Teaching and Teacher Education 59 (2016) 1–12

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Teacher candidate performance assessments: Local scoring and implications for teacher preparation program improvement

Kevin C. Bastian [a, *], Gary T. Henry [b], Yi Pan [c], Diana Lys [d]

[a] University of North Carolina at Chapel Hill, Education Policy Initiative at Carolina, Abernathy Hall, Campus Box 3279, Chapel Hill, NC 27599, USA
[b] Vanderbilt University, Peabody College, PMB #414, 230 Appleton Place, Nashville, TN 37203, USA
[c] Frank Porter Graham Child Development Institute, University of North Carolina at Chapel Hill, 105 Smith Level Road, Chapel Hill, NC 27516, USA
[d] 101Peabody Hall, Campus Box 3500, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

HIGHLIGHTS

- Locally-scored performance assessments partially-aligned with construct framework.
- Locally-scored performance assessments systematically higher than official scores.
- Locally-scored performance assessments significantly predict teacher outcomes.
- Candidate performance assessments may inform evidence-based program improvement.

ARTICLE INFO

Article history:
Received 5 October 2015
Received in revised form
6 May 2016
Accepted 16 May 2016

Keywords:
Performance assessments
Evidence-based program improvement
Construct validity
Reliability
Predictive validity

ABSTRACT

Locally-scored teacher candidate performance assessments offer teacher preparation programs (TPPs) formative performance data, common language and expectations, and information to guide program improvements. To best use these data, TPPs need to understand the validity and reliability of local scoring and assess whether scores predict candidates' performance as teachers. Examining locally-scored performance assessments, we find that local scores are significantly higher than official scores. However, local scores identify three factors partially-aligned with the assessment's construct blueprint and significantly predict teachers' performance outcomes. These analyses provide a framework for research and highlight the utility of locally-scored performance assessments for evidence-based TPP improvement.

1. Introduction

In recent years, public concern for the quality of teachers and teacher education has pushed policymakers and accreditation agencies in the United States to hold teacher preparation programs (TPPs) accountable for the effectiveness of their graduates (Crowe, 2011). For example, shortly after the implementation of the federal No Child Left Behind act in 2002, states such as Louisiana, North Carolina, and Tennessee initiated efforts to link teachers' value-added scores to the TPP from which they graduated (Bastian, Patterson, & Yi, 2015; Gansle, Noell, & Burns, 2012 Henry et al., 2011; Henry, Thompson, Fortner, Zulli, & Kershaw, 2010; Noell & Burns, 2006; Noell, Porter, Patt, & Dahir, 2008; TSBE, 2012, 2013). In 2009, the Race to the Top grant competition mandated that states seeking federal funds commit to publicly reporting TPP's effectiveness on value-added measures and closing low performing TPPs (Crowe, 2011; Henry, Kershaw, Zulli, & Smith, 2012). More recently, the United States Department of Education has proposed regulations that would require TPPs to report a variety of performance measures, including the learning outcomes for graduates' K-12 students (Federal Register 2014-28218, 2014). Likewise, the Council for the Accreditation of Educator Preparation (CAEP), the national accrediting body for educator preparation programs, requires TPPs to demonstrate the impact of their graduates on student learning,

* Corresponding author. Department of Public Policy, Education Policy Initiative at Carolina, University of North Carolina at Chapel Hill, Abernathy Hall, Campus Box 3279, Chapel Hill, NC 27599, USA.
E-mail addresses: kbastian@email.unc.edu (K.C. Bastian), gary.henry@vanderbilt.edu (G.T. Henry), yi.pan@unc.edu (Y. Pan), lys@unc.edu (D. Lys).

http://dx.doi.org/10.1016/j.tate.2016.05.008
0742-051X/

classroom instruction, and employer satisfaction (CAEP 2013).

In response to these policies and the desire of teacher educators to prepare more effective beginning teachers, TPPs have begun to reform their preparation practices and engage in continuous improvement efforts. Given the current policy context and its focus on the achievement scores of students taught by TPP graduates, it appears that the success of these reforms will be judged, at least in part, on the value-added scores of TPP graduates. By themselves, however, teacher value-added scores are insufficient to guide TPP reforms for two reasons. First, value-added scores come too late to guide TPP improvement efforts—there are often several years between teacher candidates' preparation and their entry into the workforce and impact on student learning. Second, while measuring one aspect of teachers' effectiveness, the value-added scores of program graduates do not provide information about specific teaching practices that would allow TPP faculty and staff to identify programmatic strengths and weaknesses. While some states and school districts use multiple measures—value-added, classroom observations, evaluation ratings—to assess teacher performance, these still suffer from the first problem—they come too late to guide TPP improvement.

To best drive program improvement efforts, TPPs need data on the performance of their candidates that is timely, identifies multiple domains of teaching effectiveness, and significantly predicts outcomes for teachers-of-record. At least one study suggests that many traditional measures of teacher candidate performance, such as grade point average, licensure exam scores, dispositional ratings, and student teaching ratings, do not meet all these criteria and thus, may be of limited use for evidence-based program improvement (Henry et al., 2013). In recent years, however, many teacher educators have supported the creation and widespread adoption of teacher candidate performance assessments—one of which has received widespread attention, the Teacher Performance Assessment (TPA). The TPA is a portfolio completed by teaching candidates during their student teaching experience that uses video clips of instruction, lesson plans, student work samples, and candidates' reflective commentaries to examine candidates' ability to effectively plan for instruction, teach in their content area, and assess both students and their own teaching. These assessments are scored using rubrics that have been field tested for reliability (Stanford Center for Assessment, Learning, and Equity (SCALE, 2013).[1]

While teacher candidate performance assessments could be used as a high-stakes measure for certification and/or program completion decisions (Duckor, Castellano, Tellez, Wihardini, & Wilson, 2014), performance assessments that are locally-scored by TPP faculty and staff may inform evidence-based program improvement efforts. As argued by Peck and colleagues, locally-scored performance assessments provide TPP faculty and staff with: (1) a common language for discussing candidates' performance; (2) common expectations for teacher candidate performance; (3) a forum for accepting collective responsibility for teacher candidate performance in which reforms to improve preparation practices can be developed; and (4) direct evidence of the extent to which teacher candidates demonstrate specific knowledge and skills expected by TPP faculty and staff (Peck, Singer-Gabella, Sloan, & Lin, 2014). Essentially, locally-scored performance assessments represent a promising measure for evidence-based program improvement.

