Measure 3 Candidate Competency at Program Completion

Educator Preparation Program (EPP) Student Teaching Evaluation 2023-2024

The student teaching evaluation provides a framework for teacher licensure candidates, college supervisors (faculty mentors), and clinical instructors to monitor and support student teachers' growth during the student teaching sequence. Developed in collaboration with P-12 stakeholders, the internship evaluation measures student teachers' development on competencies aligned to the Virginia Department of Education's Uniform Performance Standards for Teachers and the InTASC Model Core Teaching Standards and Learning Progressions for Teachers. *Based on small numbers in the 2023-2024 students who complete the student teaching component equaled 5, therefore, for this reporting we combined 2022-2023 student teacher rubric results with the 2023-2024 cohort to strengthen the analysis.

Administration
Throught the clinical experience college supervisors and clinical instruction complete the student teaching evaluation. All evaluators complete the student teaching evaluation in the Watermark system. The instrument includes space for evaluators to leave open-ended comments tagged to specific items on the candidates' overall performance. After completing the final evaluation, candidates participate in a conference which includes the college supervisor, the clinical instructor, and the candidate.

Use of Data
Teacher candidates have access to their weekly and final assessment results in the Watermark system. After completing the evaluation on their own, candidates review their results in preparation for a conference with the college supervisor and the clinical instructor. At the conference, everyone involved discussed the results and worked together to help the candidate set professional development goals. The EPP faculty reviews and analyzes the final student teaching evaluation results and share results to support program review and improvement.

Scoring Procedure
Table 1 provides performance level descriptors specific to each criterion and describe expected competence, skills, and performance at each level. The performance level descriptions are intended as progressions across InTASC performance levels. There are currently eight separate categories on the evaluation rating scale. We moved from a range of scores in 4 areas (1-2, 3-4, 5-6, and 7-8), however these were expanded to eight individual scoring areas due to constraints in the Watermark system. College Supervisors and Clinical Instructors found the new system confusing because there seemed to be overlap in the descriptions, and upon review of comments from the student teachers and the clinical instructors, the EPP faculty determined it is worth examining returning to a 4-point scale for 2025-2026 academic session as it would help the clinical instructors scoring more accurately. This will be discussed at our department meetings and with our Advisory Committee with our stakeholders.

Table 1
Student Teaching Evaluation Rubric Scoring

UNACCEPTABLE	UN-SATIFACTORY	EMERGING DEVELOPING	DEVELOPING	EMERGING SATISFACTORY (TARGET LEVEL OR ABOVE FOR CANDIDATES)	SATISFACTORY	EMERGING PROFCIENT	PROFICIENT
1 Point SOL not included in all lessons	**2 Points** SOL not included in all lessons	**3 Points** Classroom teacher helps candidate choose appropriate SOL	**4 Points** Classroom teacher helps candidate choose appropriate SOL	**5 Points** SOL are reviewed and aligned with lesson content with the assistance of the classroom teacher, but sub requirements are not highlighted	**6 Points** SOL are reviewed and aligned with lesson content with the assistance of the classroom teacher, but sub requirements are not highlighted	**7 Points** SOL are appropriately aligned with lesson; sub-SOL requirements are reviewed and included	**8 Points** SOL are appropriately aligned with lesson; sub-SOL requirements are reviewed and included

Progression Levels:
The student teaching evaluation is a developmental continuum, built on both the InTASC standards and the Virginia Uniform Performance Standards. The expectation is that student teachers meet the satisfactory rating in most or all areas by the final evaluation of the student teaching placement. At mid-term, the expectation is that student teachers meet the developing rating in most or all areas. The program does not expect student teachers to be proficient in every area during early clinical experiences, nor are interns expected or required to earn proficient in every area during early clinical experiences.

Establishing Evidence of Content Validity and Reliability
Through the revision process, EPP established evidence of validity for the student teaching evaluation instrument and the results and conclusions generated by this assessment. The instrument is aligned to VUPS and InTASC Model Core Teaching Standards and was designed based on the CAEP framework for assessments. Content area experts participated in the redesign of the instrument and provided feedback on revisions, including feedback on item content and application of the instrument in the student teaching experience.

A sample of supervisor teachers was recruited to provide data for the psychometric evaluation of the Randolph College Student Teacher Final Evaluation for assessing student teachers' performance. The pilot study was designed to test the internal consistency reliability of the rubric sub-scales. Participants (n = 23) were asked to complete the rubric based on a student teacher they had supervised in the last five years. The student teaching rubric contains 26 items with seven sub-scales (see Table 2). To orient participants in completing the rubric, each supervisor was asked to describe the student teacher in three to four sentences with details that made the student unique and memorable. The sample included elementary (n = 10), middle (n = 4), and high school (n = 9) teachers who supervised a student teacher between the spring of 2016 and the fall of 2022.

