Justice and Security to estimate the interrater reliability (IRR) of the National Instrument of the Juvenile Criminal Justice System, which is known by its acronym LIJ (pronounced like the English word "lie"; Van der Put et al., 2011). The LIJ is used to predict the risk of recidivism and to identify protective factors and risk factors of minors who are suspect of a criminal case. Completing the LIJ includes an officer of the Netherlands Child Care and Protection Board (rater) separately interviewing both the

juvenile (subject) and at least one caretaker to obtain answers to almost 200 questions. This procedure typically takes several workdays spent on reading the police files, conducting the two interviews, and obtaining additional information from and verifying information with, for example, social workers or teachers (see Van der Ark et al., 2018, with a summary in English on p. 5).

Estimating IRR requires that each subject is assessed by multiple raters. Three main challenges complicated investigating the IRR of the LIJ. First, a lack of time. The officers—who also have other important job responsibilities—lacked the time to obtain multiple ratings of the same juveniles. Second, the pool of raters and subjects to choose from was limited. Ecologically valid ratings require raters who are real officers and subjects who are real juveniles within the justice system. Third, recording interviews with the juveniles would be too ethically risky, but raters were required to make observations at the same time and location. The LIJ was administered on 18 different locations in The Netherlands. Obtaining multiple ratings of the same subjects was thus complicated by constrains on time and resources. From a pragmatic perspective, each juvenile was preferably assessed by a minimal number of raters from a local team.

Sampling few raters minimizes the burden on subjects and raters, but maximizes sampling variability of IRR estimates. Because the stakes of the juvenile delinquents were high, precise IRR estimates were required. Planned missing observational designs in which the raters vary across subjects enable using a larger sample of raters while keeping the burden on individual raters stable. Guidance in optimizing such a planned missing observational design to yield precise IRR estimates is currently lacking. In this chapter, we therefore discuss a simulation study that aimed to yield IRR estimates with maximal precision while minimizing the burden on raters.

1.1 Intraclass Correlation Coefficients

IRR coefficients that can accommodate incomplete data are rare (e.g., Krippendorff's α; Hayes and Krippendorff, 2007), but an advantageous choice is the intraclass correlation coefficient (ICC), which has long been used to quantify IRR (Bartko, 1966; Fleiss and Cohen, 1973; Shrout and Fleiss, 1979). A family of ICC coefficients can be derived from the broad framework of generalizability theory (GT; Cronbach et al., 1963; Brennan, 2001), which was developed for normally distributed variables (McGraw and Wong, 1996) and can be calculated from variance components estimable using a linear mixed model (Jiang, 2018):

$$Y_{sr} = \mu + \mu_s + \mu_r + \mu_{sr}, \tag{1}$$

where μ_s and μ_r are main subject and rater effects, respectively, and μ_{sr} is the subject $\times$ rater interaction (confounded with any other source of measurement error). Assuming independent effects with means of zero, the orthogonal variance components sum to the total variance of Y_{sr}:

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

An ICC quantifying absolute agreement among raters expresses the variance between subjects relative to all sources of variance (McGraw and Wong, 1996):

$$\text{ICC(A, 1)} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2}, \tag{3}$$

interpreted as the degree to which subjects' absolute scores can be generalized over raters (relevant when evaluating whether a subject meets an absolute criterion; Vispoel et al., 2018; Ten Hove et al., 2022). If, in practice, judgments would be made by averaging scores across $k > 1$ raters, reliability would be increased by reducing rater-related error: $\frac{\sigma_r^2 + \sigma_{sr}^2}{k}$.