Despite this promise, TPPs can best rely on evidence from locally-scored performance assessments when the scores: (1) measure the constructs that they were designed to measure (construct validity); (2) are reliably scored by different individuals (reliability); and (3) predict teacher candidates' performance as classroom teachers (predictive validity) (Admiraal, Hoeksma, van de Kamp, & van Duin, 2011). Extant research suggests that teacher candidate performance assessments, like TPA, can be the fulcrum that leverages an evidence-based culture; however, without data that are valid, reliable, and predict outcomes of interest, the evidence provided by locally-scored performance assessments may not guide TPPs to adopt more effective preparation practices (Peck & McDonald, 2014; Peck Gallucci, Sloan, & Lippincott, 2009).

Therefore, for this study, we partnered with the College of Education at a large public university in North Carolina (hereon referred to as Collaborating University) to evaluate the construct validity, reliability, and predictive validity of their locally-scored performance assessment portfolios. Collaborating University (CU) used the widely-adopted TPA that was developed by Stanford University and is aligned with standards for TPPs (e.g. CAEP standards) and practicing teachers (e.g. the Interstate Teacher Assessment and Support Consortium, InTASC, standards). While the edTPA has recently replaced the TPA (SCALE., 2013), this study makes three contributions to the teacher candidate performance assessment research literature. First, this study focuses on the relationship between performance assessment scores and outcomes for program graduates—entry into and exit from the profession, teacher evaluation ratings, and teacher value-added scores. Second, this study compares local TPA portfolio scores to those from the official scorer, Pearson, to assess the utility of locally-scored measures as a guide for program improvements.[2] This is especially important given the centrality of local scoring in the current research on TPP reform and establishing a culture of evidence within TPPs (Miller, Carroll, Jancic, & Markworth, 2015; Peck, Gallucci, Sloan, & Lippincott, 2009, 2014; Peck & McDonald, 2014). Finally, this study serves as a proof of concept for the type of study that individual TPPs or collections of programs can undertake to establish the utility of local scoring of teacher candidate performance assessments to guide their own program improvement efforts. With 11 states requiring teacher candidate performance assessments for program completion and/or licensure decisions and over 600 universities using teacher candidate performance assessments, it is important to provide evidence on the validity and reliability of local scoring (edTPA, 2015).

In the sections that follow, we first provide further background on teacher candidate performance assessments. Specifically, we describe the origins of teacher candidate performance assessments and the organization of the TPA. Second, we detail CU's local scoring procedures, the TPA data and sample, and the outcome measures for the predictive validity analyses. Third, we present our analyses and findings. These analyses include more rigorous factor analysis models to assess construct validity, tests to assess the similarity of ratings from locally and officially-scored portfolios, and a range of regression models to determine whether local TPA scores predict teacher outcomes. Finally, we close with a discussion of the implications of our work for TPPs and their improvement efforts, policy action, and further research.

[1] In the recently released edTPA field test report, SCALE researchers reported two measures of inter-rater reliability: (1) the adjacent agreement rate and (2) the 'Kappa-N', which adjusts for inter-rater agreement by chance. Overall, these values were relatively high—0.917 and 0.829, respectively—and are comparable to reliability rates for other well-established performance assessments (SCALE., 2013).

[2] Pearson and its Evaluation Systems Group is a commercial education assessment organization that has partnered with SCALE to officially-score teaching candidates' edTPA portfolios.

2. Background

2.1. Teacher candidate performance assessments

In the United States, teacher candidate performance assessments originate from (1) the National Research Council's call to develop broader and more authentic assessments of teacher candidates—beyond licensure exams—and their performance in the classroom (Darling-Hammond & Snyder, 2000; Mitchell, Robinson, Plake, & Knowles, 2001); (2) the National Board for Professional Teaching Standards and its performance-based framework for assessing and credentialing veteran teachers (National Board Certification); and (3) the widespread push to improve TPPs and the performance of beginning teachers (Darling-Hammond, Newton, & Wei, 2013). These performance assessments capture a broad range of knowledge and skills and directly evaluate teaching ability by requiring teaching candidates to complete a portfolio during their student teaching experience that includes curriculum plans, video clips of instruction, samples of student work and assessments, and candidate commentaries regarding their teaching decisions (Pecheone & Chung, 2006; SCALE., 2013). Public and private universities in California initially-led the teacher candidate performance assessment initiative through the creation of the Performance Assessment for California Teachers (PACT). Currently, edTPA (which replaced the TPA used in this study), developed by SCALE, is the most widely-adopted teacher candidate performance assessment, with 626 TPPs in 41 states and Washington, DC in varied stages of edTPA implementation (edTPA, 2015).

