US-China Education Review B 3 (2012) 305-315
Earlier title: US-China Education Review, ISSN 1548-6613

IRR (Inter-rater Reliability) of a COP (Classroom Observation Protocol)—A Critical Appraisal

Ning Rui
Research for Better Schools, Philadelphia, USA

Jill M. Feldman
Westat, Rockville, USA

Notwithstanding broad utility of COPs (classroom observation protocols), there has been limited documentation of the psychometric properties of even the most popular COPs. This study attempted to fill this void by closely examining the item and domain-level IRR (inter-rater reliability) of a COP that was used in a federally funded striving readers program. A combination of reliability measures (e.g., joint-probability of agreement, Cohen's kappa, polychoric correlation and intra-class correlation coefficients) was selected dependent upon which were appropriate given the scale of each item set. Results indicate that most items in physical environment, cognitive demand and students' class engagement can be assessed with moderate reliability. Items in classroom climate and instructional modes yielded mixed estimates. Recommendations were provided for possible improvement of similar instruments.

Keyword: COP (classroom observation protocol), IRR (inter-rater reliability), adolescent literacy, program evaluation

Introduction

A COP (classroom observation protocol) is an instrument used to assess and measure the quality of teaching and learning in the classroom, identify how well resources and learning environment are contributing to learning and provide suggestions on areas for possible improvement and development. Notwithstanding broad potential utility of COPs, it does not suffice that the instruments simply remain consistent internally or over reasonable periods of time. Rather, for COPs to be useful for teacher PD (professional development) evaluation, it should be shown that observers of the same class session concur substantially on the degree to which the instructor's classroom behaviors, methods and modes of interaction with students conform to a preexisting concept of what represents good teaching. In other words, observation protocols that are idiosyncratic to the observer, but not the instructor, can be limited and misleading for evaluation purposes. Unfortunately, there has been very limited documentation of the psychometric properties of even the most popular COPs currently used in evaluations of various instructional and teacher PD programs nationwide. In particular, there is little consensus about what statistical measures are best to analyze the IRR (inter-rater reliability) of this type of instruments. As funders increase demands for more rigorous government-funded evaluations of educational programs and interventions, one way that evaluators can meet these demands is by using the most appropriate statistical measures for estimating the psychometric properties of specific

Ning Rui, Ph.D., Research for Better Schools.
Jill M. Feldman, Ph.D., Westat.

protocols. The present study attempts to address this issue through a critical appraisal of a COP used in a federally funded striving readers program. The COP was developed to inform implementation fidelity ratings of a school-wide PD model designed to support middle school content area teachers' implementation of literacy strategies in ways that support the academic achievement of students who attend high poverty urban middle schools.

Background

The SRP (Striving Readers Project) under study, situated in a large high-poverty urban school district in the South, is one of the eight programs sponsored by US Department of Education to address the needs of struggling adolescent readers and includes school-wide and targeted interventions plus rigorous evaluations of each component. The CLA (Content Literacy Academy), a school-wide PD model for content area teachers, provided 180 hours of intensive training over two years to increase teachers' knowledge about and use of research-based reading strategies to improve students' achievement in reading and core content areas (mathematics, science, social studies and English language arts), especially for students attending high-poverty urban middle schools. The SR-COP (Striving Readers Classroom Observation Protocol) was developed by Research for Better Schools, a non-profit research and development organization in Philadelphia, as an instrument to record and rate observations of Striving Readers' classroom lessons as a part of the evaluation of CLA. The instrument was adapted from the CETP (Collaborative for Excellence in Teacher Preparation)—a classroom observation tool developed by Lawrenz, Huffman, and Appeldoorn (2002) at University of Minnesota.

The COP items were organized into seven domains:

(1) Physical environment;

(2) Materials/technology;

(3) Classroom climate;

(4) Instructional modes;

(5) Literacy strategies;

(6) Cognitive demand;

(7) Level of student engagement.