Equation 1 can be extended by relying on the latent response variable (LRV) interpretation of a probit model (Agresti, 2007). Assuming an observed outcome X—measured using a discrete scale with categories[3] $c = 0, \ldots, C$—is a crude indicator of an underlying continuum Y:

$$X_{sr} = c \text{ if } \tau_c < Y_{sr} \leq \tau_{c+1}, \tag{4}$$

a standard linear-model interpretation is applicable to the LRV Y_{sr}, under an identification constraint that the residual variance $\sigma_{sr}^2 = 1$. An advantage is that ICCs can be comparable across studies that used different response scales, such as binary vs. 5- or 7-point Likert scales (Zumbo et al., 2007; Vispoel et al., 2019). The LRV approach has recently been proposed for generalizability coefficients (of which ICCs are a special case; Ten Hove et al., 2022; Vispoel et al., 2018) using structural equation modeling (SEM; Vispoel et al., 2019; Ark, 2015), which Jorgensen (2021) showed could be problematic for sparse data from PMD designs. The current study investigates a generalized linear mixed model (GLMM) with a (cumulative) probit link function for ordinal outcomes, which can estimate variance components from a fully crossed design even with incomplete data.

1.2 Planned Missing Data

Planned missing data (PMD) designs were conceived to reduce participant burden in large scale surveys (Graham et al., 1996). We introduce some terminology to facilitate discussing PMD designs in the context of multirater studies. Regardless of whether the limits are monetary, we refer to the total number of ratings (N_{Ratings}) as the *budget*. A fixed budget could be limited not only by time and monetary constraints but also by the numbers of available subjects and raters. We further define *workload* as the number of subjects per rater ($N_{S/R}$) and *team size* as the number of raters per subject ($N_{R/S}$). How the budget is allocated depends on a number of features, listed in Table 1. The overall number of subjects and raters are represented by N_S and N_R. Different (sub)samples of the pool of raters might be assigned to each subject, and different (sub)samples of

[3] There are actually $C + 2$ thresholds, but the lowest and highest thresholds are fixed by definition to be the lowest and highest possible scores in the Y distribution; because the normal distribution is unbounded, $\tau_0 = -\infty$ and $\tau_{C+1} = +\infty$.

the subject pool may be assigned to each rater. In a fully crossed two-way design with complete data, $N_{\text{Ratings}} = N_S \times N_R$ because the number of subjects assigned to each rater ($N_{S/R} = N_S$) is the entire subject pool; likewise, the number of raters assigned to each subject ($N_{R/S} = N_R$) is the entire rater pool. Incomplete designs are still crossed but do not assign each rater to every subject (Ten Hove et al., 2022). Putka et al. (2008) referred to *ill-structured measurement designs* when assignment was not systematic or optimal, but thoughtfully deployed PMD designs can be economically advantageous in multirater studies with expensive or time-consuming observational protocols (e.g., Vial et al., 2019; Zee et al., 2020; Yuen et al., 2020).

Table 1. Trade-Off Among Rater-Pool Size, Subject-Pool Size, Team Size, and Workload Given a Fixed Budget

A *smaller* pool of ...	... requires assigning *more* ...
subjects [a]	raters per subject [c] (larger teams)
raters [b]	subjects per rater [d] (greater workload)
Assigning *fewer* ...	**... requires a *larger* pool of ...**
raters per subject [c] (smaller teams)	subjects [a]
subjects per rater [d] (lighter workload)	raters [b]

Note. Budget = total number of ratings ($N_{\text{Ratings}} = N_R \times N_{S/R} = N_S \times N_{R/S}$), assuming equal team sizes across subjects and equal workload across raters. When using a block design (i.e., nonoverlapping teams), additionally useful design features can be derived, albeit redundant with the features above: block size $= N_{S/R} \times N_{R/S}$ and $N_{\text{Blocks}} = \frac{N_{\text{Ratings}}}{\text{block size}}$.
[a] Subject pool: $N_S = N_R \times \frac{N_{S/R}}{N_{R/S}}$. [b] Rater pool: $N_R = N_S \times \frac{N_{R/S}}{N_{S/R}}$.
[c] Team size: $N_{R/S} = N_{S/R} \times \frac{N_R}{N_S}$. [d] Workload: $N_{S/R} = N_{R/S} \times \frac{N_S}{N_R}$.

