INTER-RATER RELIABILITY OF OPR SCORES (SPRING AND FALL 2010)

Background

Reliability is a necessary condition for validity. With assessments that are scored by more than one judge, inter-rater reliability, or the degree to which different raters use the instrument consistently, is important. There are several ways to calculate inter-rater reliability. Perhaps the most common method is to compute the Pearson correlation coefficient between scores from different raters for the same person or performance. If the evaluators agree with one another, then there should be a high correlation between the ratings given by the one evaluator and those given by the other. Since raters are assessing the same construct, it is expected that inter-rater reliability should be high. In general, "high correlations" are defined as +/-.70 to +/-1.00. In contrast, "low" correlations fall in the range of +/- .10 to +/- .39, and moderate correlations are between +/- .40 to +/- .69.

How high should inter-rater reliability be? Unfortunately, there is no straightforward answer to this question—other than to say "the higher, the better." The importance of reliability clearly depends on the nature of the decision to be made. For example, Salvia and Ysseldyke (1998, p. 163) specify a minimum reliability of .90 for assessments that are used for tracking and placement decisions, and .80 for screening decisions (e.g., recommending a student for further testing). Although providing no numbers, Linn and Gronlund (2000) argue that high reliability is mandatory when assessment-based decisions are important, final, irreversible, unconfirmable by other data, concern individuals, and have lasting consequences. In contrast, lower reliability can be tolerated when decision-making is in the early stages and the decision is of minor importance, reversible, confirmable by other evidence, concerns groups, and has temporary effects. Generally speaking, inter-rater reliability of .70 is considered the minimum acceptable level in most contexts.

Reliability is affected by the degree of variability in the data (i.e., how different scores are from each other). As variability decreases, reliability coefficients tend to decrease, regardless of the consistency of evaluators' scores. Otherwise, low agreement between raters usually indicates a problem with the criteria (ambiguous, unclear), the raters' application of the criteria (incorrect, nonsystematic), or both.

Methods

FSEHD's Observation and Progress Report is a key unit assessment in student teaching. Cooperating Teachers and College Supervisors each conduct three formal observations of candidate performance during this time. Student teaching policy stipulates that one of these observations be completed at the same time by both the College Supervisor and Cooperating Teacher. Therefore, it is possible to examine the inter-rater reliability of College Supervisors and Cooperating Teachers using this common observation.

OPR data from Spring 2010 and Fall 2010 were examined to assess the degree of inter-rater reliability between College Supervisors and Cooperating Teachers. Only those lessons observed on the same date by a candidate's College Supervisor and Cooperating Teacher were used for analysis. This yielded 239 observations (from 211 College Supervisors and Cooperating Teachers) for Spring 2010 and 126 observations (from 117 College Supervisors and Cooperating Teachers) for Fall 2010.

Thirty-eight percent of Spring 2010 College Supervisors and Cooperating Teachers had participated in Fall 2009 OPR training sessions conducted by the Offices of Assessment and Partnerships and Placements. In contrast, only 20% of Fall 2010 College Supervisors and Cooperating Teachers had participated in the same training.

To measure inter-rater reliability between Cooperating Teachers' and College Supervisors' ratings, correlations were conducted between their ratings on each OPR indicator that they evaluated in common (i.e., each indicator in the Planning, Implementation, Content, Climate, Classroom Management, and Reflection sections), in addition to the Capsule Rating. Inter-rater reliability was computed in this way for each semester: Spring 2010 and Fall 2010.

Furthermore, mean ratings for the Planning, Implementation, Content, Climate, Classroom Management, and Reflection sections were computed. Inter-rater reliability was subsequently calculated between average measures on these OPR components.

Findings

Overall, inter-rater reliability between Cooperating Teachers and College Supervisors was not particularly high. Inter-reliability for average ratings on OPR sections ranged from .59 to .75. Inter-rater reliability for individual OPR indicators ranged from .45 to .70. Inter-rater reliability for OPR Capsule Ratings ranged from .68 to .69. Additionally, inter-rater reliability tended to be higher in Spring 2010 than in Fall 2010.