While a primary purpose of teacher candidate performance assessments is to determine candidates' readiness to teach—potentially linking candidate scores to high-stakes licensure decisions—TPPs may also leverage the educative nature of performance assessments for evidence-based program improvement. As detailed by Peck and colleagues, performance assessment data, particularly locally-scored portfolios, offer TPPs a common language and expectations for candidate performance, a forum for accepting collective responsibility for candidate performance, and direct evidence of the extent to which candidates demonstrate specific knowledge and skills (Peck et al., 2014). Essentially, locally-scored teacher candidate performance assessments can serve as a foundation for a teacher educator learning community (Grossman, Wineburg, & Woolworth, 2001). Program leadership and faculty can use evidence from candidate performance assessments to identify programmatic strengths and areas for improvement, enact evidence-based program reforms, and evaluate the success of those reforms. Teacher candidate performance assessment scores that are valid, reliable, and predictive of graduate outcomes are critical to this learning community foundation.

2.2. Constructs of the TPA

In 2011–12, the TPA consisted of 12 rubrics organized into three main constructs—referred to as Tasks.[3] As shown in the construct blueprint in Table 1, each of these rubrics aligned with one of the three main TPA constructs of planning (Rubrics 1, 2, 3, 10, and 11), instruction (Rubrics 4, 5, and 9), and assessment (Rubrics 6, 7, 8, and 12). In addition, five of the TPA rubrics were cross-listed with a TPA cross-cutting theme: analysis of teaching (Rubrics 8 and 9) and academic language (Rubrics 10, 11, and 12). Evaluators score each rubric from 1 to 5, with 1 indicating a struggling candidate who is not ready to teach, 2 indicating a candidate who needs more practice, 3 indicating an acceptable level of performance to begin teaching, 4 indicating a candidate with a solid foundation of knowledge and skills, and 5 indicating a highly accomplished teaching candidate.

Based on this blueprint, it is unclear how the TPA rubrics will align with the main TPA constructs and cross-cutting themes. With three main constructs and two cross-cutting themes, it is possible that the TPA rubrics may reflect five underlying factors. It is also possible that the three main constructs will emerge or that a combination of the main constructs and cross-cutting themes will underlie the TPA rubrics. For example, using locally-scored portfolios from the PACT, a performance assessment comparable to TPA/edTPA, Duckor and colleagues found that a three domain model of planning, instruction, and metacognition (a combination of assessment, reflection, and academic language items) fit the portfolio scores well and best-identified distinct teaching skills (Duckor et al., 2014). Given the complexity of the TPA blueprint, the underlying factor structure of the 2011–12 TPA scores is difficult to predict. Therefore, we implemented an exploratory factor analysis (discussed in Section 4.1.1) to reveal the actual structure of the rubrics.

3. Method

3.1. TPA scoring procedures at Collaborating University

As a way to establish a culture of evidence and make formative improvements to teacher preparation practices, CU previously examined its teacher candidate performance data to identify areas of strength and shortcomings. As part of this work, CU found that its self-developed candidate portfolio assessment failed to identify the multiple constructs it was designed to assess and was not predictive of graduate outcomes (Henry et al., 2013). This lack of valid assessment data upon which to base program improvement efforts and respond to new accountability pressures drove CU to seek a new portfolio assessment. As a result, CU explored other performance assessments and decided to pilot TPA based upon several advantages: (1) its support from national teacher education associations; (2) its reputation as being "by the profession for the profession"; (3) its use in TPPs across many states; (4) upcoming field testing to assess the instrument's reliability and validity; and (5) its connections to National Board Certification, which resonated with CU's PK-12 partners. CU began piloting the TPA during the 2010-11 academic year in middle grades (mathematics, English, science, and history-social studies) and secondary English and history-social studies. In 2011–12, CU expanded the pilot to include elementary and special education, with local evaluators scoring candidates' portfolios.

To prepare for local TPA scoring, CU required each local evaluator to participate in nine hours of TPA training facilitated by officially-calibrated faculty—these training sessions pre-dated the local evaluation protocols developed by SCALE (Dobson, 2013). In these sessions the TPA trainers provided local evaluators with a thorough description of each TPA rubric and the criteria for each scoring level (1–5). To calibrate performance assessment scoring, the TPA trainers supplied local evaluators with sample TPA portfolios and in small groups the TPA trainers facilitated discussions regarding the quality of evidence in each portfolio and the score for each rubric. After the groups reached a consensus for each rubric score, the TPA trainers revealed the official score and the local evaluators engaged in further discussion regarding the portfolio evidence. Importantly, these local scorer trainings were in line with research evidence showing that rubric use can enhance the reliability of performance assessment scoring and promote learning—providing faculty opportunities for self-assessment based on candidates' performance against a common set of expectations (Jonsson & Svingby, 2007).

Once the local evaluation training was complete, CU assigned

[3] Here, we follow the terminology of TPA/edTPA and refer to each measure upon which teaching candidates receive a score (e.g. Engaging Students) as a rubric.

Table 1
Construct blueprint for teacher performance assessment rubrics.

Main construct	Main construct only	Cross-cutting: Analysis of teaching	Cross-cutting: Academic Language	Count of standards in main constructs
Planning	Planning for content understanding (1) Knowledge of students for planning (2) Planning for assessment (3)	–	Language demands (10) Language supports (11)	5
Instruction	Engaging students (4) Deepening student learning (5)	Analysis of teaching (9)	–	3
Assessment	Analysis of student learning (6) Feedback (7)	Using assessment results (8)	Language use (12)	4
Count of main only and cross-cutting rubrics	7	2	3	12

Note: This table places each of the rubrics into the main construct and, when applicable, into the cross-cutting theme as designated in the TPA blueprint. TPA rubric numbers are listed in parentheses.

teacher candidates' TPA portfolios to the trained local scorers. To control for bias, CU blinded scoring assignments within content areas and did not assign university supervisors or faculty to score the portfolios of the candidates they supervised during student teaching. To limit workload, CU assigned no more than five TPA portfolios to any local evaluator. If candidates received scores of 1 or 2 on any rubric, lead faculty remediated the candidates and CU allowed the candidates to revise their portfolio for scoring. The present study is limited to candidates' initial TPA scores, since we believe initial scores best capture candidates' own knowledge and ability to effectively perform key teaching tasks.

