Examining the Validity of a Student Teaching Evaluation Instrument

Sarah P. Hylton
William and Mary

Jacob D. Joseph
William and Mary

Thomas J. Ward
William and Mary

Christopher R. Gareis
William and Mary

Abstract

This study examines the validity of a student teaching evaluation instrument used in a teacher preparation program. Grounded in research-based conceptualizations of teaching and aligned to standards from relevant professional associations, the instrument is used to evaluate teacher candidates during their student teaching experience. To determine the instrument's validity, we used exploratory factor analysis and structural equation modeling to study responses from cooperating teachers, university supervisors, and student teachers. Findings partially confirm the validity of the instrument and indicate that the 30 competencies of the instrument comprise an invariant structure of four domains: planning, onstage teaching, assessment, and professionalism. Implications include instrument revision, the need for rater training, and further exploration of curricular alignment.

Keywords: evaluation instrument, student teaching, validity

Initial preparation to undertake the work of teaching relies on authentic practice in field-based settings (AACTE Clinical Practice Commission, 2018; Ball & Forzani, 2009; Darling-Hammond, 2006; Zeichner, 2010). Meaningful field-based experiences provide opportunities for student teachers (STs) to apply what they have learned during coursework, and they present situations that prompt and require STs' continued learning (Zeichner, 2010). Teacher preparation programs (TPPs) rely on partner K-12 schools to provide a setting for such clinical experiences and on cooperating teachers (CTs) and university supervisors (USs) to provide mentoring, coaching, and evaluation during that experience (Clarke, et al., 2014; Gareis & Grant, 2014).

Although mentoring is the central task of being a CT (Gareis & Grant, 2014), CTs also serve in an evaluative capacity, providing summative judgments of the quality of STs' performance and their potential to be successful teachers (Wang et al., 2003). CTs work in conjunction with USs, who represent the TPP and its affiliated interests, namely, ensuring that STs are developing as autonomous professionals who will be prepared to join the teaching profession as novice, professionally credentialed educators.

Providing CTs and USs with the tools necessary to enact their evaluative roles is an obligation for all TPPs. If STs' effectiveness is to be meaningfully evaluated during their clinical experiences, then the instrument used to make that determination must be sound. As Bryant et al. (2016) assert, the "assessment of pre-service teachers' performance in the field must include…the assurance that the assessment is valid and reliable" (p. 81). Accreditation bodies, too, emphasize the intentional development and consistent use of valid evaluation instruments (Council for the Accreditation of Educator Preparation, 2013; Sandholtz & Shea, 2012). High-stakes consequences associated with summative evaluations, such as final grades, credentialing, and hiring decisions, compound the need for trustworthy evaluation instruments (Clarke et al., 2014).

Despite the centrality of evaluation instruments in judging ST performance, the quality of such tools may be insufficient to summatively judge the performance of STs (Clarke et al., 2014), particularly with regard to their validity and reliability (Bryant et al., 2016; Choi et al., 2016). Indeed, Richmond et al. (2019) characterize the development and validation of "informative, scalable, and accepted" instruments for assessing ST performance as "a persistent challenge facing teacher education" (p. 86). Therefore, the primary purpose of this study was to investigate the validity of the student teaching evaluation instrument currently used in our

TPP. Because this study provides evidence of such validation, our methodology and findings may serve as a model for other TPPs attempting to design, revise, and/or validate their evaluation instruments.

History and Evolution of the Student Teaching Evaluation Instrument

The evaluation instrument investigated in this study is currently used by the TPP at a mid-sized public university in Virginia. The current instrument was initially developed in 2001–2002 when the TPP convened a committee to study the student teaching standards as outlined by relevant professional associations, including the Interstate Teacher Assessment and Support Consortium (InTASC), the National Council for the Accreditation of Teacher Education (NCATE), the National Council of Teachers of English (NCTE), the National Council of Teachers of Math (NCTM), the National Council of Teachers of Science (NCTS), and the Council for Exceptional Children (CEC). The committee's work was also influenced by Stronge's (2002) research-based review of qualities of effective teachers and Danielson's (1996) Framework for Teaching.