Average Ratings

Inter-rater reliability estimates across average scores different sections of the OPR by semester is displayed in Table 1:

Table 1: Inter-Rater Reliability Estimates across OPR Section by Semester

OPR Sections	Inter-Rater Reliability of CTs and CS Ratings*	
	Spring 2010 (n=233 to 239)	Fall 2010 (n=117 to 126)
Planning	.73	.68
Implementation	.73	.67
Content	.69	.59
Climate	.74	.61
Classroom Management	.75	.68
Reflection	.59	.62
Average	.71	.64

* All correlations are statistically significant at p<.001.

In Spring 2010, inter-rater reliability on all sections except Reflection were close to or surpassed the minimum requirement of .70. In contrast, the inter-rater reliability of Cooperating Teachers' and College Supervisors' mean Reflection ratings was just .59, indicating moderate inter-rater reliability. In Fall 2010, inter-rater reliability on averaged Planning, Implementation, and Classroom Management ratings approach the minimum requirement of .70. In contrast, inter-rater reliability in the areas of Content, Climate, and Reflection were much lower.

As stated earlier, inter-rater reliability was higher for averaged OPR scores than it was for scores on individual indicators. This is no doubt due to the greater heterogeneity of scores in the averaged ratings.

Ratings in Individual Indicators

During both semesters, inter-rater reliability on Planning indicators was in the moderate range (range=.59 to .70). There is little evidence of a pattern of high and low ratings across the two semesters. Indicators with higher inter-rater reliability in Spring 2010 evidenced low inter-rater reliability in Fall 2010 and vice versa (see Table 2).

Table 2: Inter-Rater Reliability on Planning Indicators

Planning Indicators	Inter-Rater Reliability	
	Spring 2010	Fall 2010
1. The design of the lesson demonstrates careful planning and organization, from appropriate set induction to closure.	.70	.58
2. Lesson objectives are measurable and observable.	.64	.60
3. The lesson plan objectives are aligned with GLEs, GSEs, and/or appropriate standards.	.63	.57
4. The instructional strategies, activities and technical resources (e.g. manipulatives, adaptive or assistive technologies, electronic technology) in this lesson plan demonstrate attention to students' experience, preparedness, and/or learning styles.	.70	.57
5. The instructional strategies, activities and technical resources (e.g. manipulatives, adaptive or assistive technologies, electronic technology) in this lesson plan demonstrate attention to issues of access, equity, and diversity for students.	.61	.54
6. The lesson design demonstrates an accurate understanding of content.	.62	.62
7. The lesson is designed to engage students in meaningful instructional tasks related to content.	.60	.68
8. The lesson is designed to be student-centered, take advantage of students' curiosity, and be highly engaging.	.64	.63
9. Formative and/or summative assessments are aligned with objectives.	.59	.66
10. The lesson incorporates flexibility and plans for reteaching and/or extension, if needed.	.61	.53

* All correlations are statistically significant at p<.001.

During both semesters, inter-rater reliability on Implementation indicators was in the moderate range (range=.59 to .65). In the two semesters, Cooperating Teachers and College Supervisors agreed least on the degree to which they observed that the candidate arranged the physical environment to maximize learning and designed/adapted relevant learning experiences that incorporate digital tools and resources to promote student learning and creativity (see Table 3).

