Measure 3 Candidate Competency at Program Completion

Student Teaching Final Evaluation Rater Correlation Results 2022-2023

Candidates build and apply their knowledge of the learner and learning throughout the EPP program, beginning in introductory courses and continuing through subsequent coursework and practicum experiences, eventually culminating in student teaching and the completion of an action research project. During student teaching, during which candidates receive frequent feedback and support from well-qualified clinical instructors and college supervisors, EPP faculty and clinical faculty use valid, reliable EPP-created assessments to evaluate candidates' dispositions as well as content knowledge, instructional planning and delivery, and professionalism. In this document, we provide data on candidate achievement in student teaching based on the InTASC Standards. During student teaching, "Learning Environment" is one of the subscales on the observation instrument used for weekly, midterm, and then final evaluation and upon which the student teaching grade is based; all completer and all subgroup means exceed the EPP target (rating of 6). Ratings by both College Supervisors and Clinical Instructors, 100% of candidates scored above this target score of 6.

Student Teaching Final Evaluation: Clinical Instructors and College Supervisors evaluate candidates' content knowledge, pedagogical knowledge, and performance on the student teaching assessment rubric. The student teaching assessment measures candidates' progression on competencies aligned to the Virginia Uniform Performance Standards and the InTASC standards. The expectation is that candidates score a 6.0, the target rating of "Proficient," on the final set of evaluations. Candidates participate in a final conference with their college supervisors (CS) and clinical instructors (CI) to review feedback and establish professional development goals for their first year of teaching.

Student teachers scored proficient (7,8) or satisfactory level (6, 5) no student fell below a 6 in any of the InTASC areas. Ratings for student teachers' demonstration of effective teaching were consistent across both raters at the end of the student teaching. Average scores in each category: Professional Knowledge: 7.4, Instructional Planning: 7.5, Active Learning: 7.3, Assessment of Learning: 7.5, Cultural Competence and Environment: 7.4, Professionalism 7.5, Set & Measure Learning Goals: 7.6. Overall, the student teachers excelled in most areas, particularly in curriculum standards, subject matter knowledge, communication, and professionalism. Candidates also demonstrated a strong commitment to student learning, as evidenced by use of data, differentiated instruction, and student involvement in goal setting. Table 1 includes the averages for each InTASC area for the cohort. The group (n=7) data was not disaggregated by licensure area. There were four secondary candidates, two elementary education candidates, and one special education candidate.

Table 1
2022-2023 Student Teacher Final Evaluation Scores by InTASC Standard

InTASC subcategories	Means (CS and CI)	Standard Deviation
1. Instruction & Subject Knowledge	7.4	.71
2. Instructional Planning	7.5	.85
3. Active Learning & Differentiation	7.3	.96
4. Expectations and Assessment	7.5	.74
5. Cultural Competence & Environment	7.4	.68
6. Professionalism	7.5	.90
7. Sets & Measures Learning Goals	7.6	.50

Our analysis of inter-rater agreement in this case study included percent of absolute agreement and percent adjacent. Percent agreement measures the percentage of scores between two raters that are exactly the same, and percent adjacent measures the number of times the scores were exactly the same plus the number of times the scores were only one level different. Percent adjacent lets the researcher know how often there is major *disagreement* between the scorers on the quality of the artifact. The percentage of agreement provides a clear and easily understood statistic (Altman, 1991). We calculated the number of times the clinical instructors and college supervisors agree on a rating, then divide by the total number of ratings, and calculated the percentage of times the ratings fall within one performance level of one another (e.g., count as agreement cases in which rater one gives Teacher-A 4 points and rater two gives Teacher-A 5 points). Results fall between 0 and 100% (Gisev et al, 2013).

Data from this cycle demonstrates Randolph College raters agree our students meet program expectations and are prepared to enter the classroom.

The EPP student teaching rubric scoring ranges (see Table 2) outline the possible scores an individual could receive on an assessment, and the levels of performance that must be demonstrated for each score to be given.

Table 2

Scoring Protocol for Student Teaching Final Evaluation Rubric

	8 7	6 5	4 3	2 1
Sample performance indicators: Examples of teacher work conducted in the performance of the standard may include, but are not limited to:	**Proficient**	**Satisfactory**	**Developing**	**Unsatisfactory**
	Effective performance independently	Performs well with assistance	Requires additional support	Unsuccessful performance

Table 3 summarizes disaggregated data on the seven InTASC standards, and shows high levels of agreement in all areas. Rater agreement on the student teacher observation rubric, a performance assessment, established how closely the clinical instructor and the college supervisor agree about the student teacher's instructional performance in a classroom setting. The clinical instructor and the college supervisor independently code an observation. If observer's codes agree this is evidence that the coding scheme is objective (i.e. similar coding for both raters). Generally, we want our data to be objective, so it is important to establish that inter-rater reliability is high. The student teaching observation rubric includes an eight-point scale with 1 indicating and 8 indicating. When raters scored a 7 (proficient) or an 8 (proficient) these were coded as agreement because they indicate proficient performance.

Table 3

Rater agreement and means by rater role for student teaching final evaluations combined items and subcategories

Categories		Mean	
	Percent Agreement	CI	CS
All items	93.8%	7.7	7.2
Instruction & Subject Knowledge	95%	7.7	7.2
Instructional Planning	90%	7.7	7.2
Active Learning & Differentiation	85%	7.5	7.1
Expectations and Assessment	95%	7.7	7.3
Cultural Competence & Environment	95%	7.6	7.3
Professionalism	100%	7.6	7.3
Sets & Measures Learning Goals	100%	7.8	7.4

The lowest rater agreement occurred in Active Learning & Differentiation, where agreement, though high, was 85%. This rating appeared low compared to the other areas where agreement ranged from 90% to 100%. Raters agreed 100% in two of the seven categories. Compared to the 2021-2022 agreement scores were slightly higher.

Raters agreed above 90% in all but two categories: Active Learning & Differentiation and Instructional Planning. Moving forward, we will meet with College Supervisors and Clinical Instructors to develop clear definitions and examples of the expectations for student teachers related to final evaluations. Practice video tapes will be used during training sessions with clinical instructors and college supervisors.

Subjectivity in scoring was reduced by using standardized scoring criteria via a student teaching rubric, which is based on the InTASC standards. In addition to the percent agreement analysis, we conducted a t-test analysis to support whether there was a significant difference between the clinical instructor and college supervisor ratings of the student teacher observations. The purpose of these analyses was to examine the reliability of scoring on the student teaching final evaluation by two different raters. A total of 246 scores were analyzed. The mean score for the clinical instructors' scores was 7.7 with a standard deviation of .67, and the mean score for the college supervisors' scores was 7.2 with a standard deviation of .82. The results indicate that there is statistical significance between the clinical instructors' scores and the college supervisors' scores, t(244) = -4.34. p = 0.00002.

Reference

Gisev, N., Bell J. S. & Chen, T. F. (2013). Interrater agreement and inter-rater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy. 9, 330–338.