After completing a crosswalk of these professional standards, frameworks, and guiding conceptual elements, the committee, comprised of general education, special education, gifted education, and educational leadership faculty members, established an agreed-upon set of competencies which represent the knowledge, skills, and dispositions STs should develop throughout their coursework and field experiences. A panel of Clinical Faculty (CTs in our affiliated K-12 schools who have taken a master's level course in CT preparation) also reviewed the competencies to ensure that they reflected the skills necessary for effective teaching and were readily understood by our K-12 partners.

Ultimately, the committee proposed an evaluation instrument of 30 competencies divided into six domains (see *Appendix*): Foundational Understanding; Ability to Plan, Organize, and Prepare to Teach; Teaching Skills; Assessment and Evaluation of Learning; Classroom Management Knowledge and Skills; and Professional Knowledge and Skills. The instrument offers performance indicators as illustrative examples of each competency; however, these indicators are not intended as an explicit list of required behaviors. Rather, the instrument allows for the possibility that an evaluator might not observe any of the performance indicators but may note other, equally valid indications of the demonstration of a competency.

Revisions to the Competencies

Over time, the original 30 competencies have undergone both minor and substantive revisions. Slight changes in wording have constituted most of the minor revisions. For instance, Competency 26 originally stated the expectation that a ST "participates in professional development" (School of Education, 2002) but was later revised to clarify that a ST "participates in *and applies* [emphasis added] professional development" (School of Education, 2016).

In 2009, an ad hoc committee in the TPP made more substantive revisions to the competencies to ensure that they reflect both the college's Diversity Statement and the TPP's commitment to preparing STs who are dedicated to advancing diversity, equity, and inclusion. Grounded in the belief that respect for and attention to diversity, equity, and inclusion pervade a teacher's core responsibilities, the committee decided to incorporate these attributes throughout the existing competencies rather than creating additional competencies or an additional domain. For instance, Competency 5 originally stated the expectation that a ST "demonstrates an understanding of the purposes and roles of K-12 education" (School of Education, 2002) but was revised to clarify the expectation that a ST "demonstrates an understanding of the purposes and

roles of PreK-12 education *in a diverse and inclusive democratic society* [emphasis added]" (School of Education, 2016). The TPP again consulted with Clinical Faculty from partnering K-12 schools to solicit feedback and suggestions on these changes.

Revisions to the Rating Scale

The TPP has also made revisions to the rating scale of the evaluation instrument. Initially, the evaluation instrument enumerated three performance levels: "Below Expectations," "Meets Expectations," and "Exceeds Expectations." Over several years of the instrument's use, USs shared that some of the competencies addressed less visible teaching tasks and were not always evident during their observations. Similarly, CTs argued that other of the competencies addressed theoretical constructs which were sometimes more challenging to observe in applied practice. Based on these concerns, "Unable to Observe" was included as an element in the rating scale beginning in 2004. In 2013, Clinical Faculty from the TPP's partnership schools voiced concerns that the scoring system did not provide a means of acknowledging STs who were moving in the right direction but not yet meeting expectations. Supported by empirical evidence (Gareis & Grant, 2014), the TPP added "Developing" as an option between "Below Expectations" and "Meets Expectations."

Uses of the Instrument

The evaluation instrument is not an observation tool. Rather, it conveys accumulated judgment about a ST's performance based on multiple sources of information, including observations and coaching conversations. The evaluation instrument is used formatively at the midpoint of the student teaching experience and summatively at the end of that student teaching experience. At both intervals in the term, the instrument is completed by the ST, the CT, and the US, thus creating a total of six evaluations of a ST's performance. The CT and US's evaluations,

combined with the ST's self-evaluation, are used formatively to provide clarity about areas for continued growth. Summatively, the instrument leads to a final judgment regarding a ST's overall teaching effectiveness, resulting in a grade of pass or fail in the student teaching experience.

Methodology and Results

Grounded in professional standards and considerable stakeholder feedback, the current evaluation instrument was created with substantial face validity. Therefore, the intent of this study was to investigate the validity of the instrument using empirical methodology, namely factor analysis. Specifically, we sought to answer two questions:

1. To what degree does a consistent factor structure emerge from the rating data?
2. To what degree does that structure reflect the a priori six-factor structure upon which the instrument was theoretically constructed?