When inter-rater agreement is low, there are usually two reasons: (1) The scale is defective; and (2) Raters need to be retrained on the rating criteria. One of the main challenges of estimating reliability for SR-COP is that the items in different domains are scaled differently. For physical environment, all five items are in a 1-4 Likert-scale; for materials/technology, there are 12 dichotomously scaled (Yes/No) items; for classroom climate, there are six categorical items in a scale of 1, 2, 3, 4 and DK (do not know); for instructional mode, literacy strategies, cognitive demand and level of student engagement, observers indicated the use of specific modes of instruction, literacy strategies, levels of cognitive demand and student engagement in each of the four 10-minute intervals of the class through transcription of detailed field notes which are then used to complete the SR-COP data matrix. Except for cognitive demand and student engagement, there may be more than one strategy (each with an associated code) that the observer can choose to describe instruction in each interval.

Methods

The SR-COP was used by 10 pairs of evaluators to collect data about classroom implementation related to use of CLA strategies during observations conducted in spring, 2008. The purpose for conducting these observations was to determine the IRR of data collected using the SR-COP among evaluators who completed a two-day training session designed to initiate team members in its use. Data collected by these 10 pairs of raters (N = 10, k = 20) were used to calculate estimates of the SR-COP's IRR.

Due to the variability of items and number of domains that comprise the SR-COP, there are a variety of measures that can be used for calculating item, domain and overall inter-rater reliabilities (see Appendix A for a comparison of the pros and cons of various reliability measures). In the present study, a combination of IRR coefficients were selected dependent upon which were appropriate given the scale of each item set. A crude measure of IRR is joint-probability of agreement and is calculated as the percent of time when both raters indicate identical ratings. However, this measure assumes that scales are nominal and does not take into account that agreement may happen by chance, hence is the least robust measure of IRR. Cohen's (1960) simple kappa coefficient is a commonly used method for estimating paired inter-rater agreement for nominal scale data and includes an estimate of the amount of agreement solely due to chance.

Cohen's simple kappa was expressed by the following equation:

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

where $p_o = \sum_{i=1}^{k} p_{ii}$, i.e., the observed proportion of agreement; and $p_e = \sum_{i=1}^{k} p_{i.}p_{.i}$, i.e., the proportion of agreement expected by chance (Fleiss, 1981). However, items in the physical environment domain are more appropriately treated as ordinal instead of nominal variables because the values in each item represent categories with some intrinsic ranking, such as "1 = crowded" to "4 = spacious" or "1 = few" to "4 = plentiful". One solution to dealing with ordinal data is to use weighted kappa, which recognizes the distance among successive categories. The weighted kappa coefficient was defined as:

$$\hat{\kappa}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}} \tag{2}$$

where $p_{o(w)} = \sum_i \sum_j w_{ij} p_{ij}$ and $p_{e(w)} = \sum_i \sum_j w_{ij} p_i p_j$.

Fleiss and Cohen's (1973) kappa coefficient weights are used to calculate the weighted kappa coefficients for items in the physical environment domain. The weight is in a quadratic form as follows:

$$w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_C - C_1)^2} \tag{3}$$

where C_i is the score for column i, and C is the number of categories or columns. When C = 2, i.e., the variable is dichotomous, the weighted kappa is reduced to the simple kappa coefficient. The quadratically weighted kappa is approximately equivalent to ICC (intra-class correlation coefficients) for ordinal data (Fleiss & Cohen, 1973) and as such is useful to measure the proportion of between-subject variance (as different from within-subject variance) associated with differences among the scores of subjects.

There are other correlation coefficients that can be used to measure the degree of agreement between two raters, depending on the assumptions made about the item scales. For example, the polychoric correlation coefficient is a good means of measuring inter-rater agreement for ordinal data. It would estimate what the correlation between raters would be if ratings were made on a continuous scale. If we assume that ratings of physical environment items are made on a continuous scale, then the use of the Pearson product-moment correlation coefficient would be appropriate. If we assume that the scores on these items are ordinal ranked, then Spearman correlation coefficient (Spearman, 1904) or Kendall's (1948) tau would be an appropriate measure of IRR. Given the structure of physical environment domain, IRR was estimated using each approach and the results were compared.

