Interrater reliability for incomplete and dependent data

ten Hove, D.

Publication date
2023

Appendices

Appendix A Performance of ICC Estimators

Table A1: Mean, SD, min and max of all Dependent Variables across Conditions with $\sigma_r^2 = 0.50$ and $\sigma_{sr}^2 = 1.00$ and Continuous Responses

Outcome	ICC	Method	M	SD	min	max
Bias	ICC(A, 1)	MCMC-HL	-0.10	0.09	-0.38	-0.01
		MLE-RE	0.00	0.02	-0.03	0.04
		MLE-CF	-0.01	0.02	-0.05	0.04
	ICC(C, 1)	MCMC-HL	-0.01	0.02	-0.06	0.01
		MLE-RE	-0.01	0.01	-0.03	-0.00
		MLE-CF	0.01	0.03	-0.02	0.10
Bias uncertainty measures	ICC(A, 1)	MCMC-HL	-0.01	0.21	-0.31	0.50
		MLE-RE	0.01	0.03	-0.08	0.06
		MLE-CF	-0.28	0.17	-0.71	-0.06
	ICC(C, 1)	MCMC-HL	-0.05	0.08	-0.28	0.06
		MLE-RE	-0.02	0.04	-0.10	0.06
		MLE-CF	-0.04	0.07	-0.22	0.04
RMSE	ICC(A, 1)	MCMC-HL	0.19	0.16	0.00	0.75
		MLE-RE	0.19	0.15	0.00	0.71
		MLE-CF	0.26	0.20	0.01	0.69
	ICC(C, 1)	MCMC-HL	0.23	0.19	0.01	0.56
		MLE-RE	0.22	0.18	0.01	0.56
		MLE-CF	0.28	0.38	0.01	1.46
Coverage	ICC(A, 1)	MCMC-HL	0.93	0.02	0.88	0.96
		MLE-RE-DM	0.91	0.04	0.79	0.96
		MLE-RE-MC	0.94	0.05	0.79	0.99
		MLE-CF-DM	0.80	0.13	0.38	0.93
		MLE-CF-MC	0.82	0.12	0.39	0.94
	ICC(C, 1)	MCMC-HL	0.95	0.01	0.93	0.97
		MLE-RE-DM	0.94	0.02	0.90	0.96
		MLE-RE-MC	0.96	0.01	0.94	0.97
		MLE-CF-DM	0.92	0.05	0.75	0.96
		MLE-CF-MC	0.95	0.02	0.89	0.96
Percentage Convergence		MCMC-HL	95.56	4.61	87.00	100.00
		MLE-RE	99.76	0.22	99.10	100.00
		MLE-CF	97.59	5.57	81.60	100.00
Estimation Time (sec.)		MCMC-HL	427.26	167.64	170.07	629.51
		MLE-RE-DM	6.97	3.78	1.95	15.73
		MLE-RE-MC	8.87	4.18	3.01	17.00
		MLE-CF-DM	1775.30	2293.72	9.26	7625.45
		MLE-CF-MC	1779.76	2296.55	9.94	7633.63

Table A2: Mean, SD, min and max of all Dependent Variables across Conditions with $\sigma_r^2 = 0.50$ and $\sigma_{sr}^2 = 1.00$ and Binary Responses

Outcome	ICC	Method	M	SD	min	max
Bias	ICC(A, 1)	MCMC-HL	0.00	0.10	-0.20	2.00
		MLE-RE	-0.21	0.09	-0.39	2.00
		MLE-CF	-0.08	0.26	-0.85	2.00
	ICC(C, 1)	MCMC-HL	0.15	0.15	-0.03	2.00
		MLE-RE	-0.26	0.08	-0.44	2.00
		MLE-CF	-0.12	0.27	-0.87	2.00
Bias uncertainty measures	ICC(A, 1)	MCMC-HL	-0.10	0.18	-0.34	2.00
		MLE-CF	-0.38	0.19	-0.72	2.00
	ICC(C, 1)	MCMC-HL	-0.18	0.13	-0.43	2.00
		MLE-CF	-0.27	0.22	-0.74	2.00
RMSE	ICC(A, 1)	MCMC-HL	3.13	1.23	0.74	5.23
		MLE-RE	3.10	1.21	0.71	5.16
		MLE-CF	1.24	1.26	0.00	4.70
	ICC(C, 1)	MCMC-HL	4.71	1.39	2.00	7.81
		MLE-RE	4.66	1.36	2.00	7.81
		MLE-CF	1.60	1.78	0.01	6.89
Coverage	ICC(A, 1)	MCMC-HL	0.95	0.02	0.89	2.00
		MLE-CF-DM	0.67	0.15	0.00	2.00
		MLE-CF-MC	0.67	0.18	0.00	2.00
	ICC(C, 1)	MCMC-HL	0.87	0.08	0.64	2.00
		MLE-CF-DM	0.74	0.16	0.00	2.00
		MLE-CF-MC	0.71	0.23	0.00	2.00
Percentage Convergence		MCMC-HL	96.01	3.95	2.00	100.00
		MLE-RE	98.33	1.13	2.00	100.00
		MLE-CF	53.68	35.81	0.00	100.00
Estimation Time (sec.)		MCMC-HL	86.73	7.35	2.00	98.15
		MLE	0.73	0.17	0.42	2.00
		MLE-CF-DM	1.56	0.75	0.81	3.58
		MLE-CF-MC	3.03	1.28	1.56	6.87

Appendix B Multilevel IRR

The contents of this document contain the Online Supplementary materials to "Inter-rater reliability for Multilevel Data: A Generalizability Theory Approach". Additional materials, such as the Stan models, the R syntax from the simulations and the applied example, and the data used for the applied example are available through the Open Science Framework: https://osf.io/bwk5t/

B.1 Population Values ICCs Simulation 1

Table B3: Population Values of Subject-Level ICCs across Conditions

	σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
k	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
	$\text{ICC}_s(\boldsymbol{A}, 1)$				$\text{ICC}_s(\boldsymbol{A}, k)$			
2	0.60	0.50	0.60	0.50	0.75	0.67	0.75	0.67
5	0.60	0.50	0.60	0.50	0.88	0.83	0.88	0.83
10	0.60	0.50	0.60	0.50	0.94	0.91	0.94	0.91
	$\text{ICC}_s(\boldsymbol{C}, 1)$				$\text{ICC}_s(\boldsymbol{C}, k)$			
2	0.67	0.67	0.67	0.67	0.80	0.80	0.80	0.80
5	0.67	0.67	0.67	0.67	0.91	0.91	0.91	0.91
10	0.67	0.67	0.67	0.67	0.95	0.95	0.95	0.95

Note. ICC = Intraclass correlation; k = Number of raters; σ_c^2 = Variance of cluster effects; σ_r^2 = Variance of rater effects.

Table B4: Population Values of Cluster-Level ICCs across Conditions

	σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
k	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
	$\text{ICC}_c(\boldsymbol{A}, 1)$				$\text{ICC}_c(\boldsymbol{A}, k)$			
2	0.33	0.20	0.61	0.43	0.50	0.33	0.76	0.60
5	0.33	0.20	0.61	0.43	0.71	0.55	0.89	0.79
10	0.33	0.20	0.61	0.43	0.83	0.71	0.94	0.88
	$\text{ICC}_c(\boldsymbol{C}, 1)$				$\text{ICC}_c(\boldsymbol{C}, k)$			
2	0.50	0.50	0.76	0.76	0.67	0.67	0.86	0.86
5	0.50	0.50	0.76	0.76	0.83	0.83	0.94	0.94
10	0.50	0.50	0.76	0.76	0.91	0.91	0.97	0.97

Note. ICC = Intraclass correlation; k = Number of raters; σ_c^2 = Variance of cluster effects; σ_r^2 = Variance of rater effects.