Sample

The data for the current investigation were extracted from the ratings of three cohorts of STs over a six-semester timeframe. The midterm and final ratings of USs, CTs, and STs were examined for a common factorial model. The data set included 1,486 cases with complete rating scales which were used in the analyses. Table 1 displays the number of cases at each time period for each group.

Table 1

Sample Size by Time and Rater

Rater	Midterm	Final
Cooperating Teacher	262	237
University Supervisor	250	251
Student Teacher	256	230

Analyses

Because the TPP developed the instrument using a theoretical model, we conducted initial analyses using confirmatory factor analysis (CFA), treating the item-level rating data as ordinal inputs. We conducted CFAs in MPus v7 (Muthén & Muthén, 2012) using diagonally weighted least squares (WLSMV) as the estimator. WLSMV is specifically designed to handle ordinal data (Li, 2016). We used three fit measures to judge the fit of our models, including the root mean square error of approximation (RMSEA), confirmatory fit index (CFI), and weighted root mean square residual (WRMR). We considered a model to fit well if the RMSEA was at or below .05, CFI was greater than .95, and WRMR was lower than 1 (DiStefano, 2016; Hu & Bentler, 1999; & MacCallum et al., 1996). When using WLSMV, typical chi-square difference testing cannot be conducted in the usual manner (Muthén & Muthén, 2012). Therefore, we relied on the fit indices as a guide for better models.

First, we conducted two CFAs for each group (i.e., STs, USs, and CTs) at each time point (i.e., midterm and final). The initial CFA tested a unidimensional model, and the second CFA tested the proposed correlated six-factor model. The unidimensional model was estimated to provide a comparison point for the six-factor model. Table 2 presents the results of the unidimensional and six-factor models. As suspected, the unidimensional model did not fit the

data well. The six-factor model was an improvement over the unidimensional model in all instances as indicted by the lower chi-square value and fit indices that achieved or approximated the established thresholds for good fit. However, four of the six-factor models evidenced a non-positive definite covariance matrix, rendering those solutions inadmissible. Based on the magnitude of the correlations among the factors, we tested both second-order and bifactor models for each group. Neither of these models fit well in more than one group.

Table 2

CFA Results for the Unidimensional and Six-Factor Models

Model	df	$\chi 2$	RMSEA	CFI	WRMR
US Midterm					
Unidimensional	405	1325.24	0.09	0.88	1.85
Six Factors	390	865.36	0.07	0.94	1.35[a]
ST Midterm					
Unidimensional	405	943.28	0.07	0.94	1.40
Six Factors	390	683.72	0.05	0.96	1.06[a]
CT Midterm					
Unidimensional	405	1085.49	0.08	0.94	1.48
Six Factors	390	735.94	0.05	0.97	1.06
US Final					
Unidimensional	405	990.24	0.07	0.94	1.42
Six Factors	390	744.68	0.06	0.96	1.12[a]
ST Final					
Unidimensional	405	799.48	0.06	0.95	1.27
Six Factors	390	602.04	0.04	0.97	1.01
CT Final					
Unidimensional	405	825.64	0.06	0.96	1.22
Six Factors	390	611.41	0.04	0.98	0.95[a]

Note: The latent variable covariance matrix was not positive definite.

Before testing for a common structure across groups, researchers typically demonstrate that the proposed structure fits the groups independently (Byrne, 2016). The lack of an

acceptable solution for the six-factor model in four of the CFAs argues against a common structure. We therefore applied an exploratory factor analysis (EFA) and structural equation modeling (SEM) approach known as Exploratory Structural Equation Modeling (ESEM) (Asparouhov & Muthén, 2009) to the data sets to determine whether a consistent factor pattern could be identified. ESEM incorporates the advantages of the less restrictive EFA and the more advanced CFA (including tests of model fit) at the same time. ESEM has shown to result in improved model fit and deflated inter-factor correlations compared to EFA (Asparouhov & Muthén, 2009; Marsh et al., 2014). The ESEMs were conducted in MPus v7 (Muthén & Muthén, 2012) using WLSMV as the estimator with geomin rotation. We examined the CFI, RMSEA, and standardized root mean square residual (SRMR) as indicators of model fit. We considered a model to be a good fit if the CFI was greater than .95, the RMSEA was at or below .05, and the SRMR was less than .05 (Hu & Bentler, 1999; MacCallum et al., 1996). Although two of the analyses (ST and CT final) indicated an acceptable three-factor model, the requirement of an acceptable model across all six samples indicated a four-factor solution was necessary.

