Georgia Educational Researcher

Volume 18 | Issue 2 Article 1

Summer 2021

Observation of Field Practice Rubric: Establishing Content Validity and Reliability

Ruchi Bhatnagar
Georgia State University, rbhatnagar@gsu.edu

Carla L. Tanguay
Georgia State University, ctanguay@gsu.edu

Caroline Sullivan
Georgia State University, csullivan@gsu.edu

Joyce E. Many
Georgia State University, jmany@gsu.edu

Recommended Citation
Bhatnagar, Ruchi; Tanguay, Carla L.; Sullivan, Caroline; and Many, Joyce E. (2021) "Observation of Field Practice Rubric: Establishing Content Validity and Reliability," *Georgia Educational Researcher*: Vol. 18 : Iss. 2 , Article 1.
DOI: 10.20429/ger.2021.180201
Available at: https://digitalcommons.georgiasouthern.edu/gerjournal/vol18/iss2/1

Observation of Field Practice Rubric: Establishing Content Validity and Reliability

Abstract

Most teacher education assessments are criticized for lacking validity and reliability. This study describes the process of developing the Observation of Field Performance rubric to assess initial teacher candidates' classroom performance and establishing the content validity as well as reliability of the rubric. A panel of content area experts determined that 10 out of 12 items of the rubric were essential and the CVR was above the acceptable range for all 12 items, indicating that the rubric had a strong content validity. Additionally, the analysis of instructors' ratings on the rubric showed that the rubric had good level of internal consistency and inter-rater reliability. Thus, this study determined that the OFP is a reliable and valid measure of candidate performance during field practice. Establishing validity and reliability not only enables teacher education programs to collect high quality assessment data, it is also crucial for program approval and accreditation decisions by national and state agencies.

Keywords

Performance assessment, content validity, reliability, teacher education rubrics

Creative Commons License

Introduction

Rubrics are widely used in teacher education to assess candidates however, most assessment tools in teacher education are home-grown (Grossman, Hammerness, McDonald, & Ronfeldt, 2008). A majority of these rubrics lack validity and reliability, thus, the data collected by these assessment tools cannot be used as dependable indices of student performance or provide information about program effectiveness (AERA, et.al., 2014; Castle & Shaklee, 2006; Grossman, et.al., 2008). As a result, CAEP and other accrediting agencies have brought focus on the importance of determining validity and reliability of the instruments that assess teacher candidates to make determination about program approval (CAEP Handbook: Initial-Level Programs, 2018). In Georgia, the Professional Standards Commission's program approval standards include use of multiple key assessments to monitor candidate progress, completer achievements, and provider operational effectiveness, using instruments that are valid and consistent (GaPSC, 2018). Thus, guidelines for program approval and accreditation lay out clear expectations for initial teacher education programs emphasizing the use of valid and reliable assessment rubrics.

Faculty from our College of Education and Human Development, which is housed in a large R1 University in a southeastern city in the United States, created a rubric called Observation of Field Performance (OFP) for assessing teacher candidates' performance in their practicum or student teaching courses. The purpose of this rubric was to collect data on various aspects of teacher candidates' performance and competencies during the midpoint and endpoint of the program to provide them formative as well as summative feedback. The rubric was created in collaboration with faculty teaching in various initial teacher preparation programs within the college, with the intention that the rubric would be a generic measure of teacher candidates' effectiveness, regardless of the content area or grade-level of teacher preparation. This study describes the process of rubric development, validity and reliability analysis, and next steps for rubric development.

Research Questions

- Does the Observation of Field Performance rubric have content validity to be an instrument for assessing candidate field performance in diverse content areas?
- Does the Observation of Field Performance rubric possess acceptable internal consistency as well as interrater reliability to be used for assessing candidate field performance in diverse content areas?

Literature Review

Rubrics articulate expectations for teacher candidates by listing criteria of proficiency and performance level descriptions across a continuum of quality, therefore, are used widely in teacher education (Andrade, 2010). Additionally, rubrics are helpful in listing the criteria for both the teacher candidates as well as assessors about the specific expectations in their work and lay out what the various performance levels would look like that describe the work at varying quality levels, from low to high (Jonsson, 2014; Panadero & Jonsson, 2013).

In our college, the OFP rubric measures candidates' performance in field in four broad areas: professional knowledge, instructional delivery, assessment of and for learning, and learning environment. These are important skills and competencies agreed upon by teacher educators and policy makers, which also recommend use of performance-based assessments (Andrade & Heritage, 2017; Bastian, Henry, Pan, & Lys, 2016; Chong & Romkey, 2016; Darling-Hammond, Newton, & Wei, 2013). In order to ensure that these rubric indicators assessed important facets of teacher candidates' preparation, each indicator on the rubric was aligned with and tagged to the Interstate Teacher Assessment and Support Consortium (InTASC) standards (CCSSO, 2013). The instructors use the OFP to rate the candidates at least twice in the program, during the practicum courses and the student teaching courses to ensure that candidates meet important criteria outlined by the InTASC standards during the preparation (CCSSO, 2013).

For home-grown rubrics, researchers recommend that once a rubric is created to the satisfaction of the faculty, the next step should be to determine if it is a valid measure of candidate proficiency that is, determining the appropriateness of the inferences that are made from the assessment (Moskal & Leydens, 2000; Bhatnagar, 2018). Validity refers to the degree to which the evidence supports that these interpretations are correct and that the manner in which the interpretations are used is appropriate (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014). Three types of evidence are commonly examined to support the validity of an assessment instrument: content, construct, and criterion (Bhatnagar, Kim, & Many, 2017; Goldhaber, Cowan, & Theobald, 2017).