Table B5: Relative Difference in Population Values of Conflated ICCs compared to Multilevel ICCs across Conditions

	σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
k	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
Subject-Level								
	$\text{ICC}_s(\boldsymbol{A}, 1)$				$\text{ICC}_s(\boldsymbol{A}, k)$			
2	2.51	4.88	9.27	15.35	1.00	1.00	5.60	5.60
5	2.51	4.88	9.27	15.35	1.00	1.00	5.60	5.60
10	2.51	4.88	9.27	15.35	1.00	1.00	5.60	5.60
	$\text{ICC}_s(\boldsymbol{C}, 1)$				$\text{ICC}_s(\boldsymbol{C}, k)$			
2	1.45	3.05	5.22	9.25	0.55	0.55	3.03	3.03
5	0.64	1.44	2.26	4.22	0.24	0.24	1.28	1.28
10	0.33	0.76	1.16	2.21	0.12	0.12	0.65	0.65
Cluster-Level								
	$\text{ICC}_c(\boldsymbol{A}, 1)$				$\text{ICC}_c(\boldsymbol{A}, k)$			
2	57.03	156.80	6.26	35.21	-6.27	-6.27	-11.20	-11.20
5	57.03	156.80	6.26	35.21	-6.27	-6.27	-11.20	-11.20
10	57.03	156.80	6.26	35.21	-6.27	-6.27	-11.20	-11.20
	$\text{ICC}_c(\boldsymbol{C}, 1)$				$\text{ICC}_c(\boldsymbol{C}, k)$			
2	33.02	98.04	3.53	21.22	-3.47	-3.47	-6.07	-6.07
5	14.59	46.16	1.53	9.68	-1.48	-1.48	-2.56	-2.56
10	7.56	24.52	0.79	5.08	-0.76	-0.76	-1.30	-1.30

Note. ICC = Intraclass correlation; k = Number of raters; σ_c^2 = Variance of cluster effects; σ_r^2 = Variance of rater effects. Let ICC_{conf} denote the conflated ICC, and let ICC_{ML} the multilevel ICC. Relative difference is computed as $\frac{\text{ICC}_{conf} - \text{ICC}_{ML}}{\text{ICC}_{ML}} * 100$.

Table B6: Relative MAP Bias of Subject-Level ICCs across Conditions

			σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
N_c	N_s	K	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
			$\text{ICC}_s(A,1)$				$\text{ICC}_s(A,k)$			
20	10	2	-0.03	0.06	-0.02	0.06	-0.01	0.04	-0.01	0.04
		5	-0.01	0.03	0.00	0.02	0.00	0.01	0.00	0.01
		10	-0.01	0.01	-0.01	0.01	0.00	0.00	0.00	0.00
	30	2	0.00	0.10	0.02	0.10	0.00	0.07	0.01	0.07
		5	0.00	0.03	0.01	0.03	0.00	0.01	0.00	0.01
		10	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00
40	10	2	-0.02	0.06	-0.02	0.06	-0.01	0.05	-0.01	0.05
		5	0.00	0.03	0.00	0.02	0.00	0.01	0.00	0.01
		10	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00
	30	2	0.00	0.09	0.02	0.09	0.00	0.06	0.01	0.06
		5	0.00	0.03	0.01	0.03	0.00	0.01	0.00	0.01
		10	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00
			$\text{ICC}_s(C,1)$				$\text{ICC}_s(C,k)$			
20	10	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
40	10	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note. ICC = Intraclass correlation; N_c = Number of clusters; N_s = Number of subjects per cluster; K = Number of raters; σ_c^2 = Variance of cluster effects; σ_r^2 = Variance of rater effects. Relative bias was computed as $\frac{\bar{\theta}-\theta}{\theta}$, where $\bar{\theta}$ denotes the average MAP estimate of a parameter (or derived ICC) across replications in a condition, and θ denotes the population parameter in that condition.

B.2 Results Simulation 1

Table B7: Relative MAP Bias of Cluster-Level ICCs across Conditions

			σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
N_c	N_s	K	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
			$\text{ICC}_c(A,1)$				$\text{ICC}_c(A,k)$			
20	10	2	-0.54	-0.41	-0.22	-0.11	-0.42	-0.31	-0.13	-0.05
		5	-0.25	-0.23	-0.06	-0.05	-0.11	-0.13	-0.01	0.00
		10	-0.14	-0.15	-0.04	-0.04	-0.02	-0.04	0.00	0.00
	30	2	-0.53	-0.38	-0.14	-0.04	-0.4	-0.29	-0.06	0.00
		5	-0.15	-0.12	-0.04	-0.02	-0.05	-0.04	-0.01	0.00
		10	-0.08	-0.07	-0.03	-0.02	-0.01	-0.01	0.00	0.00
40	10	2	-0.48	-0.33	-0.13	-0.02	-0.36	-0.25	-0.06	0.01
		5	-0.18	-0.16	-0.04	-0.01	-0.06	-0.06	-0.01	0.01
		10	-0.09	-0.10	-0.02	-0.02	-0.01	-0.02	0.00	0.00
	30	2	-0.41	-0.26	-0.05	0.04	-0.27	-0.15	-0.01	0.06
		5	-0.10	-0.05	-0.02	0.01	-0.03	0.00	0.00	0.02
		10	-0.06	-0.04	-0.01	-0.01	-0.01	0.00	0.00	0.00
			$\text{ICC}_c(C,1)$				$\text{ICC}_c(C,k)$			
20	10	2	-0.13	-0.13	-0.02	-0.02	-0.08	-0.08	-0.01	-0.01
		5	-0.09	-0.09	0.00	-0.01	-0.02	-0.02	0.00	0.00
		10	-0.05	-0.05	0.00	0.00	0.00	0.00	0.00	0.00
	30	2	-0.16	-0.16	-0.01	-0.02	-0.10	-0.11	0.00	-0.01
		5	-0.02	-0.02	0.00	0.00	0.00	0.00	0.00	0.00
		10	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
40	10	2	-0.09	-0.08	0.00	0.00	-0.05	-0.05	0.00	0.00
		5	-0.04	-0.04	0.00	0.00	-0.01	-0.01	0.00	0.00
		10	-0.01	-0.02	0.00	0.00	0.00	0.00	0.00	0.00
	30	2	-0.03	-0.03	0.01	0.00	-0.01	-0.02	0.01	0.00
		5	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
		10	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00

Note. ICC = Intraclass correlation; N_c = Number of clusters; N_s = Number of subjects per cluster; K = Number of raters; σ_c^2 = Variance of cluster effects; σ_r^2 = Variance of rater effects. Relative bias was computed as $\frac{\bar{\theta}-\theta}{\theta}$, where $\bar{\theta}$ denotes the average MAP estimate of a parameter (or derived ICC) across replications in a condition, and θ denotes the population parameter in that condition.

Table B8: Percentage of 95% BCI Coverage of Subject-Level ICCs across Conditions

			σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
N_c	$N_{s:c}$	K	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
			ICC$_s(A,1)$				ICC$_s(A,k)$			
20	10	2	95	98*	95	98*	95	98*	95	98*
		5	95	96	95	96	95	96	95	96
		10	95	95	95	95	95	95	95	95
	30	2	97*	98*	97*	98*	97*	98*	97*	98*
		5	94*	95	94	95	94*	95	94	95
		10	94	96	95	95	94	96	95	95
40	10	2	96*	98*	96	98*	96*	98*	96	98*
		5	95	97*	95	97*	95	97*	95	97*
		10	95	95	95	95	95	95	95	95
	30	2	97*	99*	98*	99*	97*	99*	98*	99*
		5	95	97*	95	97*	95	97*	95	97*
		10	95	96	96	96	95	96	96	96
			ICC$_s(C,1)$				ICC$_s(C,k)$			
20	10	2	94	95	94	94	94	95	94	94
		5	95	95	95	95	95	95	95	95
		10	95	95	95	95	95	95	95	95
	30	2	97*	96	97*	96*	97*	96	97*	96*
		5	96	95	96	95	96	95	96	95
		10	95	94	94	94	95	94	94	94
40	10	2	96	96	96	96	96	96	96	96
		5	94	94	95	94	94	94	95	94
		10	94	94	94	94	94	94	94	94
	30	2	96	96	96	96	96	96	96	96
		5	94	95	95	95	94	95	95	95
		10	95	95	95	95	95	95	95	95

Note. ICC = Intraclass correlation; N_c = Number of clusters; N_s = Number of subjects per cluster; K = Number of raters; σ_c^2 = Variance of cluster effects; σ_s^2 = Variance of subject effects; * Outside 95% Agresti-Coull confidence interval.