Five of the seven sub-scales demonstrated strong internal consistency reliability, ranging from .83 to .92 (see Table 1) and exceed research standards for reliability coefficients. The "Student academic progress" sub-scale meets the minimum cut off for internal consistency reliability (α = .70). One item, "communicates student progress in a timely manner" (7.3) demonstrated much higher variance in responses than the other two items. The "Professionalism" sub-scale did not meet the minimum acceptable cut off for internal consistency reliability. The three items assess mastery of standard oral and written English (item 6.1), professional dress (item 6.2), and professional demeanor (item 6.3). The first item (6.1) demonstrated lower variability in responses than the other two items on the sub-scale, and it also did not meet the .30 cut off for corrected item-total correlation. It may be that these items are not necessarily strongly correlated with each other. However, removing the item does not improve the overall reliability of the sub-scale.

Table 2
Student Teaching Final Evaluation Reliability

Sub-Scale	Number of items	Scale Mean	Scale Standard Deviation	Cronbach Alpha Reliability Coefficient (α)
Professional knowledge	4	26.83	4.24	.88
Instructional planning	4	25.74	4.56	.83
Instructional delivery	4	26.57	4.17	.87
Assessment of and for student learning	4	25.74	5.09	.92
Learning environment	4	27.00	4.55	.83
Professionalism	3	20.57	2.78	.52
Student academic progress	3	18.70	3.86	.76

Student Teaching Final Evaluation
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, InTASC categories 1: Learner Development, 2: Learning Differences, 3: Learning Environment, 4: Content Knowledge, 5: Application of Content, 6: Assessment, and 7: Planning for Instruction are used for weekly, midterm, and the final evaluations and upon which a major component of the student teaching grade is based.

Student Teaching Final Evaluation Scoring and Rater Agreement Results 2023-2024
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 or higher, the target rating of "Satisfactory," 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 emerging proficient (7) or satisfactory level (6) no student fell below a 6 on InTASC standard 6: Assessment or standard 7: Planning for Instruction. It is interesting to note the Clinical Instructors did not score any student below a 6 on any of the 7 InTASC standards. In contrast, the college supervisors (who are mostly full-time faculty members) scored at least two students a 4 or 5 (nothing below a 4) on standards 1: Learner Development, 2: Learner Differences, 3: Learning Environment, 4: Content Knowledge, or 5: Application of Content. Two students scored below a score of 6. 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=5) data was not disaggregated by licensure area due to the small sample size. There were two secondary

candidates, one elementary education candidate, and two special education candidates.

Table 3
2023-2024 Student Teacher Final Evaluation Scores by InTASC Standard

InTASC subcategories	Means		STDV	
	CI	CS	CI	CS
Instruction & Subject Knowledge	7.2	6.4	.70	1.0
Instructional Planning	7.5	6.3	.68	.86
Active Learning & Differentiation	7.2	6.3	.69	1.1
Expectations and Assessment	7.6	6.1	.60	1.1
Cultural Competence & Environment	7.3	6.3	.64	.98
Professionalism	7.3	7.0	.46	.53
Sets & Measures Learning Goals	7.3	6.5	.61	.83

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 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 divided 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 4 summarizes disaggregated data from 2022-2023 on the seven InTASC standards and shows high levels of agreement in all areas.Table 5 summarizes data from 2023-2-24 on the seven InTASC standards and shows there was not as high of rater agreement, however the numbers were lower. 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 4
2022-2023 Rater agreement and means by rater role for student teaching final

Categories	Mean Percent Agreement	CI CS
All items	93.8%	7.7 & 7.2
Learner Development	95%	7.7 & 7.2
Learning Differences	90%	7.7 & 7.2
Learning Environments	85%	7.5 & 7.1
Content Knowledge	95%	7.7 & 7.3
Application of Content	95%	7.6 & 7.3
Assessment	100%	7.6 & 7.3
Planning for Instruction	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.

However, agreement scores for the 2023-2024 cohort showed a lower percent of agreement as seen in Table 5.

Table 5
2023-2024 Rater agreement and means by rater role for student teaching final evaluations combined items and subcategories

Categories	Percent Agreement	CI	CS
All items	70%	7.3 &	6.3
Instruction & Subject Knowledge	80%	7.2	6.4
Learner Development	80%	7.5	6.3
Learning Differences	76%	7.2	6.3
Learning Environments	60%	7.6	6.1
Content Knowledge	60%	7.3	6.3
Application of Content	87%	7.3	7.0
Assessment	73%	7.3	6.5
Planning for Instruction			

The lowest rater agreement occurred in Learning Environments & Content Knowledge, where the agreement was 60%. This rating appeared low compared to the other areas where agreement ranged from 70% to 80%. Subjectivity in scoring during 2023-2024 appeared to increase compared to 2022-2023. This was evidence what we compared the clinical instructors (classroom teachers) and the college supervisors (college faculty).

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 135 scores were analyzed in 2023-2024. The mean score for the clinical instructors' scores was 7.4 with a standard deviation of .63, and the mean score for the college supervisors' scores was 6.4 with a standard deviation of .95. The results indicate that there is statistical significance between the clinical instructors' scores and the college supervisors' scores. The t-test yields an extremely low p-value (far below conventional significance levels, such as 0.05). This indicates that the difference in means between the two groups (CI and CS) is highly statistically significant.

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.