Results

Physical Environment

The IRR estimates for physical environment items are presented in Table 1. The estimates for the items are also averaged to create an IRR estimate for the domain. As indicated in Table 1, kappa, Pearson's *r*, Spearman's rho and Kendall's tau result in consistent IRR estimates and associated levels of statistical significance. For most items in the first domain, the Pearson's *r*, Spearman's rho and Kendall's tau fall within the expected range (0.5 to 0.8), except for "desk arrangement", which has a negative value using all reliability measures. A possible reason of low agreement on desk arrangement is that this is the only item that explicitly asks for raters' subjective opinion ("appropriateness"), compared to the other four items that tend to elicit more objective observation and judgment about the classroom set-up (e.g., whether the resources are sparsely or richly equipped, whether the space is crowded or spacious, whether the bulletin boards are rich with content-related materials and whether plenty of books are available). Therefore, a suggestion for improving the reliability for item 3 would be to use an adjective that would describe the desk arrangement more objectively.

The weighted kappa tends to be more conservative, indicating that there is moderate inter-rater agreement on resource availability, substantial agreement on classroom space, richness of materials on bulletin boards and availability of books; and no agreement on the appropriateness of desk arrangement for classroom activities/tasks. The arithmetic mean of weighted kappa for the physical environment domain is 0.515, indicating moderate domain-level agreement between raters.

Table 1

Item-Specific and Average IRR Estimates for Physical Environment

	1. Resources	2. Space	3. Desk arrangement	4. Bulletin board/walls	5. Availability of books	Overall
Weighted kappa	0.524	0.627*	-0.163^{a}	0.803$^{a\ **}$	0.786$^{a\ *}$	0.515
Pearson's *r*	0.575	0.676*	-0.364	0.834**	0.794**	0.503
Spearman's rho	0.587	0.687*	-0.385	0.832**	0.804**	0.505
Kendall's tau	0.500	0.615*	-0.371	0.804**	0.768*	0.463
Polychoric	0.653	0.816*	-0.992	1**	0.940*	0.483

Notes. $^{*}p < 0.05$, $^{**}p < 0.01$; a Corrected for unbalanced contingency tables.

To estimate kappa, both raters must use the same number of rating criteria. That is, the number of rating categories used by rater 1 should equal the number of categories used by rater 2. While calculating the kappa for items 3, 4 and 5, we found that the number of rating categories used by two raters did not match and resulted in an unbalanced contingency table (see Table 2 for item 3). The problem of incomplete use of all rating criteria is less likely with larger samples.

Table 2

Contingency Table of Rating Categories by Two Raters for Item 3

Frequency		Rater 2			Total
		2	3	4	
Rater 1	3	0	0	1	1
	4	1	5	3	9
Total		1	5	4	10

To address the problem, a strategy of correcting unbalanced kappa tables like the one developed by Crewson (2001) was used to add dummy observations to the data set for both raters to compensate for those who did not use all the categories in the rating scale (e.g., rater 1 did not use categories 1 and 2 and rater 2 did not use category 1). The resultant "balanced" contingency table is shown in Table 3. Following this, a control variable was created to classify the original observations and dummy observations as missing to effectively balance the data table and exclude dummy observations from the sample.

Table 3

Corrected Contingency Table of Rating Categories by Two Raters for Item 3

		Rater 2				Total
		1	2	3	4	
Rater 1	1	0	0	0	0	0
	2	0	0	0	0	0
	3	0	0	0	1	1
	4	0	1	5	3	9
Total		0	1	5	4	10

Materials/Technology

The second domain consists of 24 dichotomously scaled items about the availability and use of various instructional materials (e.g., books, maps, art supplies, etc.) and technology equipment (e.g., computer, printer, projector, etc.). Overall percent of agreement (p_o) is a good method to measure IRR for items using a dichotomous scale. Cohen (1960) criticized the use of overall agreement p_o because of its tendency to ignore chance-related inflation or bias, which may affect only more frequently used categories. For example, p_0 can be high even when raters randomly assign ratings based on probabilities that are equal to the observed base rates. If one rater carelessly checked "Yes" for all items and the other one gave more objective ratings, the two raters would still share a "high" agreement, if the class indeed met the criteria for 80% of items. However, Cohen's logic is also problematic, because it is unclear what advantage there is to compare an observed percent of agreement with a hypothetical value p_c which is unknown in reality. To address this dilemma, we jointly consider the percent of positive (p_+) agreement and percent of negative (p_-) agreement along with kappa coefficient shown in Table 4. If p_+ and p_- are relatively even, the likelihood of bias due to chance-based agreement is much smaller.