Randomly or systematically assigning a $N_{S/R}$ subset of the subject pool to be observed by each rater (or vice versa: a $N_{R/S}$ subset of the rater pool is assigned to observe each subject) has been shown to improve accuracy of estimated variance components used to calculate an ICC to represent IRR (Ten Hove et al., 2020, 2021). For example, Yuen et al. (2020) randomly assigned a team of two raters to each subject in a staggered fashion that maximized the overlap among raters (i.e., each possible pair of raters rated the same subject at least once). If only $N_R = 2$ raters had observed all $N_S = 29$ subjects, each rater would have a workload of $N_{S/R} = 29$. Instead, each of $N_R = 6$ raters had a substantially lower workload of only $N_{S/R} = 9$ or 10. Thus, given a fixed budget ($N_{\text{Ratings}} = 58$), sampling the same team size ($N_{R/S} = 2$) from a larger pool of $N_R = 6$ raters reduced the workload by $\frac{N_{R/S}}{N_R} = 1/3$.

The simulation study described next was designed to decide how IRR of the LIJ could be most efficiently estimated under budget constraints. The results

led Van der Ark et al. (2018) to evaluate the LIJ by assigning teams of $N_R = 4$ raters to evaluate $N_{S/R} = 2$ subjects each.

2 Method

To develop an observational design for estimating the IRR of the scales and items of the LIJ, we conducted a set of Monte Carlo simulations. We provide our R syntax for replicating our simulation on the Open Science Framework (OSF[4]).

2.1 Data generation

The two-way model in Equation 1 was used to generate normal random effects for all conditions, with $\mu = 0$ for the grand mean and all random-effect means, $\sigma_s^2 = 0.70$, $\sigma_r^2 = 0.15$, and $\sigma_{sr}^2 = 0.25$. These population variances implied a population ICC(A,1) = 0.636, denoted ρ. For ordinal conditions, thresholds $\tau_1 = -0.5$ and $\tau_2 = 0.5$ were used to discretize the normal data into $C = 3$ categories. To keep the generated data comparable across conditions, the data were always generated from a fully crossed design for a given N_S and N_R. Then, missing data patterns were imposed to yield a certain number of complete-data "blocks" (i.e., $N_{\text{Blocks}} = N_{R/S} \times N_{S/R}$) that yielded a fixed budget of $N_{\text{Ratings}} = 384$. The proportion of missing data in two-way designs[5] (i.e., $N_{S/R} > 1$) varied from 83.33% (in conditions with the largest blocks) to 98.96% (with the smallest workload $N_{S/R} = 2$ and team size $N_{R/S} = 2$, requiring the largest N_R and N_S).

Core design factors. The design factors we had most control over were workload and team size given a fixed budget, and we planned to estimate ICCs for continuous, ordinal, and binary items. So our core factors were team size ($N_{R/S} = 2, 4,$ or 8), workload ($N_{S/R} = 1, 2, 4,$ or 8), and model used for generating and analyzing data (linear or probit for continuous or discrete data, respectively), yielding $3 \times 4 \times 2 = 24$ conditions. When $N_{S/R} = 1$, we used an one-way model by removing μ_r from Equation 1 and its variance component σ_r^2 from Equation 2 because when raters are nested in subjects, μ_r is confounded with the rater $\times$ subject interaction. Thus, Eq. 3 still represents ICC(A,1).

Additional design factors. In the results section, we also describe two follow-up studies in which we varied two additional factors: The magnitude of reliability, and random versus block assignment of raters to subjects. We fully crossed these design factors with the core conditions described above, but did not cross these with each other. The results are useful for planning missing observational designs for IRR. Additional manipulations are available in the R scripts provided with the online supplementary materials.

[4] Supplemental online materials available at https://osf.io/g5hvs/

[5] When $N_{S/R} = 1$, there is no "missing-data problem" because raters are nested in (rather than crossed with) subjects.

