Measure 3 Candidate Competency at program Completion

Student Teaching Final Evaluation Rater Correlation Results 2021-2022

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 coach and mentor to review feedback and establish professional development goals for their first year of teaching.

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 1) 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 1

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 2 summarizes disaggregated data on the seven InTASC standards, and show 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. We recognize it is not possible nor cost effective to

Table 3

Table 2. Rater agreement and means by rater role for student teaching final evaluations combined items and subcategories

Categories		Mean	
	Percent Agreement	CI	CS
All items	93.2%	7.5	7.4
Instruction & Subject Knowledge	88.7%	7.4	7.4
Instructional Planning	92.5%	7.5	7.3
Active Learning & Differentiation	96.2%	7.5	7.3
Expectations and Assessment	98.1%	7.4	7.3
Cultural Competence & Environment	96.2%	7.5	7.5
Professionalism	90%	7.6	7.7
Sets & Measures Learning Goals	95%	7.6	7.4

The lowest rater agreement occurred in Instruction, where agreement, though high, was 88%. This rating appeared low compared to the other areas where agreement ranged from 90% to 98.1%. Compared to scores reported in 2020-2021, the 2021-2022 agreement scores were slightly lower. The EPP plans to redesign the rubric training protocol to include three different video samples; these will include elementary, middle, and high school.

Raters agreed above 90% in all but two categories: Instruction & subject Knowledge and Professionalism. 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 professionalism. In addition, in all methods courses candidates will This will allow us to determine if the current expectation is unclear to the raters or if we need to increase clarification of professional behaviors to the student teachers.

Subjectivity in scoring was reduced by using standardized scoring criteria via a student teaching rubric, which is based on the InTASC standards. An increase in objective scoring was achieved by training the clinical instructors and college supervisors to correctly apply scoring rubrics when observing student teachers. In addition to the percent agreement analysis, we conducted a paired t test analysis to support whether there was a significant different 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 339 paired scores were analyze. The mean score for the clinical instructors' scores was 7.5 with a standard deviation of 80 and the mean score for the college supervisors' scores was 7.4 with a standard deviation of 64. The results indicate that there is no statistical significance between the clinical instructors' scores and the college supervisors' scores, t(338) = 1.5. p = 0.12.

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.