Classroom Climate

The classroom climate domain consists of six items that are purported to assess the degree to which the instructor is able to create a respectful, inclusive and energized climate with a sense of open inquiry in the classroom. Responses were rated in a Likert scale where “1” denotes “not at all” and “4” denotes “to a great extent” with an additional “DK” category. The DK category was coded as 9. We calculated Cohen’s simple kappa (appropriate for nominal items) for each item, and then compared the estimates to those of Cohen’s weighted kappa and polychoric correlation coefficient (see Table 5).

Table 4

Overall, Positive and Negative Agreement and Kappa for Materials/Technology Items

		p_o	p_+ (Yes)	p_- (No)	Kappa
Computers	Present	0.800	0.800	0	-0.111
	Used during observation	0.700	0.200	0.500	0.348
Computer printers, scanners or digital cameras	Present	0.900	0.600	0.300	0.783*
	Used during observation	1	0.100	0.900	1**
Textbook	Present	0.400	0.300	0.100	-0.154
	Used during observation	0.900	0.400	0.500	0.800**
National geographic sets	Present	1	0	1	1**
	Used during observation	1	0	1	1**
Other books or articles	Present	0.800	0.300	0.500	0.583
	Used during observation	0.700	0.100	0.600	0.211
Other printed materials	Present	0.300	0.100	0.200	-0.129
	Used during observation	0.900	0.500	0.400	1**
TV, VCR/DVD, or radio/CD player	Present	0.900	0.900	0	n/a^{a}
	Used during observation	0.800	0.100	0.700	0.412
Interactive display/projector	Present	1	0	1	1**
	Used during observation	0.900	0	0.900	n/a^{a}
Overhead projector, LCD (liquid crystal display) projector	Present	0.700	0.600	0.100	0.286
	Used during observation	0.800	0	0.800	n/a^{a}
Tools	Present	0.700	0.300	0.400	0.400
	Used during observation	1	0.300	0.700	1**
Notebooks	Present	0.600	0	0.600	n/a^{a}
	Used during observation	0.700	0.300	0.400	0.400

Notes. $^{*}p < 0.05$, $^{**}p < 0.01$, based on two-sided test H_0: Kappa = 0; ^{a} unbalanced kappa table.

Table 5

Item-Specific and Average IRR Estimates for Classroom Climate Domain

	Structure	Active participation	Respect	Interactions	Open inquiry	Intellectual rigor	Average
Simple kappa	0.069	0.157	0.118	-0.154	0.324	0.315	0.138
Weighted kappa	0.058	-0.106	0.094	-0.261	0.749*	0.946**	0.247
Polychoric	0.469	-0.418	-0.160	-0.058	0.899	0.863	0.266

Notes. $^{*}p < 0.05$, $^{**}p < 0.01$.

As shown in Table 5, the item-specific inter-rater reliabilities are generally very low for the classroom climate domain. Only “open inquiry” and “intellectual rigor” fit into the range for “fair agreement”. When we relaxed the assumption of the item data (e.g., assume that the items are in an ordinal scale or even continuous scale), the weighted kappa and polychoric correlation coefficient gave a higher value for the “open inquiry” and "intellectual rigor". There are two possible reasons for such a low kappa. First, introducing a DK category could be problematic, if it encourages respondent to avoid giving a response (in this case, since "DK" is coded as 9, it would seriously lower the weighted kappa and polychoric if one rater gave a "DK" while the other gave a relatively lower rating, such as 1 or 2. Second, since kappa is very sensitive to possibility of chance agreement, it is likely to have a low kappa in the face of high percentage of cases that two raters agree on. It appears that Cohen's proposition of comparing the observed agreement with a hypothetical true agreement can be misleading in some cases where the statistic can mistakenly consider some of the actual agreement as "chance agreement".

Instructional Modes

For the instructional modes domain, the observers were told to familiarize themselves with the definitions of the types of instruction prior to the classroom observation. A list of 23 codes for various instruction modes (such as "LWD" = a lecture with discussion, "HOA" = the use of hands-on activities, or "DP" = drill and practice) with detailed definitions are included in the annotated guide. Since teachers often use multiple instructional methods during the class period, more than one code may be assigned to each of the four 10-minute intervals. Because of this reason, it is challenging to use traditional statistical measures to estimate the inter-rater agreement for items in this domain, where each rater has multiple answer options. Another challenge is the discrepancy of timing between raters during each interval, i.e., one rater may start the observation earlier than the other. Given this specific situation, we decided to estimate the percent of agreement for each time interval with a two-step process: (1) For each classroom during each time interval, we compared the codes assigned by two raters and computed the percent of agreement; and (2) We computed the mean inter-rater percent of agreement of all 10 classrooms. The results are shown in Table 6.

Table 6

Inter-rater Percent of Agreement on Instructional Modes for Four Intervals

	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	

Table 6 shows that the average percent of inter-rater agreement for each classroom varies widely from 0.250 to 0.908. However, the mean percent of agreement is relatively stable across four intervals, ranging from 0.500 to 0.619. The classrooms with inter-rater agreement below 0.500 all have various degree of discrepancy in timing across raters.

Literacy Strategies

We did not compute the IRR estimates for the literacy strategies domain, because no literacy strategies were observed in more than half of the classrooms. The number of classes with observable literacy activities was simply too low to make reasonable and reliable conclusions about item and domain level IRR.

Cognitive Demand

In this domain, an observer was asked to fill in the type of cognitive demand that the instructor used during each 10-minute interval of the class. A list of six codes for various cognitive demands (such as 1 = Remember, 2 = Understand, 3 = Apply, 4 = Analyze, 5 = Evaluate and 6 = Create) with detailed definitions was provided in the annotated guide. Unlike the instructional mode domain, an observer can choose only one type of cognitive demand for each time interval, thus, making the estimation of IRR easier. The six codes are associated with an ordinal sequence of cognitive levels, where a greater value represents a higher demand than the previous one. Therefore, we used Cohen's weighted kappa to estimate the item-specific IRR in this domain. The results of weighted kappa in comparison with raw percent of agreement for all four intervals are shown in Table 7.

Table 7

Inter-rater Percent of Agreement and Cohen's Weighted Kappa on Cognitive Demand for Four Intervals

	Interval 1	Interval 2	Interval 3	Interval 4	Interval mean
Weighted kappa	0.105	0.486	-0.125	0.188	0.164
Agreement (%)	0.400	0.500	0.600	0.500	0.500

Like what is shown in previous domains, the weighted kappa, which takes into account of chance agreement and ordinal structure of the data, tends to be low even in the face of relatively high raw percent of agreement. The difference of two estimates is especially large when two raters gave sharply different ratings for one interval (e.g., during interval 3, one rater gave a "1" and the other gave a "4"). In addition, the lack of synchronism in observation across raters (as evidenced by the disparity in their timing) during each interval could affect both the validity and reliability of any time-varying items.

Level of Engagement

In this domain, each observer evaluated the level of student engagement in a classroom based on a three-level scale as follows: (1) LE (low engagement): 80% or more of the students are off-task; (2) ME (mixed engagement); and (3) HE (high engagement) 80% or more of the students are engaged.

The underlying trait to be assessed, percent of students who are engaged, is a continuous and normally distributed variable. Based on this assumption, polychoric correlation would be the most appropriate measure. Polychoric correlation is used when one is trying to estimate the correlation between two theorized continuous latent variables from two ordinal scale variables. For comparison purpose, we computed the polychoric correlation coefficients as well as weighted kappa and raw percent of agreement on student engagement for all four intervals (see Table 8).