2.2 Analysis

We used Markov chain Monte Carlo (MCMC) estimation with uninformative priors, implemented in the Stan software (Carpenter et al., 2017), for each of 2000 replications within each condition. We saved the posterior mean (denoted $\hat{\rho}$) as an estimate of ρ, as well as the central 95% Bayesian credible interval (BCI) limits. In each condition, we evaluated accuracy of posterior means as point estimates (Ten Hove et al., 2020) by calculating the relative parameter bias, which is the difference between a condition's average estimate (denoted $\bar{\rho}$) and ρ, divided by ρ: $\frac{\bar{\rho}-\rho}{\rho}$. We evaluated accuracy of BCIs by calculating 95% coverage rates (i.e., proportion of replications whose intervals captured ρ) in each condition. Finally, we evaluated precision (our primary criterion for choosing an optimal design for the LIJ evaluation) of the estimates by calculating the average width of 95% BCIs in each condition.

We investigate the effects of design factors on bias and precision using fully factorial linear regression models (ANOVA) and on the coverage using fully factorial binary logistic regression models (analysis of deviance).

3 Results

For brevity, we report only medium and larger effects (i.e., Monte Carlo design factor accounts for $\eta^2_{\text{p}} > 6\%$ of variance, holding other effects constant) on bias or precision, or 6% of deviance in coverage (analogous to McFadden's[6] pseudo-R^2). More extensive results are provided on the OSF.

3.1 Core conditions

Bias was negligible across conditions ($M_{\text{bias}} = -0.01$, $SD = 0.01$), and no design factors explained more than 0.05% of variance in bias. Coverage was nominal across conditions ($M_{\text{cov}} = 0.94$, $SD = 0.01$), and no design factors explained more than 0.05% of deviance in coverage. Precision was substantially affected only by the scale (continuous or ordinal: $\eta^2_{\text{p}} = 14.74\%$) and team size ($\eta^2_{\text{p}} = 11.38\%$). ICC(A,1) was more precisely estimated for continuous data (95% BCI width: $M = 0.19_{\text{width}}$, $SD = 0.03$) than for ordinal data ($M_{\text{width}} = 0.23$, $SD = 0.02$). ICC(A,1) was more precisely estimated using teams of $N_{R/S} = 2$ ($M_{\text{width}} = 0.20$, $SD = 0.03$) or $N_{R/S} = 4$ ($M_{\text{width}} = 0.20$, $SD = 0.03$) than for teams of $N_{R/S} = 8$ ($M_{\text{width}} = 0.23$, $SD = 0.02$).

An explanation of why smaller teams yielded more precise estimates may be that—holding other factors constant—assigning smaller teams ($N_{R/S}$) maximizes N_S (Table 1). Because σ^2_s should be expected to be the largest component of an ICC in practice (e.g., for even a modest IRR ≥ 0.50), a more efficiently estimated σ^2_s could lead to a more efficiently estimated ρ. The next simulation additionally varied the amount of rater error, illuminating this explanation.

[6] Find descriptions of several types of pseudo-R^2 for logistic regression here: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

3.2 Magnitude of ICC

Rater variance was fixed to $\sigma_r^2 = 0.25$ in the core conditions, implying a modest $\rho = .636$. Because we expected ICCs to vary across LIJ items, we added conditions with more rater variance ($\sigma_r^2 = 0.70$, implying lower IRR: $\rho = .42$) and with less rater variance ($\sigma_r^2 = 0.05$, implying higher IRR: $\rho = .70$). Extending the 24 core conditions by varying $\sigma_r^2 = 0.05$, 0.25, or 0.70 yielded 72 conditions.