Table 3: Inter-Rater Reliability on Implementation Indicators

Implementation Indicators	Inter-Rater Reliability	
	Spring 2010	Fall 2010
1. The teacher candidate arranges the physical environment to maximize learning in this particular lesson.	.59	.59
2. The teacher candidate attends to individual student needs, including learning and behavioral issues.	.65	.59
3. The teacher candidate designs or adapts relevant learning experiences that incorporate digital tools and resources (e.g. manipulatives, adaptive or assistive technologies, electronic technology) to promote student learning and creativity.	.60	.56
4. The pace of the lesson is appropriate for the developmental levels/needs of the students and the purposes of the lesson.	.61	.59
5. The teacher candidate customizes and personalizes learning activities using digital tools and resources (e.g. manipulatives, adaptive or assistive technologies, electronic technology).	.62	.58
6. The teacher candidate uses multiple forms of assessment (e.g., observation, rubrics, oral questioning, etc.) to measure student learning.	.64	.56
7. The teacher candidate's questioning strategies are likely to enhance the development of student conceptual understanding/problem solving (e.g., emphasized higher order questions, appropriately used "wait time," identified prior conceptions and misconceptions).	.63	.57
8. The lesson is modified as needed based on formative assessment within the lesson.	.62	.65

* All correlations are statistically significant at p<.001.

During both semesters, inter-rater reliability on Content indicators was in the moderate range (range=.52 to .64). During both semesters, Cooperating Teachers and College Supervisors agreed least on the degree to which they observed that students were intellectually engaged with important ideas relevant to the focus of the lesson (see Table 4).

Table 4: Inter-Rater Reliability on Content Indicators

Content Indicators	Inter-Rater Reliability	
	Spring 2010	Fall 2010
The content of the lesson is significant and worthwhile.	.64	.50
The content of the lesson is appropriate for the developmental levels of the students in this class.	.60	.53
Students are intellectually engaged with important ideas relevant to the focus of the lesson.	.56	.52
The teacher candidate provides accurate content information and displays an understanding of important concepts.	.63	.53
Appropriate connections are made to other areas of the discipline, to other disciplines, and/or to real-world contexts.	.63	.60

* All correlations are statistically significant at p<.001.

During both semesters, inter-rater reliability on Climate indicators was in the moderate range (range=.45 to .69). There is little evidence of a pattern of high and low ratings across the two semesters. Indicators with higher inter-rater reliability in Spring 2010 evidenced low inter-rater reliability in Fall 2010 and vice versa (see Table 5). Of note, however, is the unexpectedly low degree of inter-rater reliability (r=.45) in Fall 2010 on the degree to which the observed lesson manifested a high proportion of student-to-student communication about the content of the lesson.

Table 5: Inter-Rater Reliability on Climate Indicators

Climate Indicators	Inter-Rater Reliability	
	Spring 2010	Fall 2010
The teacher candidate demonstrates positive relationships with his/her students through interactions, including talk, body language, comments on papers, etc.	.68	.57
There is a sense of community in the classroom. Students treat each other and the teacher candidate with respect.	.69	.60
Active participation of all is encouraged and valued.	.57	.64
The teacher candidate's language and behavior clearly demonstrate that she/he is approachable, sensitive, and supportive to all students.	.63	.51
The climate of the lesson encourages students to generate ideas, questions, conjectures, and/or propositions.	.64	.52
Intellectual rigor, constructive criticism, and the challenging of ideas are evident.	.64	.50
There was a high proportion of student-to-student communication about the content of the lesson.	.64	.45

* All correlations are statistically significant at p<.001.

During both semesters, inter-rater reliability on Classroom Management indicators was in the moderate range (range=.52 to .70). In both semesters, College Supervisors and Cooperating Teachers agreed most on the degree to which the teacher candidate had an effective way to get all students to be attentive (see Table 6). They agreed least on how well the teacher candidate circulated around the classroom to keep students on task, listen to them, and challenge them with questions.

Table 6: Inter-Rater Reliability on Classroom Management Indicators

Classroom Management Indicators	Inter-Rater Reliability	
	Spring 2010	Fall 2010
The teacher candidate has an effective way of getting all students in the class to be attentive.	.70	.64
The teacher candidate does not try to "talk over" the students.	.64	.58
The majority of class time is spent devoted to academic tasks, and time is divided in a meaningful, constructive way.	.63	.55
The teacher candidate circulates the room in order to keep students on task, to listen, and to challenge students with questions, when appropriate.	.62	.52
The teacher candidate provides clear, concise, and specific directions prior to transitions and checks for understanding before moving on to the next task or activity.	.65	.53
The teacher candidate applies a set of fair classroom rules, and behavioral interventions are based on logical consequences.	.69	.63

* All correlations are statistically significant at p<.001.