Content validity of the rubric is a crucial consideration because it reflects the extent to which a rubric incorporates the knowledge of the content area that is of interest, and assesses if the instrument adequately samples the content domain

(Goldhaber, Cowan, & Theobald, 2017; Moskal & Leydens, 2000). Validity is not a property of a data set but refers to the appropriateness of inferences from test scores or other forms of assessment and the credibility of the interpretations that are made concerning the findings of a measurement effort (CAEP Handbook: Initial-Level Programs, 2018, p. 126). An important piece of validity evidence is item validity. Item validity refers to how well the test items and rubrics function in terms of measuring what was intended to be measured; in other words, the quality of the items and rubrics (Panadero & Jonsson, 2013; & Tabachnick & Fidell, 2019). Alignment to content standards is also considered as a component of content validity evidence that supports the intended use of the assessment results (Kane, 2006). Since the OFP is aligned to the InTASC standards, which are important standards for initial teacher preparation, the rubric items create connections between (a) content standards and instruction; (b) content standards and the assessment; and (c) instruction and the assessment (Davis-Becker & Buckendahl, 2013).

Accrediting agencies like CAEP list an expectation that the educator preparation programs (EPP) should take steps to ensure the validity of the assessment, which may be: construct (the appropriateness of inferences made from test scores based on the construct), content (how well an instrument measures the construct), concurrent (how the instrument compares to other established assessments in the field), or predictive validity (the extent to which scores on this instrument compare to scores on another instrument in the field), and also explain the process used for establishing the validity. In order to be at an advanced level, the EPP is expected to report a validity coefficient for the assessment and the types of validity investigated should go beyond content validity and move toward predictive validity (CAEP Handbook: Initial-Level Programs, 2018).

A related aspect of rubric quality is reliability, which refers to the degree to which scores from a particular test are consistent from one use of the test to the next (Moskal & Leydens, 2000). Reliability is a very important piece of validity evidence; a test score could have high reliability and be valid for one purpose, but not for another purpose (Bookhart, 2019; Dawson, 2017; Jonsson & Svingby, 2007). Therefore, it is important to analyze reliability of the rubric to ensure that it is consistently used across raters to produce quality data. For the purposes of developing a reliable rubric, the following considerations are recommended to increase the clarity of a given rubric: 1) clear definition of scoring categories; 2) clear distinction between scoring categories clear; 3) clear interpretation of two raters in a similar fashion, for a given response utilizing scoring rubric (Chong & Romkey, 2016; Moskal & Leydens, 2000).

Rubric quality is based on the match of the rubric content to the outcomes being measured and the degree to which the wording in each cell of a rubric row is parallel in terms of the wording used and homogeneous in terms of the content being measured (Jonsson & Svingby, 2007). Making rubric indicators clear and concise in their expectations positively impact both interrater reliability and validity (Bhatnagar, Kim & Many, 2017; Kane, 2006). Well defined scoring categories assist in maintaining consistent scoring across raters. In order to improve the quality of selected-response tests that will be used again, poorly functioning items need to be identified so they can be fixed, eliminated, or replaced (Bhatnagar, 2018; Jonsson & Svingby, 2007), and ambiguous or misleading items need to be identified (Moss, Girard, & Haniford, 2006). Qualified raters ideally score the responses for agreement, and the rater information would be used to make changes to the rubrics (Wilson, Hallan, Pechone, & Moss, 2014). Additionally, CAEP recommends collecting student responses on an assessment and looking for patterns in the responses that might identify ambiguous or misleading wording in the rubric and make fixes as needed (CAEP Handbook: Initial-Level Programs, 2018).

Method

Development and Use of the OFP Rubric

Program faculty from the Middle and Secondary Education (MSE) and Early Childhood and Elementary Education (ECEE) departments as well as the assessment coordinator for the college collaborated to create the rubric and the descriptions of the various performance levels. The intention was to create a set of generic indicators of teacher candidate performance, which would work across all initial teacher preparation programs in the college and across all grade levels.

The OFP rubric has undergone a few iterations. The first version had 22 items on which we trained the faculty and supervisors grading students in practicum and student teaching courses in the academic year 2017-2018. Prior to the use of this rubric we tagged the rubric to the InTASC standards to ensure the alignment with standards for initial teacher preparation (CCSCO, 2013).

After the first year of implementation, we obtained feedback from instructors and supervisors and analyzed the data collected on the 22 items. Based on the feedback from the instructors and assessment data, we revised the rubric by eliminating 10 items, making the OFP a 12- item rubric. We also rephrased the language of the items so that the rubric included observable skills and competencies. In 2018-2019, we continued the process of obtaining feedback from instructors/supervisors to make improvements in the rubric language. In fall 2019, we invited faculty and supervisors from various content areas to score the content validity of the rubric.

The OFP has four broad areas, professional knowledge, instructional delivery, assessment for and of learning, and learning environment, each of which includes 2-4 items that the university supervisor or instructor assesses while observing the teacher candidate in the classroom (refer to Appendix A). For example, within Professional Knowledge, instructors rate candidates on: knowledge of learners, content knowledge, academic language, and pedagogical content knowledge. The instructors use the rubric for multiple observations during the practicum or student teaching course, while providing formative scores and feedback on the rubric throughout the semester. At the end of the semester, instructors enter the ratings for the last observation as the summative rating on the OFP rubric. Ratings on the rubric are provided on a 4-point scale (4 = advanced, 3 = proficient, 2= developing, and 1 = insufficient). The expectation at program midpoint is that candidates would get an overall rating of Level 2 or above, and at the program endpoint candidates would get an overall rating of Level 3 or above. Please refer to Appendix A for a copy of the rubric and details about the description of rubric items as well as rubric levels.