Bias was still negligible across conditions ($M_{\text{bias}} = -0.002$, $SD = 0.014$) and no design factors explained more than 0.03% of variance in bias. Coverage also still was nominal across conditions ($M_{\text{cov}} = 0.94$, $SD = 0.01$), and no design factors explained more than 0.01% of deviance in coverage. Efficiency was substantially affected by ρ, which explained $\eta_{\text{p}}^2 = 15.64\%$ of the variability in BCI width, whereas the influential factors from the core conditions explained only $\eta_{\text{p}}^2 = 5\%$ of the variability in BCI width, holding other factors constant in this extended design. The higher IRR was in the population, the more precisely it was estimated ($\rho = 0.70$: $M_{\text{width}} = 0.19$, $SD = 0.03$; $\rho = 0.64$: $M_{\text{width}} = 0.21$, $SD = 0.03$; $\rho = 0.42$: $M_{\text{width}} = 0.25$, $SD = 0.03$). This was consistent with our explanation for why smaller teams yielded more precise estimates under a fixed budget; it is not a general rule that fewer raters (per subject) yield more precision (Ten Hove et al., 2021).

3.3 Overlapping teams

In the core conditions, we imposed a missing-data structure that mimicked the blocks assigned in the LIJ study. We compared this to unstructured random assignment by randomly deleting all but $N_{R/S}$ ratings for each subject. This strategy meant that the workload could vary across raters, with an average (rather than fixed) workload of $N_{S/R}$. This design is comparable to Yuen et al. (2020), who designed a balanced workload across raters (i.e., fixed $N_{S/R}$). Because overlapping teams implies a two-way design, we omitted the $N_{S/R} = 1$ conditions. Thus, this study had a 3 ($N_{R/S} = 2, 4,$ or 8) $\times$ 3 ($N_{S/R} = 2, 4,$ or 8) $\times$ 2 (scale) $\times$ 2 (teams overlap or not) design with 36 conditions.

Bias was still negligible across conditions ($M_{\text{bias}} = -0.005$, $SD = 0.013$) and no design factors explained more than 0.11% of variance in bias. Coverage also still was nominal across conditions ($M_{\text{cov}} = 0.94$, $SD = 0.01$), and no design factors explained more than 0.04% of deviance in coverage. Precision was also not substantially affected by overlapping teams; the main and all higher-order effects combined only explained $\eta_{\text{p}}^2 = 2.74\%$ of the additional variability in BCI width beyond the core design.

4 Discussion

This chapter provided evidence from Monte Carlo simulations demonstrating how beneficial planned missing observational designs can be for expensive, time-consuming multirater studies. Results showed that MCMC estimation of MLMs

and GLMMs can provide accurate point and interval estimates of ICCs across a variety of population values, scales of measurement, and planned missing observational designs, even when the vast majority of observations of a conventional (fully crossed) two-way design are missing. Bias and coverage appeared stable across the selected design factors but we showed that the precision of ICC estimates can be maximized by using more (approximately) continuous scales of measurement and allocating smaller teams of raters to subjects. These results highlight the importance of well designed observational designs. In addition, less rater error also improved efficiency, which highlights the importance of proper training on an observational protocol to avoid substantial differences between raters.

4.1 Advice for Sample-Size Planning

In practice, researchers must weigh the costs of different design features to choose the best design for their situation (e.g., Are raters or subjects more expensive? Does the gain in efficiency warrant the additional effort?), and accounting for such costs was not explored in our simulation studies. Certain design features might also be more difficult to control than others. Holding other features constant, smaller teams and lighter workloads might require larger pools of subjects and raters, respectively, either of which might be infeasible.

In the LIJ study, for example, the number of available juveniles turned out to be quite limited. Furthermore, the greatest cost was the burden on each rater, so workload was of primary concern in the design. With a fixed budget, minimizing team size to improve precision would have been coincident with maximizing N_S, which was not feasible. However, smaller teams (larger N_S) only improved precision by a few decimal places, and workload had no discernible effect precision, so we felt justified advising the ministry to assign fewer subjects to larger teams for LIJ data collection.