Finally, in the spring of 2012, SCALE, the creator of TPA, offered CU the opportunity to submit TPA portfolios to be officially-scored by Pearson as part of the national TPA field test. Though this offer came after CU had completed all of the 2012 local scoring and after many teacher candidates had graduated, 64 of the 249 teacher candidates submitted their TPA portfolios for official scoring.

3.2. TPA scores at Collaborating University

For the present study, we relied on two sets of portfolio scores provided by CU: (1) 249 locally-scored performance assessment portfolios from the 2011–12 graduating cohort and (2) 64 officially-scored performance assessment portfolios for a subset of 2011–12 graduates who also have locally-scored portfolios. The 249 locally-scored portfolios measure the 12 TPA rubrics in effect during the 2011–12 academic year as displayed in Table 2. The TPA scores for this study include those from eight different TPA handbook areas—elementary literacy, special education, secondary English and history-social studies, and middle grades mathematics, English, science, and history-social studies—which were scored by 75 different raters at CU (an average of 3.32 portfolios per local rater). The 64 officially-scored portfolios from 2011-12 cover the 12 TPA rubrics in effect during the 2011–12 academic year[4]and come from seven different areas—all those from the locally-scored portfolios except special education.[5] In Table 2 we provide descriptive statistics—means, standard deviations, and scoring distributions—for the sample (n = 249) of local scores from 2011-12; we present means and standard deviations for the 2011–12 official scores (n = 64) in Table 4.

3.3. Outcome measures for predictive validity analyses

We include four teacher outcome measures in this study: entry into teaching, teacher attrition, teacher evaluation ratings, and teacher value-added scores. The full analysis sample includes all 249 graduates of CU in 2011–12 who have locally-scored TPA portfolios. As detailed below, however, based on data availability and the research objective, the analysis sample differs for each of the teacher outcome measures.

First, to determine whether local TPA scores predict entry into the state's teacher workforce, we relied on certified salary files provided by the North Carolina Department of Public Instruction (NCDPI). We created a dichotomous outcome variable for individuals paid as teachers in North Carolina public schools (NCPS) during the 2012–13 academic year. Our sample includes all 249 CU graduates with local TPA scores, of which 181 (73 percent) taught in NCPS in 2012–13.

Second, contingent on entering the state's public school teacher workforce in 2012–13, we assess the relationships between local TPA scores and attrition from the state's public schools. Specifically, we used salary data from the September 2013 pay period, provided by the NCDPI, to create a dichotomous outcome variable for individuals who did not return to teaching in NCPS for the 2013–14 academic year. Overall, of the 181 teachers in the sample for this analysis, 13 (7 percent) did not return to NCPS in 2013–14.

Third, to examine whether local TPA scores predict teacher evaluation ratings, we use data from the North Carolina Educator Evaluation System (NCEES), an evaluation rubric in place across NCPS in which school administrators rate teachers across five standards: (Standard 1) teachers demonstrate leadership; (Standard 2) teachers establish a respectful environment for a diverse group of students; (Standard 3) teachers know the content they teach; (Standard 4) teachers facilitate learning for their students; and (Standard 5) teachers reflect on their practice. To evaluate teachers, school administrators use at least three formal classroom observations and paper-based evidences to document key teaching behaviors and rate teachers as not demonstrated, developing, proficient, accomplished, or distinguished on each of the five NCEES standards. For these analyses the outcome variable is a 1–5 ordinal value and the sample includes the 172 individuals (95 percent of those teaching in 2012–13) with local TPA scores who

[4] In the officially-scored portfolios in 2011–12, Rubric 2 was split into two parts: Knowledge of Students and Justification for Plans. In the locally-scored 2011–12 portfolios, Rubric 2 was only Knowledge of Students.

[5] Official scoring during this stage of the TPA/edTPA field test was limited and did not include the opportunity for CU to submit teacher candidate portfolios for either the Special Education Inclusive Settings or Special Education Other Settings handbook areas. As a result, officially-scored special education portfolios are not available in this analysis. For special education teaching candidates, CU has officially-scored performance assessments for their 2013–14 and 2014-15 graduating cohorts. These data will be available for future construct validity, reliability, and predictive validity analyses.

Table 2
Descriptive statistics from 2011-12 local TPA scores (n = 249).

Rubric	TPA construct (cross-cutting theme)	Mean & standard deviation	Percentage scoring at level 1	Percentage scoring at level 2	Percentage scoring at level 3	Percentage scoring at level 4	Percentage scoring at level 5
Planning for content understanding	Planning	3.29 (0.74)	2.01	6.83	55.42	31.33	4.42
Knowledge of students for planning	Planning	3.19 (0.70)	2.01	8.43	60.24	26.91	2.41
Planning for assessment	Planning	3.35 (0.74)	2.01	5.22	53.01	34.94	4.82
Engaging students	Instruction	3.32 (0.73)	2.01	4.82	57.03	30.92	5.22
Deepening student learning	Instruction	3.25 (0.71)	2.81	3.61	62.65	26.91	4.02
Analysis of student learning	Assessment	3.26 (0.68)	1.61	4.82	62.65	26.91	4.02
Feedback	Assessment	3.27 (0.74)	2.41	6.43	56.63	30.52	4.02
Using assessment results	Assessment (analysis of teaching)	3.26 (0.74)	3.61	4.02	57.83	31.33	3.21
Analysis of teaching	Instruction (analysis of teaching)	3.21 (0.73)	2.41	6.83	62.65	23.29	4.82
Language demands	Planning (academic language)	3.06 (0.67)	2.01	11.24	67.07	17.27	2.41
Language supports	Planning (academic language)	3.24 (0.69)	1.61	5.62	63.45	25.30	4.02
Language use	Assessment (academic language)	3.18 (0.73)	2.01	8.84	61.85	22.89	4.42

Note: This table displays the mean, standard deviation (in parentheses), and scoring distribution for each of the 12 TPA rubrics from the 2011–12 year. The table also indicates to which construct, and when applicable, which cross-cutting theme (in parentheses), the rubrics belong.

Table 3
Factor loadings with the 2011–12 local TPA scores.

TPA rubric	TPA construct (cross-cutting theme)	Factor loadings with group mean-centered 2011–12 local TPA scores		
		Factor 1	Factor 2	Factor 3
Planning for content understanding	Planning	**0.777**	0.166	−0.027
Knowledge of students for planning	Planning	**0.884**	0.136	−0.156
Planning for assessment	Planning	**0.592**	0.013	0.334
Engaging students	Instruction	**0.590**	−0.082	0.353
Deepening student learning	Instruction	**0.556**	−0.028	0.382
Analysis of student learning	Assessment	0.166	**0.545**	0.211
Feedback	Assessment	0.078	**0.719**	0.100
Using assessment results	Assessment (analysis of teaching)	−0.023	**0.898**	0.043
Analysis of Teaching	Instruction (analysis of teaching)	0.078	**0.806**	−0.008
Language demands	Planning (academic language)	0.043	0.136	**0.728**
Language supports	Planning (academic language)	0.159	0.223	**0.569**
Language use	Assessment (academic language)	−0.064	0.020	**0.924**

Note: This table presents factor loadings for the 2011–12 locally-scored TPA portfolios. All factor loadings greater than 0.40 are bolded.

both taught in NCPS in 2012–13 and were evaluated by a school administrator.[6]

Finally, to examine whether local TPA scores predict teacher value-added, we relied on teachers' EVAAS (Education Value-Added Assessment System) estimates—the official measure of teacher value-added in NCPS—produced by the SAS Institute™. For NCPS there are two types of EVAAS models—the multivariate response model (MRM), a random effects model that estimates teacher value-added to student achievement on the state's End-of-Grade (grades 3–8) math and reading exams and the univariate response model (URM), a hybrid random and fixed effects model that estimates teacher value-added to student achievement on the state's End-of-Course exams (algebra I, biology, and English II), 5th and 8th grade science exams, and all other courses with final exams (e.g. U.S. history, chemistry, geometry). For these analyses we make teachers' EVAAS estimates the dependent variable and the sample includes 114 EVAAS estimates—61 MRM estimates and 53 URM estimates—for 76 unique teachers (42 percent of those with TPA scores teaching in 2012–13) with local TPA data who taught a tested-grade/subject in 2012–13.

4. Results

4.1. Analyses

4.1.1. Factor analysis—construct validity

We define construct validity as the extent to which the 12 TPA rubrics are well-aligned with the three TPA main constructs—planning, instruction, and assessment—and two cross-cutting themes—analysis of teaching and academic language (Cronbach & Meehl, 1955). Due to the complexity of the construct blueprint, we implemented exploratory factor analysis (EFA) to examine the underlying factor structure of the locally-scored TPA portfolios and to ascertain whether the local scores could be used to obtain valid and interpretable constructs. An important function of EFA is to determine the number of factors to be retained, and for this analysis, instead of using traditional methods such as Kaiser's Rule of eigenvalues larger than one or scree plot examination, we employed parallel analysis (PA; Horn, 1965). Parallel analysis is a

[6] For Standards 1, 2, 4, and 5, the range of evaluation ratings in our sample is from 2 to 4 (developing to accomplished); for Standard 3 the range of evaluation ratings in our sample is from 1 to 4 (not demonstrated to accomplished).

Table 4
Comparing locally and officially-scored portfolios from 2011-12.

Correlations between locally-scored and officially-scored portfolios		Mean standard scores and standard deviations	
TPA rubric	Correlation	Locally-scored	Officially-scored
Planning for content understanding	0.022	3.45** (0.69)	3.08 (0.76)
Knowledge of students for planning	0.143	3.40** (0.71)	2.97 (0.76)
Planning for assessment	0.203	3.41** (0.71)	3.03 (0.80)
Engaging students	0.030	3.41** (0.64)	2.92 (0.86)
Deepening student learning	0.142	3.45** (0.69)	2.75 (0.85)
Analysis of student learning	0.157	3.37** (0.58)	2.95 (0.91)
Feedback	0.018	3.51** (0.67)	2.70 (0.87)
Using assessment results	0.081	3.40** (0.68)	2.78 (1.04)
Analysis of teaching	0.092	3.36** (0.70)	2.80 (0.91)
Language demands	0.060	3.19** (0.64)	2.62 (0.66)
Language supports	0.014	3.38** (0.61)	2.86 (0.80)
Language use	0.036	3.30* (0.75)	2.92 (0.88)

Note: The left panel of this table displays correlations between the locally-scored TPA portfolios and officially-scored TPA portfolios. The right panel of this table displays the average TPA scores and standard deviations from the portfolios that were both locally and officially-scored. +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.

more rigorous, empirically-based, and preferred method to determine the number of underlying factors in a dataset (Courtney, 2013; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Horn, 1965; Thompson, 2004). To do so, PA compares the underlying factor structure of an analysis dataset with the underlying structure of randomly-generated data and retains factors in the analysis dataset if their explained variance is greater than the explained variance of corresponding factors in the randomly-generated data (Horn, 1965; Thompson, 2004).