Data collected through the OFP rubric are used to monitor overall candidate performance from the midpoint to endpoint as well as to monitor overall program performance. Candidates use the data from this assessment to create their *Action Plan* (after the midpoint assessment) and *Professional Learning Plan* (after the endpoint assessment) to continue the process of growth and development while in the program and into their first year of teaching. The goal of this assessment is to demonstrate progression in the program and readiness for the teaching profession. A *Notification & Documentation Action Plan* and conference is provided for candidates who need improvement prior to the program endpoint. Candidates' progress in meeting action plan goals is monitored by the program coordinator so that candidates have ample opportunities to demonstrate overall competency. The programs run a composite report for their cohort, based on the performance of candidates on the OFP rubric at the mid and endpoint and reflect on overall scores, the areas that candidates displayed strength in, and the areas needing improvement. Thus, the rubric provides important formative and summative feedback not only to the candidates, but also for overall program effectiveness.

Determining Content Validity of the OFP Rubric

We utilized Lawshe's (1975) method to establish the content validity of this rubric. According to this method, the Content Validity Ratio (CVR) is the extent to which an assessment procedure adequately represents the content of the curricular aim(s) being measured.

For this rubric, content includes knowledge (e.g, facts) and skills (e.g., higher order thinking competencies). Establishing content evidence is completed by employing a content panel of experts to determine (1) whether the content item is, *essential* or *not necessary;* and (2) whether the content item is measured properly or not. Drawing from the literature on content validity, we examined the number of raters needed on the panel as well as the acceptable agreement level. For example, Wilson, Pan, and Shumsky (2012) indicated that the CVR ratio drops at 8 raters (.75) which was critiqued as an anomaly. However, for 9 raters the CVR is .78 and meets the criteria for content validity. For 7 raters, the CVR is .99, which is difficult to achieve. We determined it would be ideal to have at least 9 raters from across department and content areas to score the observation rubric on a binomial scale and rate each indicator on the rubric as *essential* or *not necessary* (Lawshe, 1975). Ayre and Scally (2014) expanded on Lawshe's (1975) approach and created a reference table for CVR, based on number of raters, using the binomial probabilities of *essential* and *not necessary*.

We sent out an invitation to program coordinators to recruit subject-area experts. Finally, we were able to recruit 11 panel members and their distribution across departments was as follows: Early Childhood and Elementary Education (2), Special Education (2), Middle and Secondary Education (English Language-Arts, Mathematics, Social Studies, Science, ESOL/World Language (5), Music (1), and Art (1). The panel selected as content experts met the following criteria: at least 2 years as faculty or supervisor with a degree/certification in the designated content area and at least 2 years of experience using the OFP rubric in the field.

The formula of content validity ratio is $CVR=(N_e - N/2)/(N/2)$, in which the N_e is the number of panelists indicating "essential" and N is the total number of panelists. The numeric value of content validity ratio is determined by Lawshe (1975). We referenced the CVR table (Ayre & Scally, 2014) to determine the acceptable CVR, for each rubric indicator, for 11 raters. Raters were asked to consider the relevance of each rubric item for their content area, as well as its ease of scoring as an observable behavior, which helped us respond to our first research question about the validity of the OFP rubric.

Determining Reliability of the OFP Rubric

To ensure interrater reliability, (consistency across raters on the topic) the Associate to the Dean for Clinical Practice provides professional learning using the OFP rubric for program faculty and university supervisors to practice scoring videos of practice and calculating interrater reliability. The college has created an OFP Video Scoring Bank of videos by content area and grade band. These trainings are offered in the fall and spring semesters each year and are an opportunity for the faculty and supervisors to have shared understanding of the OFP rubric elements and understand the goals of assessment at the mid and endpoint in the program.

Reliability is defined as the extent to which measurements can be replicated. In other words, it reflects not only degree of correlation but also agreement between measurements (Shrout & Fleiss, 1979). Intraclass correlation coefficient (ICC) is an index which is calculated by mean squares (estimates of the population variances based on the variability among a given set of measures) obtained through analysis of variance (Shrout & Fleiss, 1979; Tabachnick & Fidell, 2019). ICC has been widely used to evaluate interrater, test-retest, and intra-rater reliability. In our case, we utilized the one-way random-effects model for calculating the ICC (McGraw & Wong, 1996). We randomly selected 42 raters from a larger population of raters with similar characteristics (faculty and university supervisors from various initial teacher education programs in the college). Through the one-way random-effects model we can generalize our reliability results to any raters who possess the same characteristics as the selected raters in the reliability study (Shrout & Fleiss, 1979). We used the average ratings of the 42 raters who scored 6 candidates, where for each of the 6 candidates, a set of raters is chosen at random from a population of raters. Each of these raters scored 6 teacher candidates' work samples on the OFP rubric items, but each candidate was potentially rated by different raters.

In addition, we tested the OFP rubric for internal consistency reliability, which measures if the items on the rubric assess the same general construct. Internal consistency is usually measured using Cronbach's alpha, which calculates pairwise correlations between items; and ranges between negative infinity and one, with higher values indicating higher levels of internal consistency (Tabachnick & Fidell, 2019). Very high reliabilities (0.95 or higher) are not necessarily desirable, as this might suggest that there are redundant items on the rubric (Streiner, 2003). Ideally, a Cronbach's alpha between .8 and .9 indicates a good level of internal consistency, and also suggests that each rubric item collects data on a unique aspect of candidates' proficiency.

Results

Determining the Content Validity Ratio for the OFP Rubric

During the Fall 2019 meeting, 11 content area experts (faculty as well as supervisors) from ECEE, MSE, Art, Special Education, World Language, and Health and Physical education came together to rate the indicators of the OFP rubric.

Agreement was 100% for 10 out of the 12 total rubric items. For *Use of Technology*, 3 raters noted that it was *not essential* and scored it as “0”. For the item *Classroom Safety*, 1 rater marked it as *not essential* or “0”. Thus, of the total 132 instances (11 raters multiplied by 12 rubric items), there were only 4 disagreements, bringing our proportion of agreement to 97%, which is much higher than the essential proportion of agreement of 82% (Ayre & Scally 2014). Table 1.0 indicates rating provided by 11 raters for the 12 rubric items, using the criteria 1= Essential; 0=Not Essential.

Table 1.0

CVR for 12 Observation Rubric Items

	Knowledge of Learner	Content Knowledge	Academic Knowledge	Pedagogical Content Knowledge	Learner engagement	Use of Technology	Differentiation	Assessment for Learning	Modeling/ Providing Feedback	Positive Learning Environment	Classroom Facilitation	Classroom Safety
N	11	11	11	11	11	11	11	11	11	11	11	11
Mean	1.00	1.00	1.00	1.00	1.00	.73	1.00	1.00	1.00	1.00	1.00	.91
Std. Deviation	.000	.000	.000	.000	.000	.467	.000	.000	.000	.000	.000	.302

For the rubric indicator on *Use of Technology*, the CVR was .73 and for the indicator on *Classroom Safety*, the CVR was at .91 (as seen in Table 1). According to Ayre and Scally (2014), when using Lawshe’s method of computing the Computing the Content Validity Ratio (CVR), the critical CVR for 11 raters should be at least .636 (Ayre & Scally, 2014, p. 82). For both indicators rated relatively lower (*Use of Technology* and *Classroom Safety*), the CVR was higher than a .636, indicating that all rubric elements exhibit a strong content correlation and the rubric possess valid content measures for assessing candidates’ clinical practice.

Determining the Reliability of the OFP Rubric

To run the interrater reliability analysis, we downloaded the score report for the OFP rubric completed by 42 instructors from our assessment platform. This report provided the mean scores for each of the rubric items for the 42 instructors, for their ratings of 6 candidates. The data were entered in SPSS and we used the intraclass correlation coefficient (ICC), as the method to compute the interrater reliability of the rubric. The ICC is calculated by dividing the random effect variance, by the total variance, i.e., the sum of the random effect variance and the residual variance.

$$\frac{var(\beta)}{var(a) + var(\beta) + var(\varepsilon)}$$

The reported ICC is the variance for each (random effect) group compared to the total variance of the model. The ICC, thus, assesses the reliability of ratings by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. For the inter-rater reliability, the one-way Intraclass Coefficient of .753 (p< .001) showed a good level of agreement among raters (refer Table 2.0).

Table 2.0

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.068	.022	.215	4.046	10	451	.000
Average Measures	.753	.487	.920	4.046	10	451	.000

Note: One-way random effects model where people effects are random.

We calculated internal consistency reliability using Cronbach's alpha. To determine how accurate the observed value (*x*) is in relation to the true value (*t*), the reliability of *x* is a measure of internal consistency and is the correlation coefficient *rxt* of *x* and *t*.

$$r_{xt} = \frac{stdev(t)}{stdev(x)}$$

Our analysis showed that Cronbach's alpha was .897 (p < .001), indicating high congruence with the group mean scores. In addition, Cronbach's alpha based on standardized items was .904 indicating an excellent level of consistency across the 12 rubric items, meaning that these items as a group measured a common construct of teacher candidates' field performance. Table 3.0 shows the internal-consistency analysis for the OFP rubric.

Table 3.0
Internal Consistency Analysis

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Raters
.897	.904	42

In addition, we ran an inter-item correlation analysis for the 12 rubric items to identify how closely these items aligned. Our analysis showed that for the majority of the items, the correlation was moderate, between the .4 -.7 range. A moderate level of correlation is desirable in rubric items because it indicates that items on the rubric measure a similar construct of teacher competence but are not too closely overlapping and are not redundant (Tabachnick & Fidell, 2019). Table 4.0 below shows the inter-item correlation matrix for the OFP rubric.

Table 4.0
Inter-Item Correlation Matrix for Rubric Items

	KnowOfLnr	ConKnw	AcaKnw	PedKnw	LnrEngt	UseOfTec	Differn	Assess	Modelg	PosEnv	Faciln	Safety
KnowOfLnr	1.000	.639	.558	.501	.654	.687	.599	.682	.684	.637	.669	.527
ConKnw	.639	1.000	.679	.633	.615	.611	.428	.775	.476	.724	.732	.807
AcaKnw	.558	.679	1.000	.672	.729	.624	.673	.748	.626	.756	.670	.723
PedKnw	.501	.633	.672	1.000	.535	.529	.460	.649	.698	.604	.613	.548
LnrEngt	.654	.615	.729	.535	1.000	.610	.538	.714	.718	.658	.538	.553
UseOfTec	.687	.611	.624	.529	.610	1.000	.637	.618	.649	.722	.665	.641
Differn	.599	.428	.673	.460	.538	.637	1.000	.535	.479	.487	.427	.477
Assess	.682	.775	.748	.649	.714	.618	.535	1.000	.680	.651	.587	.731
Modelg	.684	.476	.626	.698	.718	.649	.479	.680	1.000	.689	.692	.507
PosEnv	.637	.724	.756	.604	.658	.722	.487	.651	.689	1.000	.879	.736
Faciln	.669	.732	.670	.613	.538	.665	.427	.587	.692	.879	1.000	.711
Safety	.527	.807	.723	.548	.553	.641	.477	.731	.507	.736	.711	1.000