Table B9: Percentage of 95% BCI Coverage of Cluster-Level ICCs across Conditions

			σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
N_c	$N_{s:c}$	K	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
			$\text{ICC}_c(A,1)$				$\text{ICC}_c(A,k)$			
20	10	2	96	98*	94	97*	96	98*	94	97*
		5	95	96	96	96	95	96	96	96
		10	96	96	96*	96	96	96	96*	96
	30	2	95	98*	94	97*	95	98*	94	97*
		5	95	96	95	96	95	96	95	96
		10	95	95	95	95	95	95	95	95
40	10	2	94	96	95	97*	94	96	95	97*
		5	94	95	95	96*	94	95	95	96*
		10	94	94	95	96	94	94	95	96
	30	2	96	97*	96	98*	96	97*	96	98*
		5	96	97*	96*	97*	96	97*	96*	97*
		10	95	95	95	95	95	95	95	95
			$\text{ICC}_c(C,1)$				$\text{ICC}_c(C,k)$			
20	10	2	97*	96*	94*	94*	97*	96*	94*	94*
		5	95	95	95	95	95	95	95	95
		10	96	96	96	96*	96	96	96	96*
	30	2	96	96	94	95	96	96	94	95
		5	95	95	96	95	95	95	96	95
		10	95	95	95	96	95	95	95	96
40	10	2	95	95	94	94	95	95	94	94
		5	95	94	96	95	95	94	96	95
		10	95	94	95	95	95	94	95	95
	30	2	94	95	94	95	94	95	94	95
		5	95	94	96	96	95	94	96	96
		10	95	95	95	95	95	95	95	95

Note. ICC = Intraclass correlation; N_c = Number of clusters; N_s = Number of subjects per cluster; K = Number of raters; σ_c^2 = Variance of cluster effects; σ_s^2 = Variance of subject effects; * Outside 95% Agresti-Coull confidence interval.

Table B10: Relative Efficiency of Subject-Level ICCs across Conditions

			σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
N_c	N_s	K	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
			ICC$_s(A,1)$				ICC$_s(A,k)$			
20	10	2	1.98	1.51	1.96	1.51	2.15	1.62	2.13	1.62
		5	1.37	1.22	1.37	1.22	1.58	1.40	1.58	1.39
		10	1.13	1.08	1.12	1.08	1.21	1.20	1.20	1.19
	30	2	2.22	1.61	2.22	1.62	2.39	1.72	2.39	1.73
		5	1.37	1.18	1.38	1.18	1.59	1.37	1.59	1.37
		10	1.15	1.07	1.14	1.07	1.24	1.18	1.24	1.18
40	10	2	2.15	1.59	2.13	1.59	2.33	1.71	2.31	1.71
		5	1.44	1.25	1.44	1.25	1.66	1.43	1.66	1.44
		10	1.21	1.14	1.20	1.14	1.30	1.26	1.29	1.26
	30	2	2.24	1.66	2.25	1.66	2.44	1.78	2.45	1.79
		5	1.46	1.24	1.46	1.24	1.69	1.44	1.70	1.43
		10	1.24	1.15	1.24	1.15	1.35	1.28	1.35	1.28
			ICC$_s(C,1)$				ICC$_s(C,k)$			
20	10	2	0.98	0.99	0.98	0.98	0.98	0.99	0.98	0.98
		5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		10	1.01	1.02	1.02	1.02	1.01	1.01	1.01	1.01
	30	2	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05
		5	1.02	1.02	1.02	1.02	1.01	1.01	1.02	1.02
		10	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
40	10	2	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03
		5	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.99
		10	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.97
	30	2	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
		5	1.03	1.03	1.03	1.02	1.03	1.03	1.03	1.02
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. ICC = Intraclass correlation; N_c = Number of clusters; N_s = Number of subjects per cluster; K = Number of raters; σ_c^2 = Variance of cluster effects; σ_r^2 = Variance of rater effects. Relative efficiency was computed as the ratio of the average posterior *SD* relative to the *SD* of the posterior means. Preferably, this ratio equals 1.

Table B11: Relative Efficiency of Cluster-Level ICCs across Conditions

			σ_c^2= 0.16		σ_c^2= 0.50		σ_c^2= 0.16		σ_c^2= 0.50	
N_c	N_s	K	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50	σ_r^2= 0.16	σ_r^2= 0.50
			$\text{ICC}_c(A,1)$				$\text{ICC}_c(A,k)$			
20	10	2	1.38	1.26	1.40	1.24	1.41	1.29	1.42	1.26
		5	1.06	1.02	1.11	1.06	1.09	1.06	1.16	1.12
		10	1.04	1.04	1.04	1.04	1.00	1.01	0.97	1.00
	30	2	1.40	1.26	1.53	1.32	1.43	1.29	1.58	1.37
		5	1.10	1.04	1.15	1.07	1.09	1.06	1.22	1.15
		10	1.03	1.01	1.03	1.02	0.96	0.99	1.01	1.03
40	10	2	1.33	1.18	1.57	1.31	1.37	1.22	1.68	1.39
		5	1.09	1.03	1.20	1.10	1.11	1.07	1.34	1.22
		10	1.01	0.99	1.05	1.04	0.95	0.97	1.06	1.07
	30	2	1.47	1.27	1.78	1.43	1.52	1.32	1.91	1.53
		5	1.13	1.05	1.24	1.11	1.21	1.12	1.42	1.26
		10	1.05	1.01	1.10	1.06	1.06	1.06	1.15	1.13
			$\text{ICC}_c(C,1)$				$\text{ICC}_c(C,k)$			
20	10	2	1.14	1.14	0.96	0.97	1.21	1.21	0.97	0.97
		5	1.00	0.99	0.92	0.93	1.05	1.04	0.88	0.88
		10	0.99	0.99	0.94	0.93	1.00	0.99	0.81	0.77
	30	2	1.10	1.08	1.02	1.02	1.14	1.13	1.03	1.03
		5	0.97	0.97	0.97	0.96	0.90	0.90	0.92	0.92
		10	0.99	0.99	0.96	0.97	0.85	0.84	0.90	0.91
40	10	2	1.03	1.03	0.97	0.98	1.06	1.06	0.96	0.97
		5	0.97	0.96	0.96	0.96	0.93	0.92	0.92	0.91
		10	0.95	0.94	0.97	0.98	0.87	0.84	0.91	0.92
	30	2	1.01	1.00	1.00	1.00	1.02	1.01	1.00	1.00
		5	0.99	0.99	0.99	1.00	0.94	0.94	0.97	0.98
		10	1.01	1.00	1.00	1.00	0.93	0.93	0.97	0.97

Note. ICC = Intraclass correlation; N_c = Number of clusters; N_s = Number of subjects per cluster; K = Number of raters; σ_c^2 = Variance of cluster effects; σ_r^2 = Variance of rater effects. Relative efficiency was computed as the ratio of the average posterior *SD* relative to the *SD* of the posterior means. Preferably, this ratio equals 1.

B.3 Results Simulation 2

Table B12: Relative MAP Bias of ICCs Across Conditions

k_s	K	$\text{ICC}_s(A,1)$	$\text{ICC}_s(A,k)$	$\text{ICC}_s(C,1)$	$\text{ICC}_s(C,k)$
2	5	-0.02	-0.01	0.00	0.00
	10	-0.02	-0.01	0.00	0.00
3	5	-0.02	-0.01	0.00	0.00
	10	-0.01	-0.01	0.00	0.00
k_s	K	$\text{ICC}_c(A,1)$	$\text{ICC}_c(A,k)$	$\text{ICC}_c(C,1)$	$\text{ICC}_c(C,k)$
2	5	-0.29	-0.22	-0.08	-0.05
	10	-0.26	-0.20	-0.06	-0.03
3	5	-0.21	-0.12	-0.04	-0.01
	10	-0.21	-0.12	-0.04	-0.01

Note. ICC = Intraclass correlation; k_s = Number of raters per subject; K = Total number of raters (size of the rater pool). Relative bias was computed as $\frac{\bar{\theta}-\theta}{\theta}$, where $\bar{\theta}$ denotes the average MAP estimate of a parameter (or derived ICC) across replications in a condition, and θ denotes the population parameter in that condition.

Table B13: Percentage of 95% BCI Coverage of ICCs Across Conditions

k_s	K	$\text{ICC}_s(A,1)$	$\text{ICC}_s(A,k)$	$\text{ICC}_s(C,1)$	$\text{ICC}_s(C,k)$
2	5	95	95	96	96
	10	95	95	96	96
3	5	95	95	96	96
	10	95	95	96	96
k_s	K	$\text{ICC}_c(A,1)$	$\text{ICC}_c(A,k)$	$\text{ICC}_c(C,1)$	$\text{ICC}_c(C,k)$
2	5	96	96	95	95
	10	95	95	95	95
3	5	95	95	94	94
	10	95	95	96	96

Note. ICC = Intraclass correlation; k_s = Number of raters per subject; K = Total number of raters (size of the rater pool).

Table B14: Relative Efficiency of ICCs Across Conditions

k_s	K	$\text{ICC}_s(A,1)$	$\text{ICC}_s(A,k)$	$\text{ICC}_s(C,1)$	$\text{ICC}_s(C,k)$
2	5	1.29	1.36	1.03	1.04
	10	1.29	1.36	1.03	1.03
3	5	1.37	1.50	1.04	1.03
	10	1.32	1.44	1.00	1.00
k_s	K	$\text{ICC}_c(A,1)$	$\text{ICC}_c(A,k)$	$\text{ICC}_c(C,1)$	$\text{ICC}_c(C,k)$
2	5	1.09	1.11	1.04	1.07
	10	1.10	1.11	1.04	1.06
3	5	1.06	1.05	0.97	0.96
	10	1.07	1.08	1.02	1.02

Note. ICC = Intraclass correlation; k_s = Number of raters per subject; K = Total number of raters (size of the rater pool). Relative efficiency was computed as the ratio of the average posterior *SD* relative to the *SD* of the posterior means. Preferably, this ratio equals 1.

B.4 Maximum Likelihood Estimates and additional MCMC Estimates Application

Table B15: Variance Decomposition as Estimated with MLE

	Conflated	MLM
Student	1.63	1.40
Teacher	–	0.23
Rater	0.26	0.26
Student × Rater	0.48	0.44
Teacher × Rater	–	0.03

Note. MLM = Multilevel Model.

Table B16: IRR coefficients as Estimated with MCMC

	ICC(A,1)		ICC(A,k)		ICC(C,1)		ICC(C,k)	
	MAP	BCI	MAP	BCI	MAP	BCI	MAP	BCI
Conflated ($k = 3$)	.68	[.49, .75]	.87	[.74, .90]	.77	[.73, .81]	.91	[.89, .93]
Student Level ($k = 3$)	.66	[.46, .74]	.85	[.72, .89]	.76	[.71, .81]	.91	[.88, .93]
Teacher Level ($k = 3$)	.39	[.10, .68]	.71	[.24, .86]	.90	[.62, .96]	.96	[.83, .98]
Teacher Level ($k = 5$)	–	–	.80	[.35, .91]	–	–	.98	[.89, .99]

Note. ICC = Intraclass correlation; MAP = maximum a posteriori; BCI = Bayesian credible interval; k = number of raters.

Table B17: IRR coefficients as Estimated with MLE

	ICC(A,1)	ICC(A,k)	ICC(C,1)	ICC(C,k)
Conflated ($k = 3$)	.69	.87	.77	.91
Student Level ($k = 3$)	.67	.86	.76	.90
Teacher Level ($k = 3$)	.45	.71	.88	.96
Teacher Level ($k = 5$)	–	.80	–	.97

Note. ICC = Intraclass correlation; k = number of raters.

References

AERA, APA, & NCME. (2018). *Standards for educational and psychological testing*. American Educational Research Association.

Agresti, A. (2010). *Analysis of ordinal categorical data*. Wiley.

Agresti, A. (2018). *An introduction to categorical data analysis* (3rd ed.). Wiley.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. http://doi.org/https://doi.org/10.1080/00031305.1998.10480550

Aitchison, J. A. (1986/2003). *The statistical analysis of compositional data*. Chapman & Hall/Blackburn Press.

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework* (Doctoral dissertation). https://open.library.ubc.ca

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37. http://doi.org/10.1016/j.jsp.2009.10.001

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. http://doi.org/10.2466/pr0.1966.19.1.3

Bartko, J. J. (1974). Corrective note to: "The intraclass correlation coefficient as a measure of reliability.". *Psychological Reports*, *34*(2), 418. http://doi.org/10.2466/pr0.1974.34.2.418

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using `lme4`. *Journal of Statistical Software*, *67*, 1–48. http://doi.org/10.18637/jss.v067.i01

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association*, *61*(1), 228–235. http://doi.org/10.2307/2283057

Bijlsma, H. J. E., Glas, C. A. W., & Visscher, A. J. (2022). Factors related to differences in digitally measured student perceptions of teaching quality. *School Effectiveness and School Improvement*, *33*(3), 1–21. http://doi.org/10.1080/09243453.2021.2023584

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: Amee guide no. 68. *Medical Teacher*, *34*(11), 960–992. http://doi.org/10.3109/0142159X.2012.703791

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26*(4), 364–375. http://doi.org/10.1177/014662102237794

Bodner, N., Tuerlinckx, F., Bosmans, G., & Ceulemans, E. (2021). Accounting for auto-dependency in binary dyadic time series data: A comparison of model- and permutation-based approaches for testing pairwise associations. *British Journal of Mathematical and Statistical Psychology*, *74*(1), 86-109. https://doi.org/10.1111/bmsp.12222

Brennan, R. L. (2001a). *Generalizability theory*. Springer.

Brennan, R. L. (2001b). Manual for `urGENOVA` version 2.1 [Computer software manual]. https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs The University of Iowa.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. http://doi.org/10.1016/j.jsp.2013.11.008

Bürkner, P.-C. (2017). `brms`: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. http://doi.org/10.18637/jss.v080.i01

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*, *19*(3), 447–469. http://doi.org/10.1111/j.1467-9507.2009.00562.x

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. http://doi.org/10.18637/jss.v076.i01

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., ... Borges, B. (2022). shiny: Web application framework for r *[Computer Software]* [Computer

software manual]. Retrieved from https://CRAN.R-project.org/package=shiny (R package version 1.7.3)

Chen, M., Zee, M., & Roorda, D. L. (2022). Assessing student–teacher relationship quality in cross-cultural contexts: Psychometric properties of student–teacher relationship drawings. *European Journal of Developmental Psychology*, *19*(5), 770–784. http://doi.org/10.1080/17405629.2021.1952862

Cho, S.-J., & Preacher, K. J. (2016). Measurement error correction formula for cluster-level group differences in cluster randomized and observational studies. *Educational and Psychological Measurement*, *76*(5), 771–786. http://doi.org/10.1177/0013164415612255

Cho, S.-J., Shen, J., & Naveiras, M. (2019). Multilevel reliability measures of latent scores within an item response theory framework. *Multivariate Behavioral Research*, *54*(6), 856–881. http://doi.org/10.1080/00273171.2019.1596780

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. http://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. http://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. http://doi.org/10.1037/h0026256

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. http://doi.org/10.1037/0012-1649.35.5.1179

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322-–328. http://doi.org/10.1037/0033-2909.88.2.322

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. http://doi.org/10.1111/j.2044-8317.1963.tb00206.x

De Leeuw, J., Van der Heijden, P. G. M., & Verboon, P. (1990). A latent time–budget model. *Statistica Neerlandica*, *44*(1), 1–22. http://doi.org/10.1111/j.1467-9574.1990.tb01268.x

De Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*(1), 45–49. http://doi.org/10.1016/j.jclinepi.2017.01.013

Eliasziw, M., Young, S. L., Woodbury, M. G., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. *Physical Therapy*, *74*(8), 777—788. http://doi.org/10.1093/ptj/74.8.777

Faber, J. M., Glas, C. A. W., & Visscher, A. J. (2018). Differentiated instruction in a data-based decision-making context. *School Effectiveness and School Improvement*, *29*(1), 43–63. http://doi.org/10.1080/09243453.2017.1366342

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology*, *11*(1), 13–22. http://doi.org/10.1027/1614-2241/a000086

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, *30*(1), 71—76. http://doi.org/10.1177/001316447003000106

Fisher, R. A. (1954). *Statistical methods for research workers* (12th ed.). Oliver and Boyd.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. http://doi.org/10.1037/h0031619

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third ed.). Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fürst, G. (2020). Measuring creativity with planned missing data. *The Journal of Creative Behavior*, *54*(1), 150–164. http://doi.org/10.1002/jocb.352

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,

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. http://doi.org/10.1037/a0032138

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://projecteuclid.org/euclid.ba/1340371048

Gelman, A. (2019). *Prior choice recommendations.* https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. http://doi.org/10.1214/ss/1177011136

Goble, P., Sandilos, L. E., & Pianta, R. C. (2019). Gains in teacher-child interaction quality and children's school readiness skills: Does it matter where teachers start? *Journal of School Psychology*, *73*, 101–113. http://doi.org/10.1016/j.jsp.2019.03.006

Graham, J. W., Hofer, S. M., & 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. http://doi.org/10.1207/s15327906mbr3102_3

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. http://doi.org/10.1037/1082-989X.11.4.323

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. http://doi.org/10.20982/tqmp.08.1.p023

Harmsen, R., Helms-Lorenz, M., Maulana, R., & Van Veen, K. (2018). The relationship between beginning teachers' stress causes, stress responses, teaching behaviour and

attrition. *Teachers and Teaching*, *24*(6), 626–643. http://doi.org/10.1080/13540602.2018.1465404

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. http://doi.org/10.1080/19312450709336664

Helms-Lorenz, M., van de Grift, W., & Maulana, R. (2016). Longitudinal effects of induction on teaching skills and attrition rates of beginning teachers. *School Effectiveness and School Improvement*, *27*(2), 178–204. http://doi.org/10.1080/09243453.2015.1035731

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. http://doi.org/10.3758/s13423-013-0572-3

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, *100*(469), 286–295. http://doi.org/10.1198/016214504000001015

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*, 1593–1623. https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf

Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer. http://doi.org/10.1007/978-3-642-72087-1_17

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. http://doi.org/10.1037/pspi0000097

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, *24*(1), article 5. http://doi.org/10.7275/5065-gc10

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. http://doi.org/10.1037/pspp0000369

IBM Corp. (2021). *IBM SPSS Statistics for Windows (version 28.0)* [Computer Software]. https://www.ibm.com/products/spss-statistics

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*(2), 277–289. http://doi.org/10.1177/00131640121971239

JASP Team. (2022). *JASP (Version 0.16.4)* [Computer Software]. https://jasp-stats.org/

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R. *Methodology*, *14*(3), 133–142. http://doi.org/10.1027/1614-2241/A000149

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. http://doi.org/10.3390/psych3020011

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *`semTools`: Useful tools for structural equation modeling.* https://CRAN.R-project.org/package=semTools (R package version 0.5-5)

Jorgensen, T. D., Rhemtulla, M., Schoemann, A. M., McPherson, B., Wu, W., & Little, T. D. (2014). Optimal assignment methods in three-form planned missing data designs for longitudinal panel studies. *International Journal of Behavioral Development*, *38*(5), 397–410. http://doi.org/10.1177/0165025414531094

Jorgensen, T. D., Van der Ark, L. A., & Ten Hove, D. (in press). Factors affecting efficiency of interrater reliability estimates from planned missing data designs. In M. Wiberg, J. González, D. Molenaar, H. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 86th annual meeting of the Psychometric Society, Bologna, Italy, 2022.* Springer.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136–153. http://doi.org/10.1080/10705510701758406

Kendall, M. G. (1948). *Rank correlation methods.* Griffin.

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. http://doi.org/10.1177/0265407596132007

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* (Vol. 18, pp. 142–182). Academic Press. http://doi.org/10.1016/S0065-2601(08)60144-6

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*(1), 126–137. http://doi.org/10.1037//0022-3514.83.1.126

Ketokivi, M. (2019). Avoiding bias and fallacy in survey research: A behavioral multilevel approach. *Journal of Operations Management*, *65*(4), 380–402. http://doi.org/10.1002/joom.1011

Kivisalu, T. M., Lewey, J. H., Shaffer, T. W., & Canfield, M. L. (2016). An investigation of interrater reliability for the rorschach performance assessment system (R–PAS) in a nonpatient U.S. sample. *Journal of Personality Assessment*, *98*(4), 382–390. http://doi.org/10.1080/00223891.2015.1118380

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. http://doi.org/10.1016/j.jcm.2016.02.012

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2021). Range-preserving confidence intervals and significance tests for scalability coefficients in Mokken scale analysis. In M. Wiberg, D. Molenaar, J. González, H. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology* (pp. 175–185). Springer. http://doi.org/10.1007/978-3-030-74772-5_16

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology*, *64*(6), 701–702. http://doi.org/10.1016/j.jclinepi.2010.12.001

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, *2*(1), 139–150. http://doi.org/10.2307/270787

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. sage.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology*, *12*(4), 139–144. http://doi.org/10.1027/1614-2241/a000119

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752. http://doi.org/10.1177/1094428112457829

Lai, M. H. C. (2020). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, *26*(1), 90–102. http://doi.org/10.1037/met0000287

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. http://doi.org/10.2307/2529310

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365–377. http://doi.org/10.1037/h0031643

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley Online Library.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2013). On the joys of missing data. *Journal of Pediatric Psychology*, *39*(2), 151–162. http://doi.org/10.1093/jpepsy/jst048

LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management*, *41*(2), 692–717. http://doi.org/10.1177/0149206314554215

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444–467. http://doi.org/10.1037/a0024376

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. http://doi.org/10.1037/a0029252

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99–128. http://doi.org/10.1207/s15327906mbr3901_4

Majdandžić, M., de Vente, W., Möller, E. L., & Bögels, S. M. (2021). Severity of fathers' and mothers' anxiety disorders predicts their observed and self-rated parenting behavior.
(In preparation)

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. http://doi.org/10.1111/infa.12125

Malloy, T. E., & Kenny, D. A. (1986). The social relations model: An integrative method for personality research. *Journal of Personality*, *54*(1), 199–225. http://doi.org/10.1111/j.1467-6494.1986.tb00393.x

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, *66*(2), 379–386. http://doi.org/10.2466/pr0.1990.66.2.379

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(3), 290–299. http://doi.org/10.1080/10705519609540045

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, *116*(535), 651-–655. http://doi.org/10.1192/bjp.116.535.651

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. http://doi.org/10.1037/1082-989X.1.1.30

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314. http://doi.org/10.1007/s10648-014-9287-x

Meredith, M., & Kruschke, J. (2018). *HDInterval: Highest (posterior) density intervals* [Computer Software]. https://CRAN.R-project.org/package=HDInterval

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*(1), 129–149. http://doi.org/10.1207/s15327906mbr3901_5

Molenaar, D., Uluman, M., Tavşancıl, E., & De Boeck, P. (2021). The hierarchical rater thresholds model for multiple raters and multiple items. *Open Education Studies*, *3*(1), 33–48. http://doi.org/10.1515/edu-2020-0105

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. http://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. http://doi.org/10.1007/s11336-020-09728-z

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*(1), 27–29. http://doi.org/10.1080/00031305.1992.10475842

Pearson, K. (1985). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(1), 240—242. http://www.jstor.org/stable/115794

Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E.-J. (2022). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*, *57*(4), 620–641. http://doi.org/10.1080/00273171.2021.1891855

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902. http://doi.org/10.1214/12-BA730

Poncet, P. (2019). `modeest`: *Mode estimation* [Computer Software]. https://CRAN.R-project.org/package=modeest

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research* (pp. 90–105). Springer. http://doi.org/10.1007/978-1-349-19051-5_6

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77–98. http://doi.org/10.1080/19312458.2012.679848

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. http://doi.org/10.1037/a0020141

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959–981. http://doi.org/10.1037/0021-9010.93.5.959

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer Software]. https://www.R-project.org/ R Foundation for Statistical Computing.

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer Software]. https://www.R-project.org/ Vienna, Austria.

Rajaratnam, N. (1960). Reliability formulas for independent decision data when reliability data are matched. *Psychometrika*, *25*(3), 261–271. http://doi.org/10.1007/BF02289730

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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. http://doi.org/10.1037/pas0000754

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. http://doi.org/10.1037/a0029315

Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review*, *22*(1), 17—25. http://www.jstor.org/stable/2088760

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5:589965*. http://doi.org/10.3389/feduc.2020.589965

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://doi.org/10.18637/jss.v048.i02

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling*, *11*(3), 424–451. http://doi.org/10.1207/s15328007sem1103_7

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. http://doi.org/10.1177/0956797617727121

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147. http://doi.org/10.1037%2F1082-989X.7.2.147

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932. http://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. http://doi.org/10.1037/0033-2909.86.2.420

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. http://doi.org/10.1177/0956797611417632

Simpkins, S. D., & Parke, R. D. (2002). Do friends and nonfriends behave differently? A social relations analysis of children's behavior. *Merrill-Palmer Quarterly*, *48*(3), 263–283. https://www.jstor.org/stable/23093770

Smid, S. C., McNeish, D., Miočević, M., & Van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161. http://doi.org/10.1080/10705511.2019.1577140

Snijders, T. A. B., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, *6*(4), 471–486. http://doi.org/10.1111/j.1475-6811.1999.tb00204.x

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72–101. http://doi.org/10.2307/1412159

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation* (Vol. 13). Wiley.

Stan Development Team. (2018). *rstan: The R interface to Stan* [Computer Software]. https://mc-stan.org/users/interfaces/rstan.html

Stan Development Team. (2020). *RStan: the R interface to Stan.* http://mc-stan.org/ (R package version 2.21.2)

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. http://doi.org/10.3102/1076998616646200

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika*, *40*(1), 105–110. http://doi.org/10.2307/2333101

Surfsara. (n.d.). *Lisa computer cluster.* https://www.surf.nl/files/2019-03/lisa-compute-cluster.pdf

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). *Interrater reliability for dyad-level predictors in network data.* (Paper presented at the XXXVIII Sunbelt 2018 Conference, Utrecht, the Netherlands)

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). On the usefulness of interrater reliability coefficients. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 82th annual meeting of the Psychometric Society, Zurich, Switzerland, 2019.* (pp. 67–75). Springer. http://doi.org/10.1007/978-3-319-77249-3_6

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, J. González, D. Molenaar, H. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019.* (pp. 79–93). Springer. http://doi.org/10.1007/978-3-030-43469-4_7

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2022a). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, *27*(4), 650–666. http://doi.org/10.1037/met0000391

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2022b). *Supplementary materials to 'updated guidelines on selecting an ICC for interrater reliability'.* http://doi.org/10.17605/OSF.IO/8J26U

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2022c). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*. http://doi.org/10.1037/met0000516 (Advanced online publication)

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (Under Reviewa). Estimating intraclass correlation coefficients for interrater reliability from incomplete observational designs: A simulation study and a tutorial.

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (Under Reviewb). Interrater reliability for interdependent network data: A generalizability theory approach.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Vanbelle, S. (2017). Comparing dependent kappa coefficients obtained on multilevel data. *Biometrical Journal*, *59*(5), 1016–1034. http://doi.org/10.1002/bimj.201600093

Vanbelle, S., Mutsvari, T., Declerck, D., & Lesaffre, E. (2012). Hierarchical modeling of agreement. *Statistics in Medicine*, *31*(5), 3667–3680. http://doi.org/10.1002/sim.5424

Van der Put, C. E., Deković, M., Stams, G. J. J. M., Van der Laan, P. H., Hoeve, M., & Van Amelsfort, L. (2011). Changes in risk factors during adolescence: Implications for risk assessment. *Criminal Justice and Behavior*, *38*(3), 248–262. http://doi.org/10.1177/0093854810391757

Van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, *30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

Van Duijn, M. A. J., Snijders, T. A. B., & Zijlstra, B. J. H. (2004). p2: A random effects model with covariates for directed graphs. *Statistica Neerlandica*, *58*(2), 234–254. http://doi.org/10.1046/j.0039-0402.2003.00258.x

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. http://doi.org/10.1016/j.jmp.2018.12.004

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. http://doi.org/10.1111/j.0006-341X.2005.031040.x

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018a). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. http://doi.org/10.1037/met0000107

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018b). Using generalizability theory to disattenuate correlation coefficients for multiple sources of measurement error. *Multivariate Behavioral Research*, *53*(2), 481–501. http://doi.org/10.1080/00273171.2018.1457938

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods*, *24*(2), 153–178. http://doi.org/10.1037/met0000177

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*(1), 108–131. http://doi.org/10.1037/0021-9010.90.1.108

Wang, T., & Merkle, E. C. (2018). `merDeriv`: Derivative computations for linear mixed effects models with application to robust standard errors. *Journal of Statistical Software*, *87*(1), 1–16. http://doi.org/10.18637/jss.v087.c01

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. http://doi.org/10.1037/0022-3514.37.10.1742

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. S. Rao & S. Sinharay (Eds.), *Handbook of statistics, volume 26, psychometrics* (pp. 81–124). Elsevier. http://doi.org/10.1016/S0169-7161(06)26004-8

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. http://doi.org/10.1002/sim.6113

Yuen, J. K., Kelley, A. S., Gelfman, L. P., Lindenberger, E. E., Smith, C. B., Arnold, R. M., ... 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. http://doi.org/10.1016/j.jpainsymman.2019.09.001

Zee, M., & Roorda, D. L. (2017). *Leerkracht-leerling relatietekeningen in het basisonderwijs. trainings- en coderingshandleiding [student–teacher relationship drawings in elementary school. training and coding manual].* Amsterdam, the Netherlands: University of Amsterdam.

Zee, M., Rudasill, K. M., & 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. http://doi.org/10.1086/708661

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association*, *36*(1), 419–480. http://doi.org/10.1080/23808985.2013.11679142

Zijlstra, B. J. H. (2017). Regression of directed graphs on independent effects for density and reciprocity. *The Journal of Mathematical Sociology*, *41*(4), 185–192. http://doi.org/10.1080/0022250X.2017.1387858

Summary

The central question of this dissertation was how to estimate the interrater reliability (IRR) from incomplete and dependent observational data. This topic is important 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. In observational research, information about attributes is obtained through external raters. In practice, scores based on observations are used to make decisions about subjects. 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. 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. IRR expresses the degree to which these observed ratings are independent of raters, and is imperative for observational practice and research. It provides information about the ability to differentiate between subjects based on ratings, and bounds the precision and validity of ratings.

The standard observational design, for which most IRR coefficients are defined, is a complete (two-way) design. In such a design, multiple independent subjects are each rated by multiple independent raters, and these raters are the same for each subject. However, in social and behavioural observational research, the raters are often not the same for each subject, resulting in incomplete data. Also, the rated subjects are often nested within clusters or relationships, resulting in dependent data. Dependent data have various facets that may be of theoretical interest to researchers. For example, multilevel data contain subject and cluster-level components, and interdependent network data contain actor, partner, and relationship components about which researchers may formulate research questions.

None of the conceptualizations of IRR can be readily used for both incomplete and dependent data. Ignoring the incompleteness or the distinct facets of theoretical interest in dependent data yields IRR estimates that are of limited value. For example, to investigate the degree to which the observed subject and cluster scores in multilevel data, or the actor, partner, and relationship effects in social network data are independent of raters, IRR should be estimated for each of these facets of interest separately. Also, estimation

methods were needed to accommodate missing observations and dependence structures in dependent data.