In addition to employing the more rigorous PA method to determine the number of retained factors, we also investigated factor analysis approaches that adjust for the clustering of locally-scored TPA portfolios. Specifically, CU's 249 locally-scored portfolios are nested within 75 local raters; without appropriate adjustments for this clustering, the assumption of independent observations may be violated and EFA results biased (Longford & Muthen, 1992; Reise, Ventura, Nuechterlein, & Kim, 2005). To overcome this challenge, we first examined the intra-class correlation of scores for each TPA standard. As shown in Appendix Table 1, the intra-class correlations ranged from 0.009 to 0.316 and there was significant between-rater variance for eight of the 12 TPA rubrics (six at the 0.05 level and two at the 0.10 level). While it is possible that these between rater differences are due to differences in the quality of TPA portfolios, rather than differences in the rating practices of scorers, we followed the suggestion of Reise and colleagues and group mean-centered TPA scores and then conducted EFA on the within-rater correlation matrix (Reise et al., 2005). We utilize these group-mean centered factor results in our predictive validity analyses.

Using both PA and clustering corrections, we employed the principal factor method to fit our factor model (Fabrigar et al., 1999). Compared with maximum likelihood methods, this model-fitting approach requires no distributional assumptions and is less likely to produce improper solutions. Regarding rotation options, we hypothesized that factors of TPA scores would be correlated with each other, as they measure components of an integrated teaching process, and began with a non-orthogonal rotation (promax). To assess the use of the promax rotation, versus the varimax (orthogonal) rotation, we examined the correlations among factors from promax rotation analyses. These correlations were all above 0.32, and thus, following the guidelines of Brown (2009), we implemented non-orthogonal factor analysis rotations. Finally, we conducted the PA using the *paran* package in R version 3.1.0 (Dinno, 2012; R Core Team, 2014); we implemented EFA using SAS 9.3 (SAS Institute, 2011).

4.1.2. Comparing local and official scores—reliability

We define reliability as the extent of agreement between locally and officially-scored TPA rubric scores (Saal, Downey, & Lahey, 1980). To compare the local versus official TPA scoring, we first examined the correlations between the local and official scores for each TPA rubric and used paired *t*-tests to assess whether there were statistically significant differences in the mean values for the two sets of scores. For each TPA rubric, these analyses determine whether the local scores are systematically higher or lower than the official TPA scores. Second, since local scores could systematically differ from official scores yet still reliably identify a candidate's relative placement in the TPA scoring distribution, we estimated a Spearman rank order correlation. For this analysis we summed the local and official scores across the 12 TPA rubrics and assessed the extent to which candidates with a high or low total score, locally, were similarly scored, officially. Due to sample size limitations, we did not conduct an EFA on the 64 TPA portfolios with official scores.

4.1.3. Teacher outcome analyses—predictive validity

We define predictive validity as the extent to which locally-scored TPA measures significantly predict outcomes for graduates as classroom teachers. To understand the relationships between the local TPA scores and teacher outcomes, we began by examining the bivariate correlations between the four outcomes and (1) the TPA constructs identified through EFA and (2) the standardized total score across all 12 locally-scored TPA rubrics. We estimated point-biserial correlations for the binary outcomes (entering and exiting the teacher workforce), Spearman rank order correlations for the categorical outcomes (evaluation ratings), and Pearson correlations for the continuous outcomes (value-added estimates). Next, we employed a set of regression models—logistic, ordered logit, and ordinary least squares (OLS) depending upon the dependent variable—to assess the multivariate relationship between teacher outcomes and local TPA scores. In these regression models, we specified the key independent variables as either the TPA constructs identified by factor analysis or the standardized total score (standardized across all 249 locally-scored portfolios). Below, we detail our three regression approaches to address each research outcome.

First, to estimate the relationship between local TPA scores and teachers' entry into or exit from the NCPS workforce, we specified a logistic regression model where becoming a teacher in 2012–13 or exiting teaching (not returning to NCPS in 2013–14) is a binary outcome. We included robust standard errors in the entry into

teaching models and cluster-adjusted standard errors, at the school level, for the attrition analyses. Coefficients from these models indicate the extent to which a one standard deviation increase in a TPA factor or the TPA total score impact the odds of workforce entry or exit.

Second, to determine whether local TPA scores predict teachers' evaluation ratings, we specified separate ordered logistic regression models for each of the five professional teaching standards in North Carolina, where the outcome variable is a teacher's 1–5 (not demonstrated through distinguished) evaluation score. In these models, we adjusted for nesting within schools by clustering standard errors at the school level. Coefficients from these models indicate the extent to which a one standard deviation increase in a TPA factor or the TPA total score impact the odds of rating higher on North Carolina's evaluation standards.

Finally, to examine whether local TPA scores predict teachers' value-added estimates, we specified an OLS regression model with teachers' EVAAS estimates as the outcome variable. For these analyses we specified one model that pools teacher value-added estimates from the MRM and URM data.[7] We then performed separate analyses for the MRM and URM data. In this way we estimate the relationships between TPA scores and all available value-added data and then determine whether TPA scores differentially predict teacher effectiveness on End-of-Grade math and reading assessments (MRM) or End-of-Course, 5th and 8th grade science, and final exams (URM). In all these models we cluster-adjusted standard errors at the school level. Coefficients from these models indicate the extent to which a one standard deviation increase in a TPA factor or the TPA total score predict teachers' value-added to student achievement.

4.2. Findings

4.2.1. Construct validity

Following Reise and colleagues, we began by using EFA to examine the factor structure of the group mean-centered local TPA scores (Reise et al., 2005). Parallel analysis on this group mean-centered locally-scored data revealed a three factor structure. Table 3 shows that TPA rubrics 1–5 loaded onto the first factor, TPA rubrics 6–9 loaded onto the second factor, and TPA rubrics 10–12 loaded onto the third factor. Comparing the TPA constructs with these group mean-centered factor analysis results, we find that the three factor structure is only partially consistent with the TPA construct blueprint. This result is comparable to that of Duckor and colleagues' analysis of the construct validity of locally-scored PACT portfolios (Duckor et al., 2014).