Thus, our analysis of content validity and reliability of the OFP rubric show that the rubric possesses strong content validity as well as good level of inter-rater reliability and internal consistency. The items of the rubric measure a similar construct but are not too closely related to be considered redundant.

Discussion

Rubrics are helpful in teacher education for making expectations for teacher candidates explicit and to communicate in concrete terms what competencies are expected (Andrade & Heritage, 2017; Brookhart, 2019). Rubrics thus scaffold learning in both formative and summative ways helping candidates and instructors to keep track of progress made over a period of time (Andrade, 2010; Darling-Hammond, Newton & Wei, 2013).

Our college of education and human development has used the OFP rubric in practicum and student teaching courses to assess teacher candidates and it was important for us to determine if this rubric was collecting valid and reliable data for our programs. The rubric was tagged beforehand with the InTASC standards to ensure alignment national initial teacher education standards (InTASC, 2013). The content matter experts who evaluated the OFP rubric items found it to be a valid measure of field performance. The overall agreement as well as item level agreement for a panel of 11 raters was higher than the acceptable level of CVR mentioned by Ayre and Scally (2014), with 10 out of 12 items having 100% agreement among the panel as being *essential*. Based on the feedback received from the faculty and supervisors using this rubric in their courses, we paid attention to the perceived issues with the lower rated elements of the rubric, namely: *Use of Technology* and *Classroom Safety*.

Upon discussion with instructors, we added a clarification statement within the element, *Use of Technology*, that the candidate, "Integrates technology to facilitate learning; involves learners in use of technology; provides rationale if technology is not used." Additionally, for the element of *Classroom Safety*, we added a clarification that it pertained only to Science labs and physical education. Even though from the content validity perspective, all rubric elements were higher than the critical CVR of .636, we are hopeful that these changes to the rubric elements will make it a stronger instrument and will enhance its validity as an assessment tool. Overall, based on the ratings of instructors across content areas and grade-levels, the OFP rubric appears to have a strong content validity, reflecting adequate sampling of the content domains expected to be measured during field placement courses (Ayre & Scally, 2014; Lawshe, 1975).

The ICC of .753 indicated a good level of agreement among the 42 instructors, showing that there was a high degree of consistency across raters in understanding the various rubric items and scoring of the items (Koo & Li, 2016). The internal consistency of the rubric was excellent at .897, showing the rubric items were well aligned and measured the same construct. The college intends to continue an ongoing monitoring of inter-rater reliability indices of the OFP rubric, training of instructors to ensure a shared understanding of the language of rubric and expectations at various points, and continue to obtain feedback on the use of OFP as a formative as well as summative rubric. The next step for the OFP would be to move beyond content validity and establish construct and predictive validity of the instrument.

One limitation of this study was that although we had 42 raters utilize the OFP rubric to score 6 students on our assessment portal, all instructors did not score the exact same set of students, which led us to use the One-Way Intraclass Coefficient for inter-rater reliability. If we could have arranged for all instructors to score the same 6 students, we would have used a 2-way random effects model. However, differences in the content specializations of various instructors prevented us from having all instructors score the same 6 student work samples. Our next steps would be to determine predictive validity of the rubric and understand if performance of teacher candidates on the OFP predicts in any way their performance on the GACE (Georgia content test for teacher certification).

The results from the content validity and reliability analysis of the OFP rubric demonstrate that it is an instrument that collects data which is valuable for providing feedback to the candidates as well as to program faculty about the competencies of the candidates in the field. This home-grown rubric was developed from the insight of program faculty, was refined over a period of time based on instructors' feedback, and also was more specific to our context. These qualities created a greater buy-in for the rubric, as compared to other externally developed rubrics (Bhatnagar, 2018; Margolis & Doring, 2013). As the state of Georgia moves away from the use of edTPA ®, the establishment of validity and reliability of the OFP was an important step in our college's effort to use rubrics that collect high quality data about our initial teacher candidates and programs. From the program approval perspective too, it is important that when edTPA, a valid and reliable assessment is phased out, it is replaced by a rubric that also has established validity and reliability (GaPSC, 2020). Our process of developing the OFP rubric and conducting validity and reliability study also has implications for other teacher education programs in the state as well as the country. Other colleges of education who wish to develop rubrics that collect data on important facets of teacher preparation, while making the rubric specific to their needs and context can learn from our experience, and can also utilize the OFP rubric as one of their performance-based assessments (Bhatnagar, Kim & Many, 2017; Darling-Hammond, Newton, & Wei, 2013). Carefully designed rubrics that are analytic, task specific, and measure aspects deemed important by the field, have the potential to provide valid and reliable data for teacher candidates as well as teacher education programs for continuous improvement.

References

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington, DC: Author.

Andrade, H.L. (2010). Students as the definitive source of formative assessment: academic self- assessment and the self-regulation of learning, in *Handbook of Formative Assessment*, H. L. Andrade and G. J. Cizek (Eds.) (New York, NY: Routledge), 90–105.