Table 8

Polychoric Correlation, Weighted Kappa and Raw Percent of Agreement on Level of Engagement

	Interval 1	Interval 2	Interval 3	Interval 4
Polychoric	0.752	n/a [a]	0.997	1.000
Weighted kappa	0.314	n/a [a]	0.546	0.778
Agreement (%)	0.600	0.400	0.600	0.800

Note. [a] Statistics not computed because of extremely unbalanced contingency table (all standard errors are zero).

In general, polychoric correlations show a high degree of agreement between the raters for student engagement observation. Both kappa and polychoric coefficient are not computed for time interval 2. This is due to the fact that one rater gave a rating of 3 (high engagement) to all classrooms and the other gave ratings of 1, 2 and 3 to various classrooms. This creates an extremely unbalanced contingency table as follows, where two rows sum to zero (see Table 9).

Table 9

Unbalanced Contingency Table for Ratings on Level of Engagement

Frequency		Rater 2			Total
		1	2	3	
	1	0	0	0	0
Rater 1	2	0	0	0	0
	3	1	5	4	10
Total		1	5	4	10

To avoid the problem of unbalanced contingency tables in the future, one can consider expanding the number of categories in order to assess the levels of student engagement more accurately. For example, one can use decile (1/10) or quartile (1/4) as a unit to quantify the proportion of students engaged in a classroom, which would represent the continuous and normally distributed latent trait more precisely. In addition, the small sample size in this study makes it less likely to detect significant kappa estimates, which should be taken into consideration in future study designs.

Use of ICC to Measure IRR (Inter-rater Reliability)

Two main issues that have not been addressed by Cohen's kappa or percent of agreement are the measure's sensitivity to sample size (a lot of kappa estimates are not significant in this study due to the small *N*) and different pairs of raters being assigned to different classrooms. One solution to these two problems is to use ICC as a measure of IRR, which is superior to traditional correlation coefficients when sample size is small ($N < 15$). ICC can be interpreted as the proportion of variance in a subject's true score that is accounted by between-subject variance, ranging from 0 to 1. In other words, we would expect a high ICC, if there is little variation between the ratings given to the same subject by different observers. Shrout and Fleiss (1979) introduced three classes of ICC for reliability, depending on whether the same observers rate each subject in a study. In the case of MSR-COP (Striving Readers Classroom Observation Protocol) study, two raters were randomly selected from a rater pool to observe a particular classroom, i.e., the pair of observers for one classroom was not necessarily the same as the pair who observed another classroom. For this type of design, Shrout and Fleiss (1979) recommended using one-way ANOVA (analysis of variance) to compute ICC, where classroom is considered as a random effect and observer as measurement error. Unfortunately, ICC only applies to cases where the outcome variable is on an interval scale or continuous scale. Only items in the physical environment and level of engagement domains can be considered approximately on an interval scale. For illustration purpose, we computed ICCs for all five items in the physical environment domain, as shown in Table 10.

Compared to the results in Table 1, the 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. Other than that, it does not seem that ICCs are especially useful for reliability studies of COPs.

Table 10

Item-Specific ICCs for Physical Environment

	1. Resources	2. Space	3. Desk arrangement	4. Bulletin board/walls	5. Availability of books	Mean
Intra-class coefficient [a]	0.557	0.646*	-0.429	0.818**	0.805**	0.479
95% CI (confidence interval)	(-0.037, 0.866)	(0.103, 0.897)	(-0.809, 0.226)	(0.451, 0.951)	(0.420, 0.947)	

Notes. $N = 10$; [a] Intra-class coefficients are Shrout and Fleiss's (1979) indices; $^{*}p < 0.05$; $^{**}p < 0.01$.

Conclusions and Discussions

Results indicate that, with proper staff training and use of the observation protocol, most items in physical environment, cognitive demand and student engagement of reading lessons can be assessed with at least moderate reliability within the context of the Striving Readers program. Items in classroom climate and instructional modes yielded mixed estimates. The IRR for literacy strategies is not reported due to high proportion of missing data. The results are of particular interest due to the increased need to conduct cross-state multi-site evaluation of literacy programs.