A common four-factor model was created by randomly selecting 50 cases from each of the six groups and conducting a constrained four-factor ESEM using WLSMV as the estimator with geomin rotation. The use of 300 cases ensured sufficient power in the analysis without over reliance on any particular sample. Table 3 presents the final factor solution with significant factor loadings in bold. The loadings shows that some of the original factors (e.g., “Professional Dispositions” and “Assessment and Evaluation for Learning”) were maintained as distinct, other factors were merged (e.g., “Teaching Skills” and “Classroom Management Knowledge and Skills”), and some factors had their competencies distributed over new factors (e.g., “Foundational Understanding”). At the competency level, some competencies

had weak association to the new factors (specifically, Competencies 3 and 5), and there was distinct cross loading for other competencies (Competencies 7, 19, and 20). Table 4 presents the factor reliabilities and correlations. The values indicate that the new factors have substantial reliability and, as expected, correlate significantly with each other.

Table 3

Common Factor Solution Significant Loadings in Bold

Domain and Competency	1 Onstage Teaching	2 Professionalism	3 Planning	4 Assessment
Foundational Understanding				
1. Demonstrates understanding of subject matter and pedagogical knowledge for instruction.	.090	.113	**.440**	.075
2. Demonstrates understanding of how students learn and develop and provides learning opportunities that support students' intellectual, social, and personal development.	**.601**	-.040	.102	.115
3. Demonstrates understanding of the central role of language and literacy in student learning.	.192	.158	**.253**	.187
4. Demonstrates understanding of how all students differ in their experiences and their approaches to learning.	**.667**	.038	-.077	.133
5. Demonstrates an understanding of the purposes and roles of PreK-12 education in a diverse and inclusive democratic society.	.077	**.300**	.142	.174
Ability to Plan, Organize, and Prepare for Teaching				
6. Plans lessons that align with local, state, and national standards.	-.244	.046	**.919**	.025
7. Selects appropriate instructional strategies/activities aligned to instructional	**.505**	-.194	**.467**	.014

goals and responsive to diverse student needs.				
8. Selects appropriate materials/resources aligned to instructional goals and that are reflective of diverse perspectives.	.179	-.121	**.641**	.008
Teaching Skills				
9. Teaches based on planned lessons.	.048	.115	**.592**	.058
10. Provides for individual differences.	**.607**	.013	-.091	.206
11. Uses motivational strategies to promote learning for all students.	**.871**	.030	-.169	.004
12. Engages students actively in learning.	**.737**	-.007	.088	-.071
13. Uses a variety of effective teaching strategies.	**.532**	-.050	.150	.054
14. Helps students develop thinking skills that promote learning.	**.505**	.064	.094	.151
15. Monitors student learning.	**.354**	.106	.038	.105
Assessment and Evaluation for Learning				
16. Creates and selects appropriate assessments for learning.	.097	-.103	.139	**.668**
17. Implements assessments for learning.	-.062	.049	.025	**.852**
18. Interprets/uses assessment results to make instructional decisions.	.060	.117	-.031	**.672**
Classroom Management Knowledge and Skills				
19. Builds positive rapport with and among students, fostering an environment that values and encourages respect for diversity.	**.572**	**.350**	-.090	-.199
20. Organizes for effective teaching.	**.323**	.135	**.347**	-.018
21. Demonstrates use of effective routines and procedures.	**.522**	.118	.190	-.033
22. Demonstrates efficient and effective use of time.	**.549**	.060	.178	-.005

23. Maintains a physically and emotionally safe learning environment for all students.	**.433**	.172	.076	-.135
24. Responds appropriately and equitably to student behaviors.	**.646**	.117	-.098	.043
Professional Dispositions				
25. Demonstrates professional demeanor and ethical behavior.	.073	**.620**	.156	-.183
26. Participates in and applies professional development.	-.123	**.618**	-.064	.173
27. Demonstrates effective oral and written communication.	-.018	**.620**	.159	.040
28. Reflects actively and continuously upon practice, leading to enhanced teaching and learning for all students.	.052	**.443**	.213	.072
29. Cooperates, collaborates, and fosters relationships with families and other members of the community.	.127	**.706**	-.181	.061
30. Demonstrates potential for teacher leadership.	.122	**.461**	-.006	.059

Table 4

Factor Correlations and Omega Reliabilities

Factor/Domain	**Onstage Teaching**	**Professionalism**	**Planning**	**Assessment**
Onstage Teaching	.94			
Professionalism	.74	.89		
Planning	.78	.65	.87	
Assessment	.68	.61	.67	.85

Note: Values in the diagonal are omega reliability coefficients.