Andrade, H., & Heritage, M. (2017). *Using assessment to enhance learning, achievement, and academic self-regulation*. New York, NY: Routledge.

Ayre, C., & Scally, A.J. (2014). Critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 47* (1), 79-86.

Bastian, K., Henry, G., Pan, Y., & Lys, D. (2016). Teacher candidate performance assessments: Local scoring for teacher preparation program improvement. *Teaching and Teacher Education, 59*, 1-12.

Bhatnagar, R. (2018). Student teachers' assessments involving three role groups: Challenges and possibilities. *The European Educational Researcher, 1(2),* 77-97.

Bhatnagar, R., Kim, J., & Many, J. E. (2017). An instrument to study state-wide implementation of edTPA: Validating the levels of edTPA integration survey. *Journal of Research in Education*, *27*(1), 24–33. Retrieved from https://www.eeraorganization.org/jre-spring-2017

Brookhart, S.M. (2019). *Educational assessment of students*. (8th ed.) New York, NY: Pearson Education, Inc.

Castle, S., & Shaklee, B. (Eds.). (2006). *Assessing teacher performance: Performance based assessment in teacher education*. Lanham, MD: Rowman & Littlefield.

Chong, A., & Romkey, L. (2016). Testing inter-rater reliability in rubrics for large scale undergraduate independent projects. *Canadian Engineering Education Association, 16,* 1-12.

Council of Chief State School Officers. (2013). InTASC model core teaching standards and learning progressions for teachers 1.0. Retrieved from https://ccsso.org/resource-library/intasc-model-core-teaching-standards-and-learning-progressions-teachers-10

Council for the Accreditation of Educator Preparation (2018). *CAEP handbook: Initial level programs*. Washington, DC: Author.

Darling-Hammond, L., Newton, S.P., & Wei, R.C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability, 25*(3), 179-204.

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research, and practice. *Assessment & Evaluation in Higher Education,* 42, 347–360. doi: 10.1080/02602938.2015.1111294

Davis-Becker, S.L., & Buckendahl, C.W. (2013). A proposed framework for evaluating alignment studies. *Educational Measurement: Issues & Practice, 32*(1), 23–33.

Georgia Profession Standards Commission (2020). *edTPA Proposed Rule Changes*. Atlanta, Ga: Author.

Georgia Profession Standards Commission (2018). *Georgia Standards for the Approval of Educator Preparation Providers and Educator Preparation Programs*. Atlanta, Ga: Author.

Goldhaber, D., Cowan, J., Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education, 68*, 377–393. doi: 10.1177/0022487117702582

Grossman, P., Hammerness, K., McDonald, M., & Ronfeldt, M. (2008). Constructing coherence: Structural predictors of perceptions of coherence in NYC teacher education programs. *Journal of Teacher Education, 59*, 273-287.

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment Evaluation in Higher Education, 39*, 840–852. doi: 10.1080/02602938.2013.875117

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Education Research Review, 2*, 130–144. doi:10.1016/j.edurev.2007.05.002

Kane, M.T. (2006). Validation. *Educational Measurement, 4*, 17–64. Westport, CT: Praeger.

Koo, T.K., & Li, M.Y. (2016). A guideline for selecting and reporting intra class correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2): 155–163.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Margolis, J., & Doring, A. (2013). National assessments for student teachers: Documenting teaching readiness to the tipping point. *Action in Teacher Education, 35*(4), 271-285.

Martone, A., & Sireci, S.G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332–1361.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10).

Moss, P., Girard, B., & Haniford, L. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162. doi: 10.3102/0091732X030001109

Panadero, E., and Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Popham, W.J. (2017). *Classroom assessment: What teachers need to know*. (8th ed.) New York, NY: Pearson Education, Inc.

Shrout, P.E.Y., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*: 420-428.

Streiner, D.L. (2003) Starting at the beginning: an introduction to coefficient alpha and internal consistency, *Journal of Personality Assessment, 80*, 99-103.

Tabachnick, B.G. & Fidell, L.S. (2019). *Using Multivariate Statistics*. (7th ed). New York, NY: Pearson Education, Inc.

Wilson, M., Hallam, P.J., Pecheone, R.L., & Moss, P.A. (2014). Evaluating the validity of portfolio assessments for licensure decisions. *Education Policy Analysis Archives, 22*(6). Retrieved from http://dx.doi.org/10.14507/epaa.v22n6.2014

Wilson, F.R., Pan, W., & Schumsky, D.A. (2012). Recalculation of critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210.

Appendix A

1.1b Key Assessment 3: Observation of Field Performance (OFP)
Program-Level Key Assessment
(Content Knowledge & Instructional Practice)

Teacher Candidate: Observer: Date:

School: Subject/Lesson Topic: Grade Level:

Directions: This rubric is aligned to INTASC and TAPS Standards. The first page provides opportunity for an overall summary of Observed Strengths/Improvement/Comments. In the feedback section, please write specific evidence and/or comments observed for each indicator throughout the lesson. The Rubric is included for reference. Mentor Teachers may use this rubric to observe and provide regular feedback. University Supervisors use this rubric to observe, provide feedback, and enter final practicum and student teaching observation scores on the electronic rubric via LiveText. The Teacher Candidate should scan the handwritten documents or upload word-processed copies of each observation to LiveText.