Walter, Eliasziw, and Donner (1998) provided guidance on computing the optimal sample size in IRR studies where IRR is measured using intra-class correlation ρ. Based on Walter et al.'s (1998) method, the optimal sample size k is a function of ρ and number of ratings received by each subject. This may be applied to reliability studies involving a subject being assessed either by different raters simultaneously or the same rater over time. Table 11 is retrieved from Walter et al.'s (1998) paper, which provides examples of applying this approach. For instance, in order to detect an IRR of at least 0.60 for a COP with null hypothesis $\rho_0 = 0$ and number of raters per subject $n = 2$, one need at least 14 classrooms. However, one can reduce the required number of classrooms by assigning more observers to each classroom, or choose the strategy that is more cost-effective.

Table 11

An Example of Required Sample Size (k) Using Walter et al.'s (1998) Approximate and Exact Methods

ρ_0	ρ_1	n	K_{approx}	K_{exact}	Difference
0.0	0.2	20	5.05	5.00	0.05
0.0	0.4	10	4.31	4.30	0.01
0.0	0.4	3	16.37	16.06	0.31
0.0	0.6	2	13.87	14.13	-0.26
0.0	0.8	10	2.00	2.20	-0.20
0.2	0.6	2	26.71	26.99	-0.28
0.2	0.8	2	8.70	8.94	-0.24
0.4	0.6	5	35.05	34.01	1.04
0.8	0.9	10	22.61	21.72	0.89

Developing a universal standard measure of IRR for COPs that is acceptable from a psychometrician's perspective has been difficult. However, the challenges will continue to excite the imagination of conscientious education researchers and evaluators. Unless we engage in the effort to generate more scientific measures with acceptable psychometric properties that are used in instructional and PD program evaluation, we will leave the important questions about what works to the ill-informed advocates or opponents of education reform. It is a privilege to initiate studies of this kind to ensure that high-quality process and outcome measures are applied in government-funded evaluation projects that are intended to help the public make wise decisions.

References

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Education and Psychological Measurement, 20*, 37-46.

Crewson, E. C. (2001). *A correction for unbalanced kappa tables*. SUGI (SAS Users Group International) Paper 194-26. Retrieved July 17, 2008, from http://www2.sas.com/proceedings/sugi26/p194-26.pdf

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: John Wiley.

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

Fleiss, J. L., & Davies, M. (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology, 115*, 841-845.

Kendall, M. G. (1948). *Rank correlation methods*. Charles Griffin & Company Limited.

Lawrenz, F., Huffman, D., & Appeldoorn, K. (2002). *Classroom observation videotape guide.* College of Education and Human Development, University of Minnesota.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Shrout, P. E., & Fleiss. J. L. (1979). Intra-class correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-427.

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17*, 101-110.

Appendix A

Table A1

A Comparison of Various Measures of IRR

IRR measure	Description	Scale of items	Pros	Cons
Joint-prob of agreement	% of time two rates give identical ratings	Nominal	Most simple	Least robust, does not account for chance agreement
Simple kappa	$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}$	Nominal	Account for chance agreement	Treat data as nominal; may underestimate IRR
Weighted kappa	$\hat{\kappa}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$	Ordinal	Recognize distance between categories	
Pearson's *r*	$r = \frac{\mathrm{SP}}{\sqrt{\mathrm{SS}_{\mathrm{X}}\mathrm{SS}_{\mathrm{Y}}}}$	Assumed continuous	Simple	Tend to overestimate
Spearman's *ρ*	Rank order correlation	Ordinal	For smaller sample size (< 30 pairs)	
Polychoric correlation	Correlation between two observed ordinal but theorized continuous variables	Ordinal	Applied latent continuous scales with small # of response categories	The IRR tends to be attenuated with smaller # of response categories
ICC	Ratio of between-groups variance to total variance	Interval	Preferred over Pearson's *r* when *N* < 15; can be used for > two raters	