Overall, inter-rater reliability indices were lower in Reflection than in any other OPR section. During both semesters, inter-rater reliability on Reflection indicators was in the moderate range (range=.51 to .61). See Table 7.

Table 7: Inter-Rater Reliability on Reflection Indicators

Reflection Indicators	Inter-Rater Reliability	
	Spring 2010	Fall 2010
1. The teacher candidate describes how s/he made decisions for planning and implementation.	.57	.57
2. The teacher candidate discusses the strengths and weaknesses of the lesson and generates appropriate ideas for possible improvements.	.58	.61
3. The teacher candidate accurately analyzes and assesses student engagement, progress toward meeting the lesson objectives, and classroom management issues.	.57	.57
4. The teacher candidate is aware of how his/her demeanor, actions, and reactions affect the classroom climate and individual students.	.51	.61
5. Based on this lesson, the teacher candidate sets concrete goals (e.g. related to flexibility, pace, response to behavioral issues, etc.) s/he will focus on for future lessons.	.54	.60

* All correlations are statistically significant at p<.001.

Discussion

Across the board, inter-rater reliability among College Supervisors' and Cooperating Teachers' OPR scores on a lesson they observe in common barely meets or does not meet the minimum requirement of .70. Inter-rater reliability on averaged ratings of OPR sections ranged from .59 to .75. In contrast, inter-rater reliability on individual indicators ranged from .45 to .70, indicating "moderate" levels of inter-rater reliability. College Supervisors and Cooperating Teachers exhibited the lowest amounts of agreement on ratings in the Reflection section of the OPR.

As reliability coefficients tend to decrease when scores have low variability (regardless of the consistency of evaluators' scores), it is quite possible these inter-rater reliability are somewhat deflated. However, it is unlikely that the moderate levels of inter-rater reliability observed are deflated to such a degree as to mask a high level of agreement among College Supervisors and Cooperating Teachers. In contrast, these findings probably indicate a problem with the criteria (ambiguous, unclear), the raters' application of the criteria (incorrect, nonsystematic), or both. This is a more likely explanation for the moderate levels of inter-rater reliability observed in Spring 2010 and Fall 2010.

In general, inter-rater reliability was higher in Spring 2010 than in Fall 2010. This may be influenced by the fact that a higher proportion of College Supervisors and Cooperating Teachers had participated in training on the OPR in Spring 2010 (38%) than in Fall 2010 (20%). These findings suggest that training on the instrument is associated with higher inter-rater reliability.

While College Supervisors' and Cooperating Teachers' OPR scores on the lesson they observe in common are important and concern individuals, it is essential to note that they are not final, irreversible, or unconfirmable by other data, and they do not have lasting consequences in and of themselves. As noted earlier, Linn and Gronlund (2000) argue that lower inter-rater reliability may be more tolerable in this type of situation.

Recommendations

Offices of Assessment and Partnerships and Placements:

- Implement additional training sessions on how to use the OPR and other student teaching assessments
- Consider developing an online training module focusing on the same topics so that raters can complete the OPR (and other) training off-site
- Consider developing (with programs) performance descriptions for each performance level on OPR rating scale
- Provide program-specific IRR data to programs that request them so that programs can reflect on the status of inter-rater reliability in their program and address how to increase it

Programs:

- Discuss rating expectations at beginning of student teaching experience
- Monitor and discuss CSs' OPR ratings with them on all OPRs
- Discuss ratings of jointly observed lesson and reasons for divergent ratings on the same indicators
- Request program-specific IRR data so that faculty can reflect on the status of inter-rater reliability in their program and how to address areas of low inter-rater reliability

References

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and evaluation in teaching* (8th ed.). New York: Macmillan.

Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin.