

# Interrater Reliability of a Classroom Observation Protocol A Critical Appraisal

Ning Rui • Jill Feldman  
Research for Better Schools

AERA Annual Meeting, San Diego, CA  
April 16, 2009



# Introduction: Why is the Study Important?

- Constructing a COP presents many challenges and pitfalls: observer bias, obtrusiveness, contextual variances, labor costs, reliability and validity (Dirr, 2003).
- Technical quality (reliability and validity) is critical to the development of any instruments for documenting classroom practices;
- Funders are increasing demands for reporting psychometric properties of measures used in program evaluation.
- Observation protocols that are idiosyncratic to the observer can be limited and misleading for program evaluation purposes.
- There is little consensus available about the best statistical measures for inter-rater reliability of COPs.



# Contribution of This Study

- Report the interrater reliability (IRR) of a federally funded Striving Readers project
- Compare the pros and cons of a variety of measures for estimating item and domain IRRs
- Discuss implications for practice
  - Appropriate IRR measures for COPs
  - Optimal sample size for an IRR study
  - Cost-effective strategies in assignment of observers
  - Barriers to effective assessment (missing data, lack of synchronism across raters)



# Project Background

- The Striving Readers Project (SRP) is one of eight programs sponsored by US Department of Education to address the needs of struggling adolescent readers.
- Striving Readers COP was developed by RBS to record classroom activities of teachers participating in Striving Readers professional.
- The version of COP under study includes seven domains: Physical Environment, Materials/Technology, Classroom Climate, Instructional Modes, Literacy Strategies, Cognitive Demand, and Level of Student Engagement.



# Methods

- Ten pairs of evaluators (N= 10, k= 20) completed a 1.5-day training for use of the SR-COP prior to conducting the classroom observation in Spring 2008.
- Items in different domains are scaled differently
  - Dichotomous items (Yes/No)
  - Likert-scale items
  - Items that require coding based on established rubrics
- Various reliability measures were applied and compared based on their appropriateness for items in each domain.



## A Review of Various IRR Measures

IRR Measure	Description	Scale of Items	Pros	Cons
Joint-prob. of agreement	% of time 2 raters give identical ratings	Nominal	Most simple	Least robust, doesn't account for chance agreement
Simple kappa	$\hat{K} = \frac{p_o - p_e}{1 - p_e}$	Nominal	Account for chance agreement	Treat data as nominal; may underestimate IRR
Weighted kappa	$\hat{K}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$	Ordinal	Recognize distance between categories	Complicated to compute
Pearson's r	$r = \frac{SP}{\sqrt{SS_X SS_Y}}$	Assumed continuous	Simple	Tend to overestimate

## A Review of Various IRR Measures (cont.)

IRR Measure	Description	Scale of Items	Pros	Cons
Spearman's $\rho$	Rank order correlation	Ordinal	For smaller sample size (<30 pairs)	Doesn't account for magnitude of differences across raters
Polychoric correlation	Correlation between two observed ordinal but theorized continuous variables	Ordinal	Apply to latent continuous scales with small # of response categories	IRR tends to be attenuated with smaller # of response categories
Intraclass correlation (ICC)	Ratio of between-groups variance to total variance	Interval	Preferred over Pearson's $r$ when $N < 15$ ; account for differences for individual segments	



# Results: Physical Environment

	1. Resources	2. Space	3. Desk Arrangement	4. Bulletin Board/Walls	5. Availability of Books	Overall
weighted kappa	.524	.627*	-.163 <sup>a</sup>	.803 <sup>a</sup> **	.786 <sup>a</sup> *	.515
Pearson's r	.575	.676*	-.364	.834**	.794**	.503
Spearman's rho	.587	.687*	-.385	.832**	.804**	.505
Kendall's tau	.500	.615*	-.371	.804**	.768*	.463
Polychoric	.653	.816*	-.992	1**	.940*	.483
ICC	.557	.646*	-.429	.818**	.805**	.479
95% CI	(-.037, .866)	(.103, .897)	(-.809, .226)	(.451, .951)	(.420, .947)	

Note. \*  $p < .05$ , \*\*  $p < .01$ .

<sup>a</sup> Corrected for unbalanced contingency tables.



# Results: Materials/ Technology

		$p_o$	$p_+$ (yes)	$p_-$ (No)	kappa
Computers	Present	.800	.800	0	-.111
	Used during observation	.700	.200	.500	.348
Computer printers, scanners, or digital cameras	Present	.900	.600	.300	.783*
	Used during observation	1	.100	.900	1**
Textbook	Present	.400	.300	.100	-.154
	Used during observation	.900	.400	.500	.800**
National Geographic sets	Present	1	0	1	1**
	Used during observation	1	0	1	1**
Other books or articles	Present	.800	.300	.500	.583
	Used during observation	.700	.100	.600	.211
Other printed materials	Present	.300	.100	.200	-.129
	Used during observation	.900	.500	.400	1**



## Results: Materials/Technology (cont.)

		$p_o$	$p_+$ (yes)	$p_-$ (No)	kappa
TV, VCR/DVD, or radio/CD player	Present	.900	.900	0	n/a <sup>a</sup>
	Used during observation	.800	.100	.700	.412
Interactive display/ projector	Present	1	0	1	1**
	Used during observation	.900	0	.900	n/a <sup>a</sup>
Overhead projector, LCD projector	Present	.700	.600	.100	.286
	Used during observation	.800	0	.800	n/a <sup>a</sup>
Tools	Present	.700	.300	.400	.400
	Used during observation	1	.300	.700	1**
Notebooks	Present	.600	0	.600	n/a <sup>a</sup>
	Used during observation	.700	.300	.400	.400

Note. \*  $p < .05$ , \*\*  $p < .01$ , based on two-sided test  $H_0: \text{kappa} = 0$ ,<sup>a</sup> unbalanced kappa table.



# Results: Classroom Climate

	Structure	Active Participation	Respect	Interactions	Open Inquiry	Intellectual rigor	Average
Simple Kappa	.069	.157	.118	-.154	.324	.315	.138
Weighted kappa	.058	-.106	.094	-.261	.749*	.946**	.247
Polychoric	.469	-.418	-.160	-.058	.899	.863	.266

Note. \*  $p < .05$ , \*\*  $p < .01$ .



# Results: Instructional Modes

	Interval 1	Interval 2	Interval 3	Interval 4	Interval Mean per Classroom
Classroom 1	0.167	0.333	0.000	0.500	0.250
Classroom 2	0.800	0.600	0.800	0.800	0.750
Classroom 3	0.333	0.750	0.667	0.750	0.625
Classroom 4	1.000	0.500	0.000	0.000	0.375
Classroom 5	0.000	0.600	0.600	0.750	0.488
Classroom 6	0.667	0.571	0.667	0.667	0.643
Classroom 7	0.500	0.500	0.667	0.000	0.417
Classroom 8	0.800	1.000	1.000	0.833	0.908
Classroom 9	0.750	1.000	0.750	0.200	0.675
Classroom 10	0.600	0.333	0.500	0.500	0.483
Classroom Mean per Interval	0.562	0.619	0.565	0.500	



# Results: Cognitive Demand

	Interval 1	Interval 2	Interval 3	Interval 4	Interval Mean
Weighted kappa	.105	.486	-.125	.188	.164
% agreement	.400	.500	.600	.500	.500



# Results: Student Engagement

	Interval 1	Interval 2	Interval 3	Interval 4
<b>Polychoric</b>	<b>.752</b>	<b>n/a<sup>a</sup></b>	<b>.997</b>	<b>1.000</b>
Weighted kappa	.314	n/a <sup>a</sup>	.546	.778
% agreement	.600	.400	.600	.800

<sup>a</sup> Statistics not computed because of extremely unbalanced contingency table (all standard errors are zero).



# Conclusion

- Results indicate that some items can be assessed with moderate reliability;
  - Physical Environment
  - Cognitive Demand
  - Student Engagement
- Analysis of items in Classroom Climate and Instructional Modes yielded mixed results
- Insufficient data precluded us from estimating IRR and making reasonable conclusions about items in Literacy Strategies
- Weighted kappa seems to be more suitable for establishing IRR in similar instruments
- ICCs are very close to Pearson's  $r$  and Spearman's  $\rho$  because their assumptions about the underlying variable are very similar: the data may be considered interval or continuous.
- Narrative descriptions and logs are helpful in decoding and explaining disparities between raters.



# Lingering Issues and Takeaways

- Conflict of labor costs and sample size requirement to formally establish IRR of the COP
- Water, Eliasziw, and Donner (1998) proposed method of computing optimal sample size as a function of ICC and # of ratings received per subject
- Conflict of requirement for multiple observers per classroom and disruption of normal class environment (Suggested strategy: arranging multiple visits to the classroom by different observers)
- Need for increased control for observer bias, contextual variances, and data quality (Suggested strategies: Focus on training of observers to a high confidence level in use of the COP and use combined approaches to help remedy potential pitfalls)