Discussion

The original six-factor model classified the stated competencies into six domains. The study results, however, indicate that these competencies are more appropriately grouped in a common four-factor model. Our review and analysis of this outcome led us to label the four new factors as "Professionalism," "Assessment," "Onstage Teaching," and "Planning."

Professionalism and Assessment

Within the common four-factor model, two of the original domains ("Professional Dispositions" and "Assessment and Evaluation for Learning") remained distinct, with all the competencies initially ascribed to those domains continuing to correlate with those factors. That is, Competencies 25–30 continue to group as "Professionalism," and Competencies 16–18 as "Assessment" in the common four-factor model. Essentially, these results indicate that raters using the instrument operationally conceptualize each of these sets of competencies as unified factors (or "domains" in the language of the TPP). Both Stronge (2002) and Danielson's (1996)

frameworks emphasize similar sets of competencies with regard to professional competencies, and Stronge also includes a similar set of competencies with regard to assessment.

We submit that the clear, unequivocal language used in the competencies comprising these two domains helps to distinguish the competencies as clearly belonging to their respective factors. For instance, the word "assessment" is used in each of the competencies affiliated with that factor, and the order of the three competencies suggests an assessment process. Likewise, the competencies associated with "Professionalism" use words such as "professional," "communicate," "cooperate," "collaborate," "reflect," "relationship," and "leadership," which may signal their inclusion in that factor.

Onstage Teaching

Both Stronge (2002) and Danielson (1996) maintain separate domains for instruction and classroom environment; our original six-factor model did as well. However, ESEM confirms that the raters using our instrument do not distinguish between teaching and classroom management as separate factors but rather perceive them to be two parts of one whole, suggesting that the raters conceive of classroom management as integral to effective instruction. These results lead us to conclude that the tasks encompassed by the competencies in these particular domains are those teacher behaviors that occur during class time and are thus most visible to others. We have labeled these readily visible tasks "Onstage Teaching" as a means of differentiating them from the tasks of teaching that occur when teachers are not actively working with a class of students (e.g., planning, reflection, assessment, feedback, etc.), tasks which Macfarlane (2007) terms "offstage" (p. 49). Although classroom management's inclusion as a subset of "Onstage Teaching" could suggest a diminished view of

its importance, we contend that, at least for the raters using this instrument, classroom management is perceived as essential to, and inseparable from, effective instruction.

Redistributed Competencies

In the common four-factor model, several of the competencies redistributed to factors other than those to which they were originally assigned. This is the case with the first five competencies which were grouped in the original six-factor model as "Foundational Understanding." Unlike the other five original domains, which encompass observable competencies, this domain focuses more on understandings that are developmental and foundational. The ESEM results, however, indicate that raters using this instrument do not perceive these five competencies as comprising a unified factor. Instead, the competencies distributed across three other factors in the common four-factor model. Competency 1 redistributed to "Planning," a change which might be explained by the competency's language about understanding subject area content and pedagogy. Competencies 2 and 4 correlated strongly with "Onstage Teaching." Competency 2 calls for STs to "provide learning opportunities," and Competency 4 requires STs to demonstrate their understanding of student differences. In both cases, the language suggests instruction, an idea supported by raters' perception that these two competencies belong in "Onstage Teaching."

Competency 3, which focuses on a ST's understanding of the central role of language in learning, did not correlate strongly with any factor. Given that the competency appears to be an outlier, there are three possible explanations: Either (a) the item is not relevant to effective teaching, (b) it *is* relevant but unlike any of the other competencies, or (c) its meaning is unclear. We contend that the second explanation best applies: Though unique, this item is

foundationally imperative as it emphasizes the need for STs to understand that language, as the primary means by which people express thought, is essential to all teaching and learning.

Competency 5, which addresses the role of public education in a democratic society, also did not correlate strongly with any factor. As an apparent outlier, the same three explanations are possible: irrelevant, unique, or unclear. We hold that two of these are likely: the item is both unclear and unlike other items. Our review of Competency 5 suggests that the use of multiple conceptual terms creates a complex statement whose meaning may not be readily apprehended by the raters using the instrument. In addition to being linguistically complex, Competency 5 is also unique. We contend, however, that its focus on understanding the purpose and role of public education, though unlike any of the other competencies, is nonetheless relevant to effective teaching.

Shared Competencies

In three cases, a competency is "shared" between two of the new factors. Competencies 7 and 20 are shared between "Onstage Teaching" and "Planning." Competency 7, originally allocated to the "Ability to Plan, Organize, and Prepare for Teaching" domain, still has a strong connection to "Planning" but aligns even more strongly with "Onstage Teaching." The language of the competency is mixed, emphasizing the planning domain with reference to selecting appropriate strategies and emphasizing the instructional domain with reference to instructional strategies and activities. Competency 20, initially allocated to "Classroom Management Knowledge and Skills," has a weak relationship to both "Onstage Teaching" and "Planning." The shortest of all the competencies in terms of wording, it nonetheless still has double-barreled language that might cause it to group with either of those categories. For instance, the word "organizes" suggests planning whereas "effective teaching" indicates

alignment with "Onstage Teaching." Competency 19, originally allocated to "Classroom Management Knowledge and Skills," still aligns most strongly with "Onstage Teaching;" however, the results indicate that some raters consider this competency to more appropriately fit with "Professionalism." Phrases such as "positive rapport" and "environment" may create a problem with clarity about the intention of the competency. In all cases, the mixed message of the competency may explain why raters perceive it as grouping with two possible factors.

Diversity: An Integrated Construct

As noted previously, significant changes were made to the wording of the competencies in 2009 to reflect the university's Diversity Statement. The intentional decision to incorporate language that addresses diversity, equity, and inclusion throughout the instrument rather than creating an additional domain for diversity is supported by the ESEM results. Despite terms and phrases such as "diverse," "diversity," "inclusive," "equitably," and "all students" appearing in ten of the competencies, these competencies did not correlate in the common four-factor model to create a fifth factor. Rather, with the exception of Competency 5, they all remained in factors similar to their original domains, evidence that affirms the committee's decision.

Implications, Limitations, and Recommendations

The empirical methodology of this study has provided important insights into a judgment of the validity of this student teaching evaluation instrument. Here we explore three pragmatic implications that extend from the findings.

Revision of the Instrument

The instrument could be strengthened through further revision, both in terms of the competencies themselves as well as the arrangement of those competencies into particular domains. The clarity of several of the competencies is problematic, particularly those that are

redistributed to new factors (Competencies 1, 2, and 4), those that are shared among the four factors (Competencies 7, 19, and 20), and those that are weak contributors to any factor (Competencies 3 and 5). Revision of these competencies for greater clarity and precision is certainly warranted, with these revisions potentially leading to stronger correlation of the various competencies to the four factors. Such revisions should draw on updated professional standards and may also consider the updated models of both Stronge and Danielson. Rearranging the competencies into the four factors indicated by the common four-factor model is another means of improving the validity of the instrument as restructuring the competencies to represent the findings of this study may provide a more integrated view of the act of teaching.

Another possible revision might be to consider how to continue to elevate respect for and attention to diversity, equity, and social justice. Guided by the belief that these principles pervade all teaching responsibilities, the 2009 revision committee chose to embed them throughout the competencies rather than to create a separate domain for them. As noted earlier, this decision appears to be supported by the data from this study. However, the phrasing of the competencies themselves might be strengthened to better reflect the university's commitment to diversity, equity, and social justice. Faculty have recently adopted a more robust Diversity Statement which explicitly addresses antiracism and social justice, and we anticipate that this new statement will prompt further discussion about the evaluation instrument and how it might be revised to ensure that our TPP graduates are committed to these principles.