Overlapping raters does not seem to have any (dis)advantage, so researchers can feel free to randomly assign raters to subjects using whichever algorithm best fits their needs. Overlapping raters might be more feasible if the ratings need not be conducted at a fixed time point; for example, if the subjects have been recorded, or if the observation is made on objects (like critics judging artwork or experts evaluating the face validity of a measurement instrument). Systematic overlap is not necessary, but might be more desirable to ensure balanced workload across raters (see Yuen et al., 2020). In contrast, random assignment to blocks (within which subjects and raters are fully crossed) would be more feasible when live observations of the same event must be made at the same time, as in the LIJ evaluation.

We conducted these simulations for a specific setting (evaluating IRR of the LIJ), and showed that some general design factors can improve the efficiency of ICC estimates. We hope that our example helps other researchers make such decisions, but future research is needed to provide advice for other scenarios that have different priorities for working within a budget.

Bibliography

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley, Hoboken NJ, 2 edition.

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework*. PhD thesis, University of British Columbia, Vancouver, BC, CA.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19(1):3–11.

Brennan, R. L. (2001). *Generalizability Theory*. Springer, New York, NY.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16(2):137–163.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Graham, J. W., Hofer, S. M., and MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31(2):197–218.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R: A `lme4` package application. *Methodology*, 14(3):133–142.

Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych*, 3(2):113–133.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

Putka, D. J., Le, H., McCloy, R. A., and Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, 93(5):959–981.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Ten Hove, D., Jorgensen, T. D., and Van der Ark, L. A. (2020). Comparing hyperprior distributions to estimate variance components for interrater reliability coefficients. In Wiberg, M., Molenaar, D., González, J., Böckenholt, U., and Kim, J.-S., editors, *Quantitative psychology: The 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019*, pages 79–93, New York, NY. Springer.

Ten Hove, D., Jorgensen, T. D., and Van der Ark, L. A. (2021). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*.

Ten Hove, D., Jorgensen, T. D., and Van der Ark, L. A. (2022). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*.

Van der Ark, L. A., Van Leeuwen, J. L., and Jorgensen, T. D. (2018). Interbeoordelaarsbetrouwbaarheid LIJ: Onderzoek naar de interbeoordelaarsbetrouwbaarheid van het landelijk instrumentarium jeugdstrafrechtketen [Interrater reliability LIJ: Research on the interrater reliability of the national instrument of the juvenile criminal justice system]. Technical report, Wetenschappelijk Onderzoek- en Documentatiecentrum, The Hague, the Netherlands. Retrieved from http://hdl.handle.net/20.500.12832/2267.

Van der Put, C., Spanjaard, H., Van Domburgh, L., Doreleijers, T., Lodewijks, H., Ferwerda, H., Bolt, R., and Stams, G. J. (2011). Ontwikkeling van het landelijke instrumentarium jeugdstrafrechtketen (LIJ) [Development of the national instrument of the juvenile criminal justice system]. *Kind & Adolescent Praktijk*, 10(2):76–83.

Vial, A., Assink, M., Stams, G. J. J. M., and Van der Put, C. (2019). Safety and risk assessment in child welfare: A reliability study using multiple measures. *Journal of Child and Family Studies*, 28:3533–3544.

Vispoel, W. P., Morris, C. A., and Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, 23(1):1–26.

Vispoel, W. P., Morris, C. A., and Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods*, 24(2):153–178.

Yuen, J. K., Kelley, A. S., Gelfman, L. P., Lindenberger, E. E., Smith, C. B., Arnold, R. M., Calton, B., Schell, J., and Berns, S. H. (2020). Development and validation of the ACP-CAT for assessing the quality of advance care planning communication. *Journal of Pain and Symptom Management*, 59(1):1–8.

Zee, M., Rudasill, K. M., and Roorda, D. L. (2020). "Draw me a picture": Student–teacher relationship drawings by children displaying externalizing, internalizing, or prosocial behavior. *The Elementary School Journal*, 120(4):636–666.

Zumbo, B. D., Gadermann, A. M., and Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, 6(1):21–29.