The first factor contains three rubrics in the planning construct—planning for content understanding, knowledge of students for planning, and planning for assessment—and two rubrics from the instruction construct—engaging students and deepening student learning. We refer to this first factor as *Planning and Instruction*. The second factor includes three rubrics from the assessment construct—analysis of student learning, feedback, and using assessment results—and one rubric from the instruction construct—analysis of teaching. We refer to this second factor as *Analysis and Feedback* and note that two of these rubrics, using assessment results and analysis of teaching, are part of the *Analysis of Teaching* cross-cutting theme. In the recently conducted field test of edTPA, the *Analysis of Effective Teaching* rubric also loaded with the assessment construct rather than the instruction construct (SCALE., 2013). Finally, the third factor contains two rubrics from the planning construct—language demands and language supports—and one rubric from the assessment construct—language use. Given that these three TPA rubrics comprise the *Academic Language* cross-cutting theme, we refer to the third factor as *Academic Language.*

Overall, we conclude that the underlying measures are only partially aligned with the three main TPA constructs and two cross-cutting themes. Two of the main constructs are combined into a single latent variable, *Planning and Instruction*. Another of the main constructs, *Assessment,* is present as a latent variable but combined with the cross-cutting theme, *Analysis of Teaching*, and labeled as *Analysis and Feedback*. Finally, one of the cross-cutting themes, *Academic Language*, is present and completely consistent with the conceptual blueprint. In our predictive validity analyses, we examine the relationship between these three latent factors from the locally-scored TPA data and teacher outcomes.

4.2.2. Reliability

To address our second research question we started with bivariate correlations and paired *t*-tests to compare the TPA rubric scores of the 64 CU teacher candidates whose portfolios were both locally and officially scored. The left panel of Table 4 displays correlations between each locally and officially-scored rubric. These correlations range between 0.014 and 0.203—with eight correlations less than 0.10—and none of the correlations is statistically significant. The right panel of Table 4 shows mean comparisons and *t*-test results. For all 12 TPA rubrics, the local scores are significantly higher than the official scores.

Even though the local scores are systematically higher than the official scores, local scores still may reliably identify a candidate's relative placement in the TPA scoring distribution. However, the Spearman rank order correlation between the local total score and official total score was 0.101 and statistically insignificant (p-value of 0.435). Further analyses, in which we separately divided the local and official total scores into tertiles, shows that of the 22 locally-scored portfolios in the bottom tertile, 11 (50 percent) were in the bottom tertile, 6 (27 percent) were in the middle tertile, and 5 (23 percent) were in the top tertile for official scoring. Of the 18 locally-scored portfolios in the top tertile, 6 (33 percent) were in the top tertile, 8 (44 percent) were in the middle tertile, and 4 (22 percent) were in the bottom tertile for official scoring. Overall, while the factor analysis results indicate that local scoring is aligned with some of the key constructs/themes of TPA, local scores, in comparison to official scores, are systemically higher and do not reliably identify high or low scoring candidates.

4.2.3. Predictive validity

To assess the predictive validity of the locally-scored TPA portfolios, we began by examining the bivariate correlations between our TPA measures and the teacher outcomes. As shown in Table 5, the *Academic Language* factor is positively and significantly correlated with entry into the teacher workforce, while none of the TPA measures is significantly correlated with teacher attrition (conditional on entry into teaching in North Carolina). Regarding teacher evaluation ratings, the *Planning and Instruction* factor and the standardized total score are both positively and significantly correlated with Standard 1 (teachers demonstrate leadership), Standard 3 (teachers know the content they teach), Standard 4 (teachers facilitate learning for their students), and Standard 5 (teachers reflect on their practice). In every significant relationship, the correlation with the total score is larger than the correlation with the *Planning and Instruction* factor. Only Standard

[7] Because the distribution of EVAAS estimates differs between the MRM and URM data, we include an indicator variable for URM observations in these pooled analyses.

Table 5
Correlations between local TPA measures from 2011-12 and teacher outcome variables.

TPA measure	Becomes a teacher	Exits NCPS	Standard 1 leadership	Standard 2 classroom environment	Standard 3 content knowledge	Standard 4 facilitating student learning	Standard 5 reflecting on practice	Overall EVAAS	EVAAS MRM	EVAAS URM
Factor 1: Planning and instruction	0.010	−0.088	0.178*	0.097	0.131+	0.183*	0.215**	0.158+	0.007	0.265*
Factor 2: Analysis and feedback	0.064	−0.099	0.105	0.067	0.085	0.100	0.115	0.015	0.044	−0.017
Factor 3: Academic language	0.152*	−0.090	0.078	0.007	0.099	0.117	0.109	0.021	−0.061	0.140
Std. Total score	0.093	0.032	0.197**	0.075	0.198**	0.227**	0.239**	0.107	−0.019	0.102

Note: For all binary outcomes (becomes a teacher and exits NCPS) we use point-biserial correlations; for categorical outcomes (teacher evaluation ratings) we use Spearman correlations; for continuous outcomes (EVAAS teacher value-added estimates) we use Pearson correlations. +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.

Table 6
Workforce entry and Attrition outcomes.

	Becomes a teacher in NCPS	Exits NCPS
Factor 1: Planning and instruction	0.729 (0.105)	0.931 (0.816)
Factor 2: Analysis and feedback	1.047 (0.798)	0.808 (0.416)
Factor 3: Academic language	1.666* (0.021)	0.844 (0.604)
Std. Total score	1.232 (0.155)	1.138 (0.530)
Cases	249	181

Note: Cells report odds ratios and p-values in parentheses. +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.