Observed Strengths:

Suggestions for Improvement:

Overall Comments:

Observer's Signature

Teacher Candidate Signature

INDICATOR	***Observation Notes & Levels of Proficiency*** **Advanced = 4; Proficient = 3; Developing = 2; Insufficient = 1**	
PROFESSIONAL KNOWEDGE		
1-PK: **Knowledge of the Learner**		1PK: ____
2-PK: **Content Knowledge**		2PK: ____
3-PK: **Academic Language**		3 PK: ____
4-PK: Pedagogical Content Knowledge		4 PK: ____
INSTRUCTIONAL DELIVERY		
1-ID: **Learner Engagement**.		1ID: ____
2-ID: **Use of Technology**		2ID: ____
3-ID: **Differentiation/UDL**		3ID: ____
ASSESSMENT OF AND FOR LEARNING		
1-AL: **Assessment for Learning**		1AL: ____
2-AL: **Provides / Models Feedback**		2AL: ____
LEARNING ENVIRONMENT		
1-LE: **Positive Learning Environment**		1LE: ____
2-LE: **Classroom Facilitation**		2LE: ____
3-LE: ***Classroom Safety**		3LE: ____
STRENGTHS: **AREAS FOR IMPROVEMENT:** **FOCUS FOR NEXT LESSON (TAPS #):**		TOTAL POINTS: ________ RATING: ________

*POINTS	RATING
35-44	Advanced
26-34	Proficient
18-25	Developing
11-17	Insufficient

INDICATOR		***Advanced (4)***	***Proficient (3)***	***Developing (2)***	***Insufficient (1)***
PROFESSIONAL KNOWLEDGE					
1-PK	**Knowledge of the Learner:** Builds upon learners' existing academic, developmental, linguistic, personal, cultural/community strengths, needs, and experiences. * *Including, but not limited to, race, ethnicity, language, religion, socioeconomic status, gender, sexual orientation/expression, national origin, or exceptionality. GA-TAPS-2014.1 GA-TAPS-2014.3 GA-TAPS-2014.4 INTASC-2013.1 INTASC-2013.2 INTASC-2013.8	Maximizes learner's prior knowledge by integrating lesson objectives with learners' academic, personal, developmental, linguistic, **AND** cultural/community strengths, needs, **AND** experiences.	Uses learner's prior knowledge by integrating lesson objectives with learners' academic, personal, developmental, **AND** linguistic, **AND/OR** cultural/community strengths, needs, **AND/OR** experiences.	Uses learner's prior knowledge by integrating lesson objectives with learners' academic, personal, developmental, **OR** linguistic, cultural/community strengths, needs, **OR** experiences.	Does not use learner's prior knowledge by integrating lesson objectives with learners' academic, personal, developmental, linguistic, **OR** cultural/community strengths, needs, **OR** experiences.
2-PK	**Content Knowledge:** Demonstrates accurate and current content knowledge in authentic contexts. GA-TAPS-2014.1 INTASC-2013.4	Demonstrates accurate **AND** current content knowledge in authentic contexts.	Demonstrates accurate **AND** current content knowledge.	Demonstrates accurate content knowledge.	Demonstrates inaccurate **OR** outdated content knowledge.
3-PK	**Academic Language:** Models and facilitates learners' use of language supports to meet academic language demands to access content. GA-TAPS-2014.1 INTASC-2013.4 INTASC-2013.5	Models and facilitates the whole class/a small group AND individual learners in using language supports to meet academic language demands to access the content.	Models and facilitates the whole class/a small group of learners in using language supports to meet academic language demands to access the content.	Models by using language supports to present academic language demands to the whole class, to a small group, or to individuals.	Does not model using language supports to present academic language demands.
4-PK	**Pedagogical Content Knowledge:** Develops learner conceptual understanding; anticipates and resolves learner misconceptions. GA-TAPS-2014.1 GA-TAPS-2014.3 INTASC-2013.5	Develops learner conceptual understanding; **AND** anticipates **AND** resolves learner misconceptions.	Develops learner conceptual understanding; **AND** anticipates **OR** resolves learner misconceptions.	Develops learner acquisition of knowledge/skills.	Does not develop learner acquisition of knowledge/skills.
INSTRUCTIONAL DELIVERY					