Training

Regardless of what revisions are made, efforts to promote a common understanding among all raters of the expectations of STs, as articulated by the competencies, is imperative (Bryant et al., 2016; AACTE, 2018). Systematic training for all who use the instrument would

provide a means of achieving this common understanding. A study of the TPP's Clinical Faculty Program, which offers training for teachers recruited to serve as CTs, concluded that explicitly, systematically, and intentionally training CTs to understand and utilize the student teaching evaluation instrument resulted in more accurate evaluations of STs (Gareis & Grant, 2014).

Currently, USs and STs in our TPP do not receive similar explicit, systematic, or intentional training on the rationale for, construction of, and use of the instrument. Extending this training to all rater groups would likely improve rater clarity about the intentions of the competencies and increase the likelihood that the language and structure of the instrument are not barriers to applying it as intended.

Teacher Preparation Program Curriculum

Finally, given the interdependent nature of assessment and curriculum (Gareis & Grant, 2015; Wiggins & McTighe, 2005), our study of this student teaching evaluation instrument must consider the implications of that assessment for the program's curriculum. Given the results from this study, we question whether the instrument is aligned with the TPP's scope and sequence of coursework. The question of curricular alignment is particularly salient for the two competencies that did not correlate strongly with any of the four factors and raises questions about whether these elements of teaching are receiving sufficient attention in the curriculum. These findings, then, imply that policy decisions regarding curriculum must be framed by the instrument and, conversely, that revisions to the instrument should shape conversation regarding the curriculum.

Limitations and Recommendations

This study is subject to several limitations. First, because the study is specific to this TPP, its results are not generalizable. Furthermore, the study does not investigate the perspectives of those stakeholders who use the instrument or their reasoning behind their conceptualizations of

the instrument, nor does it consider the influence of the stakeholders' demographics, such as years of experience or previous contextual experiences. The study is also limited in that it does not investigate the impact of potential differences between the midterm evaluation (used formatively) and the final evaluation (used summatively). Nonetheless, this study may provide instructive insight to other TPPs regarding their own instrument development and validation.

Future research might focus on a qualitative or mixed method study that further investigates the perspectives and conceptualizations of stakeholders who use the instrument and the role of demographics in influencing their use of the instrument.

Conclusion

The creation of this student teaching evaluation instrument began with qualitative consideration of the constructs for which it is intended to provide evidence. The work, undertaken by a committee of faculty experts drawing upon professionally recognized standards and frameworks, culminated in a conceptual framework of 30 competencies organized into six domains. After more than 15 years of use and refinement of the instrument, this quantitative study sought to examine how those same competencies and domains manifest in the student teaching evaluation instrument through its use by the three rater groups. Although findings indicate that the three rater groups do not perceive the competencies as holding together in the same way the committee originally conceived of them, there is an invariant structure close to the original upon which the three rater groups all agree. Although there are implications in these differences, the instrument nonetheless serves to provide reasonably valid and reliable evidence of STs' competencies. With attention to minor revisions, expanded training opportunities, and closer alignment of components of the TPP curriculum to the evaluation criteria, the TPP can

increase the validity of the student teacher evaluation instrument and the efficacy of the inferences and actions resulting from its use.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

AACTE Clinical Practice Commission. (2018). *A pivot toward clinical practice, its lexicon, and the renewal of educator preparation: A report of the AACTE clinical practice commission* [White paper]. Retrieved January 29, 2018. from AACTE: file:///C:/Users/author/Downloads/cpc-full-report-final.pdf

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60*, 497–511. https://doi.org/10.1177/0022487109348479

Bryant, C. L., Maarouf, S., Burcham, J., & Greer, D. (2016). The examination of a teacher candidate assessment rubric: A confirmatory factor analysis. *Teaching and Teacher Education*, *57*, 79–96. https://doi.org/10.1016/j.tate.2016.03.012

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.

Council for the Accreditation of Educator Preparation. (2013). *CAEP accreditation standards*. http://caepnet.org/standards/introduction

Choi, H., Benson, N. F., & Shudak, N. J. (2016). Assessment of teacher candidate dispositions: Evidence of reliability and validity. *Teacher Education Quarterly*, *43*(3), 71—89.

Clarke, A., Triggs, V., & Nielsen, W. (2014). Cooperating teacher participation in teacher education: A review of the literature. *Review of Educational Research*, *84*, 163–202. https://doi.org/10.3102/0034654313499618

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. ASCD.