Using the framework of Generalizability theory (GT), this dissertation provides definitions and estimation methods of IRR for incomplete and dependent data. In **Chapter 1**, I introduced the topics of IRR, incomplete data, and multilevel data, and discussed the need for new methods. In **Chapter 2**, I illustrated and discussed the issue of an abundance of IRR coefficients that follow from different conceptualizations of IRR. In **Chapter 3**, I explained why the intraclass correlation coefficients (ICCs) are probably the best candidate for defining IRR, and I used GT to extend the different definitions of ICCs to incomplete observational designs and to provide updated guidelines on when to use which ICC definition. ICCs are traditionally estimated with ANOVA-based methods, which are not straightforward for incomplete or dependent data. Markov chain Monte Carlo (MCMC) estimation of hierarchical linear models can handle such data, but requires the definition of hyperprior distributions. In **Chapter 4**, I investigated the effect of different hyperprior distributions on ICC estimates under different conditions that are common to observational studies. In **Chapter 5**, I described a simulation study in which I compared the MCMC approach to estimating ICCs from incomplete data with two maximum likelihood approaches. In **Chapter 6**, I generalized the definitions and estimation methods for IRR to multilevel data, and I evaluated the IRR estimates in a simulation study. In **Chapter 7**, I generalized the definitions and estimation methods of IRR to the more complex case of interdependent network data, and I evaluated the IRR estimates in a simulation study. In **Chapter 8**, I discussed the implications of this dissertation for observational studies, and suggested directions for future IRR research. I argue that ICCs are very useful IRR estimators, especially when scores based on observations are used for relative purposes. I therefore highly recommend adopting ICCs for interrater consistency as a standard measure for IRR.

This dissertation not only provides definitions and estimation methods of IRR for incomplete and nested data, it also guides researchers in using these methods. Throughout the different chapters, I illustrated the methods with empirical data from clinical and developmental domains, using free software that I provide on the Open Science Framework. Also, I provide advice about planning observational studies. Ultimately, these methods reach a wide range of researchers, to eventually lead to more reliable observational research.

Summary in Dutch/Samenvatting

De centrale vraag van dit proefschrift was hoe de interbeoordelaarsbetrouwbaarheid (IBB) kan worden geschat uit onvolledige en afhankelijke observationele data. Dit onderwerp is van belang voor observatie-studies waarin kenmerken van subjecten gemeten worden. Voorbeelden van dergelijke kenmerken zijn leesvaardigheid, didactische vaardigheden, spelvaardigheid, opvoedingsgedrag en academische vaardigheden. Omdat deze kenmerken niet direct waarneembaar zijn, gebruiken professionals en wetenschappers methoden zoals zelfrapportage vragenlijsten of observaties om informatie over deze kenmerken te verkrijgen. Bij observaties wordt informatie over kenmerken verkregen door externe beoordelaars. In de praktijk worden scores op basis van observaties gebruikt om beslissingen te nemen over subjecten. In onderzoek worden scores op basis van observaties gebruikt om de relatie tussen verschillende kenmerken te onderzoeken, of om het effect van externe criteria op de kenmerken van subjecten te onderzoeken. Idealiter is de variatie in beoordelingen van kenmerken van subjecten dus afkomstig van verschillen tussen deze subjecten, en zo min mogelijk van verschillen tussen de beoordelaars die de subjecten beoordelen. De IBB drukt uit in welke mate de waargenomen beoordelingen onafhankelijk zijn van de beoordelaars, en is van essentieel belang voor de observatiepraktijk en wetenschappelijk onderzoek. IBB geeft informatie over het vermogen om onderscheid te maken tussen subjecten op basis van beoordelingen, en begrenst de precisie en de validiteit van de beoordelingen.

Het standaard observationele onderzoeksdesign, waarvoor de meeste IBB-coëfficiënten zijn gedefinieerd, is een compleet (tweeweg) design. In een dergelijke onderzoeksdesign worden verschillende onafhankelijke subjecten beoordeeld door verschillende onafhankelijke beoordelaars, en deze beoordelaars zijn voor elke persoon dezelfde. In observationeel sociaal- en gedragswetenschappelijk onderzoek zijn de beoordelaars echter vaak niet voor elke persoon dezelfde, waardoor de data onvolledig zijn. Ook zijn de beoordeelde subjecten vaak genest binnen clusters of relaties, waardoor afhankelijke data ontstaan. Afhankelijke data hebben verschillende facetten die van theoretisch belang kunnen zijn voor onderzoekers. Zo bevatten multilevel-data componenten op persoon- en clusterniveau, en bevatten wederzijdsafhankelijke netwerkdata actor-, partner- en relatiecomponenten waarover onderzoekers onderzoeksvragen formuleren.

Geen van de conceptualiseringen van IBB kan worden gebruikt voor zowel onvolledige

als afhankelijke data. Het negeren van de onvolledigheid van data of de verschillende theoretisch belangrijke facetten in afhankelijke data levert IBB-schattingen op die van beperkte waarde zijn. Om bijvoorbeeld te onderzoeken in hoeverre de waargenomen persoon- en clusterscores in multilevel data, of de actor-, partner- en relatie-effecten in sociale netwerkdata onafhankelijk zijn van beoordelaars, moet de IBB voor elk van deze facetten afzonderlijk worden geschat. Ook waren schattingsmethoden nodig die om kunnen gaan met ontbrekende data en afhankelijkheidsstructuren in afhankelijke data.

Gebruikmakend van de generalizeerbaarheidstheorie (GT) geeft dit proefschrift definities en schattingsmethoden voor de IBB voor onvolledige en afhankelijke data. In **Hoofdstuk 1** introduceer ik de onderwerpen IBB, incomplete data en multilevel data, en bespreek ik de behoefte aan nieuwe methoden. In **Hoofdstuk 2** illustreer en bespreek ik het probleem van een overvloed aan IBB-coëfficiënten die voortvloeien uit verschillende conceptualiseringen van IBB. In **Hoofdstuk 3** leg ik uit waarom de intraklasse correlatiecoëfficiënten (ICCs) waarschijnlijk de beste kandidaat zijn om IBB te definiëren, en gebruik ik GT om de verschillende definities van ICCs uit te breiden tot onvolledige observationele designs, en om actuele richtlijnen te geven over wanneer welke ICC-definitie moet worden gebruikt. ICCs worden traditioneel geschat met op variantieanalyse gebaseerde methoden, die niet direct toepasbaar zijn voor onvolledige of afhankelijke data. Markov chain Monte Carlo (MCMC) schatting van hiërarchische lineaire modellen is wel toepasbaar, maar vereist de definitie van hyperprior verdelingen. In **Hoofdstuk 4** onderzoek ik het effect van verschillende hyperprior verdelingen op ICC schattingen onder verschillende condities die gebruikelijk zijn bij observationele studies. In **Hoofdstuk 5** beschrijf ik een simulatiestudie waarin ik de MCMC-methode voor het schatten van ICCs op basis van onvolledige data vergeleek met twee op maximum likelihood gebaseerde schattingsmethoden. In **hoofdstuk 6** generaliseer ik de definities en schattingsmethoden voor IBB naar multilevel data en evalueer ik de IBB-schattingen voor multilevel data in een simulatiestudie. In **Hoofdstuk 7** generaliseer ik de definities en schattingsmethoden voor IBB naar onderling afhankelijke netwerkdata, en evalueer ik de IBB-schattingen in een simulatiestudie. In **Hoofdstuk 8** bespreek ik de implicaties van dit proefschrift voor observationele studies, en doe ik suggesties voor toekomstig IBB-onderzoek. In dit hoofdstuk betoog ik dat ICCs uitermate bruikbare IBB-schatters zijn, vooral wanneer observationele data wordt gebruikt voor relatieve doeleinden. Ik beveel de ICCs voor interbeoordelaarsconsistentie daarom van harte aan als standaardmaat voor IBB.