INDICATOR		*Advanced (4)*	*Proficient (3)*	*Developing (2)*	*Insufficient (1)*
1-ID	**<u>Learner Engagement:</u>** Engages learners in active learning by developing higher order, critical/creative thinking through inquiry-based learning promoting diverse perspectives/experiences. GA-TAPS-2014.3 INTASC-2013.8	Engages learners in active learning by developing higher order, critical/creative thinking through inquiry-based student-centered learning **AND** promotes diverse perspectives/experiences.	Engages learners in active learning by developing higher order, critical/creative thinking through teacher-facilitated learning **AND** promotes diverse perspectives/experiences.	Directs learners to acquire knowledge **AND** skills through teacher-directed learning.	Directs learners to acquire knowledge **OR** skills through teacher-directed learning.
2-ID	**<u>Use of Technology:</u>** Integrates technology to facilitate learning; involves learners in use of technology; provides rationale if technology is not used. GA-TAPS-2014.3 INTASC-2013.8	Uses appropriate technology to facilitate learning **AND** involves learners in innovative use of technology.	Use appropriate technology to facilitate learning **AND** involves learner in using technology.	Uses appropriate technology to support instruction (e.g., lesson plans, instructional materials, assessments).	Does not use technology to support instruction.
3-ID	**<u>Differentiation:</u>** Provides appropriate accommodations and/or modifications for individual learners with various levels of language development, IEP, EIP, 504, EL-TPC plans; employs principles of Universal Design for Learning (UDL)/whole group differentiation and those who require remediation/extension of learning. GA-TAPS-2014.4 INTASC-2013.2 INTASC-2013.7	Provides appropriate accommodations and/or modifications for individual learners in the class with various levels of language development, IEP, EIP, 504, EL-TPC plans; **AND** employs principles of UDL including students who require remediation/extension of learning.	Provides appropriate accommodations and/or modifications for individual learners in the class with various levels of language development, IEP, EIP, 504, EL-TPC plans; **AND** employs principles of UDL.	Provides appropriate accommodations and/or modifications for individual learners in the class with various levels of language development, IEP, EIP, 504, EL-TPC plans; does not employ principles of UDL.	Does not provide appropriate accommodations and/or modifications for individual learners in the class with various levels of language development, IEP, EIP, 504, EL-TPC plans; does not employ principles of UDL.
ASSESSMENT OF AND FOR LEARNING					
1-AL	**<u>Assessment for Learning:</u>** Uses assessment tools for both formative and summative purposes to facilitate learning and to adjust instruction. GA-TAPS-2014.5 GA-TAPS-2014.6 INTASC-2013.6 INTASC-2013.7	Uses appropriate formative/summative assessment tools to document learners' prior knowledge **AND** new learning to facilitate learning. Adjusts instruction for the whole class, groups of learners, and/or individuals.	Uses appropriate formative/summative assessment tools to document learners' prior knowledge **OR** new learning to facilitate learning. Adjusts instruction for the whole class, groups of learners, and/or individuals.	Uses appropriate formative/summative assessment tools to document learners' prior knowledge **OR** new learning to facilitate learning.	Uses inappropriate formative/summative assessment tools **OR** does not document learners' prior knowledge or new learning to facilitate learning.

INDICATOR		***Advanced (4)***	***Proficient (3)***	***Developing (2)***	***Insufficient (1)***
2-AL	**Provides/Models Feedback:** Provides feedback to learners; models use of feedback to address strengths, needs, and strategies for improvement/extension of learning. GA-TAPS-2014.6 INTASC-2013.6	Provides feedback to learners **AND** models use of feedback to address strengths, needs, **AND** strategies for improvement / extension of learning.	Provides feedback to learners **AND** models use of feedback to address strengths, needs, **OR** strategies for improvement / extension of learning.	Provides feedback to learners **OR** models use of feedback to address strengths, needs, **OR** strategies for improvement / extension of learning.	Provides superficial / insufficient feedback to learners **AND** does not model use of feedback to address strengths, needs, **OR** strategies for improvement / extension of learning.
LEARNING ENVIRONMENT					
1-LE	**Positive Learning Environment:** Promotes a positive and safe learning community. GA-TAPS-2014.7 GA-TAPS-2014.8 INTASC-2013.3	Promotes a positive (trusting, caring, and respectful) and safe learning community; facilitates learners in self-regulation, positive student-student and teacher-student interactions; maintains high expectations; and establishes a physically safe space.	Promotes a positive (trusting, caring, and respectful) and safe learning community: facilitates learners in self-regulation and positive teacher-student interactions; maintains high expectations; and establishes a physically safe space.	Promotes a positive (trusting, caring, and respectful) and safe learning environment: facilitates positive teacher-student interactions; maintains high expectations; and establishes a physically safe space.	Promotes a learning environment that does not facilitate learners in self-regulation, or positive student-student or teacher-student interactions, **AND/OR** does not establish a physically safe space.
2-LE	**Classroom Facilitation:** Maximizes learning by organizing, classroom community expectations, time, space, and materials; and by responding to disruptions in an equitable, timely manner using appropriate verbal/non-verbal communication. GA-TAPS-2014.7 GA-TAPS-2014.8 INTASC-2013.3	Maximizes learning by organizing classroom community expectations, time, space, **AND** materials; **AND** responds to disruptions using appropriate verbal/non-verbal communication.	Maximizes learning by organizing classroom community expectations, time, space, **AND/OR** materials; **AND** responds to disruptions using appropriate verbal/non-verbal communication.	Directs learning by organizing classroom community expectations, time, space, **AND/OR** materials; **AND/OR** responds to disruptions using appropriate verbal/non-verbal communication.	Does not organize classroom community expectations, time, space, **OR** materials, **AND** does not respond to disruptions using appropriate verbal/non-verbal communication.

INDICATOR		***Advanced (4)***	***Proficient (3)***	***Developing (2)***	***Insufficient (1)***
3-LE	**Classroom Safety:** Establishes and maintains a safe classroom environment ***(Science labs and physical education only).** GA-TAPS-2014.7 INTASC-2013.3	Enforces classroom, school **AND** community safety rules **AND** policies relevant to the content with written, visual, **AND** oral procedures. **Specific to science classrooms:** enforces required OSHA safety standards.	Enforces classroom, school **AND/OR** community safety rules **AND/OR** policies relevant to the content with written, visual, **AND/OR** oral procedures. **Specific to science classrooms:** enforces required OSHA safety standards.	Enforces classroom, school **OR** community safety rules **OR** policies relevant to the content with written, visual, **OR** oral procedures. **Specific to science classrooms:** enforces required OSHA safety standards.	Does not enforce, **OR** ineffectively enforces, classroom, school **OR** community safety rules **OR** policies relevant to the content with written, visual, **OR** oral procedures. **Specific to science classrooms:** enforces required OSHA safety standards.