Darling-Hammond, L. (2006). Constructing 21st century teacher education. *Journal of Teacher Education*, *57,* 300–314. https://doi.org/10.1177/0022487105285962.

DiStefano, C. (2016). Examining fit with structural equation models. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advances* (pp. 166–196). Hogrefe.

Gareis, C. R., & Grant, L. W. (2014). The efficacy of training cooperating teachers. *Teaching and Teacher Education*, *39*, 77–88. https://doi.org/10.1016/j.tate.2013.12.007

Gareis, C. R., & Grant, L. W. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning* (2nd ed.). Eye on Education.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118

Li, C. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Macfarlane, B. (2007). Beyond performance in teaching excellence. In A. Skelton (Ed.), *International perspectives on teaching excellence* (pp. 48–59). Routledge.

Marsh, H., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*(1), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide statistical analysis with latent variables* (7th ed.). Author.

Richmond, G., Salazar, M., & Jones, N. (2019). Assessment and the future of teacher education. *Journal of Teacher Education, 70*(2), 86-89. https://doi.org/10.1177/0022487118824331

Sandholtz, J. H., & Shea, L. M. (2012). Predicting performance: A comparison of university supervisor predictions and teacher candidates' scores on a teaching performance assessment. *Journal of Teacher Education*, *63*, 39-50. https://doi.org/10.1177/0022487111421175–

School of Education. (2002). *Handbook for practica & student teaching experiences*.

School of Education. (2016). *Handbook for practica & student teaching experiences*.

Stronge, J. H. (2002). *Qualities of effective teachers*. ASCD.

Wang, A. H., Coleman, A. B., Coley, R. J., & Phelps, R. P. (2003*). Preparing teachers around the world*. Educational Testing Service.

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (Exp. 2nd ed.). ASCD.

Zeichner, K. (2010). Rethinking the connections between campus courses and field experiences in college- and university-based teacher education. *Journal of Teacher Education, 61*, 89–99. https://doi.org/10.1177/0022487109347671

Appendix

Domains and Competencies of the Student Teaching Evaluation Instrument

Domain 1: Foundational Understanding

1. Demonstrates understanding of subject matter and pedagogical knowledge for instruction.
2. Demonstrates understanding of how students learn and develop and provides learning opportunities that support students' intellectual, social, and personal development.
3. Demonstrates understanding of the central role of language and literacy in student learning.
4. Demonstrates understanding of how all students differ in their experiences and their approaches to learning.
5. Demonstrates an understanding of the purposes and roles of PreK-12 education in a diverse and inclusive democratic society.

Domain 2: Ability to Plan, Organize, and Prepare for Teaching

6. Plans lessons that align with local, state, and national standards.
7. Selects appropriate instructional strategies/activities aligned to instructional goals and responsive to diverse student needs.
8. Selects appropriate materials/resources aligned to instructional goals and that are reflective of diverse perspectives.

Domain 3: Teaching Skills

9. Teaches based on planned lessons.
10. Provides for individual differences.
11. Uses motivational strategies to promote learning for all students.
12. Engages students actively in learning.
13. Uses a variety of effective teaching strategies.
14. Helps students develop thinking skills that promote learning.
15. Monitors student learning.

Domain 4: Assessment and Evaluation for Learning

16. Creates and selects appropriate assessments for learning.
17. Implements assessments for learning.
18. Interprets/uses assessment results to make instructional decisions.

Domain 5: Classroom Management Knowledge and Skills

19. Builds positive rapport with and among students, fostering an environment that values and encourages respect for diversity.
20. Organizes for effective teaching.
21. Demonstrates use of effective routines and procedures.
22. Demonstrates efficient and effective use of time.
23. Maintains a physically and emotionally safe learning environment for all students.

24. Responds appropriately and equitably to student behavior.

Domain 6: Professional Dispositions

25. Demonstrates professional demeanor and ethical behavior.
26. Participates in and applies professional development.
27. Demonstrates effective oral and written communication.
28. Reflects actively and continuously upon practice, leading to enhanced teaching and learning for all students.
29. Demonstrates potential for teacher leadership.
30. Cooperates, collaborates, and fosters relationships with families and other members of the community.