Mijn proefschrift resulteerde niet alleen in definities en schattingsmethoden van IBB voor onvolledige en afhankelijke data, maar begeleidt onderzoekers ook in het gebruik van deze methoden. Door de hoofdstukken heen worden de methoden geïllustreerd met empirische data uit de klinische- en ontwikkelingspedagogiek, met behulp van gratis software die ik beschikbaar stel op het Open Science Framework. Ook geef ik in de verschillende hoofdstukken advies voor het plannen van observationele studies. Hopelijk bereiken deze methoden een breed scala aan onderzoekers, om uiteindelijk te leiden tot betrouwbaarder observationeel onderzoek.

Publications

Chapter 2 is published as:

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). On the usefulness of interrater reliability coefficients. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 82nd annual meeting of the Psychometric Society, Zurich, Switzerland, 2017* (pp. 67–75). Springer. http://doi.org/10.1007/978-3-319-77249-3_6

DtH, TDJ, and LAvdA designed the study, DtH conducted the analyses, wrote the draft, and prepared the manuscript for submission. TDJ and LAvdA provided feedback on the manuscript.

Chapter 3 is published as:

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2022). Updated guidelines on selecting an ICC for interrater reliability, with applications to incomplete observational designs. *Psychological Methods.* [Advanced Online Publication] http://doi.org/10.1037/met0000516

DtH designed the study, wrote the first draft, revised it to an article, and prepared it for submission. TDJ and LAvdA provided feedback on the manuscript.

Chapter 4 is published as:

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, J. González, D. Molenaar, H. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019 (pp. 79–93).* NY: Springer. http://doi.org/10.1007/978-3-030-43469-4_7

DtH designed the study, programmed the simulation, wrote the first draft, revised it to an article, and prepared it for submission. TDJ and LAvdA provided feedback on the simulation design and the manuscript.

Chapter 5 is under review as:

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

DtH designed the study, programmed the simulation, conducted the analyses, wrote the first draft, revised it to an article, and prepared it for submission. TDJ and LAvdA provided feedback on the design and the manuscript.

Chapter 6 is published as:

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. http://doi.org/10.1037/met0000391

DtH designed the study, programmed the simulation, conducted the analyses, wrote the first draft, revised it to an article, and prepared it for submission. TDJ and LAvdA provided feedback on the design and the manuscript.

Chapter 7 is under review as:

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

DtH, TDJ, and LAvdA conceptualized the study. DtH designed and programmed the simulation, conducted the analyses, wrote the first draft, revised it to an article, and prepared it for submission. TDJ and LAvdA provided feedback on the design and the manuscript.

Acknowledgements/Dankwoord

De laatste alinea's van dit boekwerk wijd ik aan alle mensen die van onschatbare waarde zijn geweest voor de totstandkoming van dit proefschrift.

Andries and Terrence, I am glad to have you as my (co-)promotors, and proud to call you my academic parents. Thank you for making me enthusiastic about interrater reliability, and starting this journey with me. You guided my PhD trajectory phenomenally. So well, in fact, that it seemed to go without obstacles. You had an eye for the product, but above all for me as a researcher, teacher and human being. Your feedback took my work to the next level, and I learned incredibly much from your psychometric knowledge. You gave me space to develop in my own way, both inside and outside my dissertation project, while always being there for me with advice and encouragement. Thank you for your guidance!

Leden van de promotiecommissie, ik voel mij vereerd dat u als diverse groep experts bereid was om dit boekwerk te beoordelen en dat u mij hier op 21 april over wilt bevragen. Ik kijk ernaar uit!

Ik ben onwijs dankbaar voor de academische omgeving waarin ik dit proefschrift mocht schrijven. Letty, Kees-Jan, Hannelies, Laura, Elisa en Judith, dank voor de gezelligheid op onze kamer, het vieren van kleine en grote successen, het meeleven in pittige tijden, de bergen etenswaren en kannen sangria, het sparren over methodologische vraagstukken en het delen van ervaringen. Collega's van Methoden en Technieken, dank voor alle inspirerende praatjes, de fijne samenwerkingen in onderwijs en onderzoek, de gezellige congressen en alle adviezen. Mengdi, Sophia, and Tania, it's been six and a half years since we started at the UvA. I am so glad that we are still walking this academic road together, albeit from a distance and each in our own subfields. Elise, we hebben nooit samengewerkt en toch voelt het alsof je mijn hele project met mij mee liep (soms zelf letterlijk). Dank je wel voor het delen van je enthousiasme over onderzoek *en* onderwijs, je luisterend oor, je goede adviezen en je gezelligheid! Andere collega's aan de UvA, van POW en hen die ik daarbuiten leerde kennen via de promovendiraad: jullie visies op onderwijs en onderzoek, manieren van werken en gezelligheid hebben mijn promotietraject enorm verrijkt. Studenten die ik tijdens mijn promotietraject heb lesgegeven en begeleid, bedankt voor de welkome afwisseling van het schrijven van een proefschrift, jullie kritische vragen en leergierigheid.

Collega-onderzoekers die aanklopten met vragen over interbeoordelaarsbetrouw-

baarheid, jullie vragen motiveerden en inspireerden mij om door te werken aan dit onderwerp. Dank hiervoor! Daarnaast waren er mensen die mij uitdaagden en inspireerden door mij mee te nemen in hun onderzoeksprojecten. Door deze 'afleidingen' verbreedde mijn blikveld en bleven de academische wereld en het schrijven van dit proefschrift interessant. Kees-Jan, Tasos, Hannelies, Lennert, Edita, Riet en Han, dank dat jullie mij meenamen naar de wereld van bifactor modellen en intelligentieonderzoek. Ik ben niet alleen trots op de studies die dit heeft opgeleverd, maar vooral ook blij met alle praktische en theoretische lessen die ik leer tijdens onze samenwerking. Terrence en Suzanne, dank dat ik met jullie mocht meedenken over het meten en modelleren van multilevel constructen. Cees, door onze samenwerking kwam ik in aanraking met de wereld van het onderwijskundig meten buiten de universiteit. Dank hiervoor! Dit was en blijft een welkome verademing naast het werk binnen de muren van de universiteit. Nieuwe collega's aan de VU, bedankt voor het warme welkom! De vele decemberborrels hebben de overstap van de UvA naar de VU zeker vergemakkelijkt. Ik ben blij met alle samenwerkingen die al zijn opgestart in onderwijs en onderzoek.

Lieve vrienden, familie en schoonfamilie, dank dat jullie het leven kleur geven! Ik voel mij bevoorrecht met zoveel gezellige en lieve mensen om mij heen. Ik voel mij gesteund door al jullie aanmoedigingen en ik geniet en ontspan tijdens onze uitgebreide etentjes, borrels, wandeltochten, sportsessies en spelletjesavonden. Thom, speciale dank aan jou voor de prachtige vormgeving van dit proefschrift! Lieve Limburgers, in de vier jaar dat ik aan dit boekwerk werkte is er nogal wat lief en leed gedeeld. Dank jullie wel, voor alle afleiding en ontspanning (of dat nou in Limburg is of elders), maar nog veel meer, dat jullie er altijd zijn als iets te vieren of te huilen valt. Samen maken we er een feestje van, zetten we de schouders eronder, komen we in de winter bij van alle chaos en gaan we met frisse energie het nieuwe jaar in!

Pap en mam, dank dat jullie altijd voor mij klaarstaan. Voor praktische zaken, advies, aanmoediging, of om gewoon neer te ploffen op de bank en bij te komen in drukke tijden. De ga-het-maar-makeninstelling die jullie mij meegaven is de drijvende kracht achter dit boek geweest (wat overigens ook geldt voor zoveel andere goede dingen in mijn leven). Daarvoor ben ik jullie onwijs dankbaar! Wendy, Luc, Jeroen, Jolien en Stefan, dank voor het gezellige zooitje dat wij samen zijn en hoe we elkaar enorm kunnen uitdagen, maar vooral ook helpen als het nodig is.

Lieve Lars, ik ben zo blij dat jij in mijn leven bent. Dank je wel dat je altijd naast mij staat. Samen met jou is alles leuker, beklim ik hoge bergen en geniet ik volop. Ik kijk onwijs uit naar alles wat er nog op ons pad komt.