2 (classroom environment) is not significantly correlated with *Planning and Instruction* or the teacher candidates' total score. In addition, the *Planning and Instruction* factor is positively and significantly correlated with overall teacher value-added (pooling MRM and URM data) and teacher value-added using the URM. The correlation with the overall value-added score appears to be driven by the significant correlation between *Planning and Instruction* and the URM scores, which are based on 5th and 8th grade science exams and high school grades End-of-Course and final exams.

Turning to multivariate regression models, the left panel of Table 6 indicates that the *Academic Language* factor significantly predicts entry into the NCPS teacher workforce. Holding the other factors at their mean values, candidates with an *Academic Language* factor score two standard deviations below the mean have a predicted probability of 50 percent for entering the teacher workforce. As a comparison, candidates with an *Academic Language* factor score two standard deviations above the mean have a predicted probability of nearly 90 percent for entering the teacher workforce. While none of the remaining coefficients in the left panel of Table 6 are statistically significant, the odds ratio for the *Planning and Instruction* factor approaches statistical significance at the $\alpha < 0.10$, which may suggest that candidates with higher *Planning and Instruction* values are less likely to enter NCPS.

Contingent on entering the teacher workforce in 2012–13, the logistic regression results from the right panel of Table 6 show that neither the TPA factors nor the standardized total score significantly predict teacher attrition. Here, we note a limitation of this analysis—only 13 teachers (out of 181) did not return to NCPS in 2013–14—and suggest that a longer time period, which can be expected to yield more exiting teachers, may be required to estimate relationships between TPA scores and attrition.

For our first measure of teacher performance, ordered logistic regression results in Table 7 indicate that the *Planning and Instruction* factor significantly predicts higher teacher evaluation ratings for Standard 1 (Leadership), Standard 4 (Facilitating Student Learning), and Standard 5 (Reflecting on Practice). The significant relationship between *Planning and Instruction* and Standard 4 is expected, since many of the teacher actions and competencies that comprise Standard 4—teachers know their students and plan appropriate instruction, teachers use a variety of methods to engage students, and teachers help students develop critical thinking skills—are well-aligned with the TPA rubrics loading onto the *Planning and Instruction* factor. Additionally, Table 7 shows that the standardized total score variable significantly predicts higher evaluation ratings across all five standards. To make these odds ratios more interpretable, Fig. 1 displays predicted probabilities for receiving an evaluation rating of developing or accomplished at three different values of the standardized total score variable (please see Appendix Table 2 for more predicted probability values). For instance, teachers with a total score two standard deviations below the mean have a 30 percent predicted probability of receiving a rating of developing (below proficient on the evaluation rating scale) on Standard 4 and only a six percent predicted probability of receiving an accomplished rating (above proficient on the evaluation rating scale); at the other end of the distribution, teachers with a total score two standard deviations above the mean have a five percent predicted probability of rating at developing and a 33 percent predicted probability of rating as accomplished.

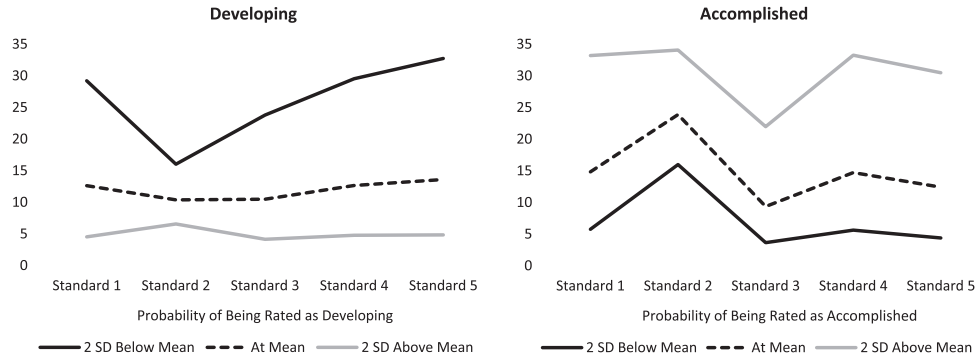
To investigate the extent to which these significant associations with evaluation ratings may be due to the sorting of teacher candidates with higher TPA scores into K-12 schools with more advantaged students, rather than the effectiveness of the teachers, we re-ran the ordered logistic regression models controlling for the percentage of minority and free and reduced-price lunch students at the school. These results (shown in Appendix Table 3) are robust to the inclusion of school controls—only the total score variable for Standard 2 loses statistical significance—suggesting that the local TPA scores for teacher candidates predict evaluation ratings when the candidates become teachers-of-record.

Finally, for our second measure of teacher performance, results in Table 8 indicate that the *Planning and Instruction* factor is significantly associated with teacher value-added in analyses limited to the URM estimates. Specifically, a one standard deviation increase in the *Planning and Instruction* factor is associated with students gaining an additional 1.4 scale score points on their End-of-Grade, End-of-Course, and final exams in URM-eligible courses. When including variables for the percentage of minority and free and reduced-price lunch students at the school to control for the

Table 7
Teacher evaluation ratings in 2012–13.

	Standard 1 leadership	Standard 2 classroom environment	Standard 3 content knowledge	Standard 4 facilitating student learning	Standard 5 reflecting on practice
Factor 1: Planning and instruction	1.649* (0.016)	1.240 (0.278)	1.362 (0.179)	1.559* (0.024)	1.550* (0.041)
Factor 2: Analysis and feedback	0.915 (0.647)	0.927 (0.701)	0.900 (0.661)	0.912 (0.659)	0.853 (0.475)
Factor 3: Academic language	0.990 (0.955)	1.062 (0.754)	1.162 (0.444)	1.122 (0.583)	1.136 (0.502)
Std. Total score	1.689** (0.001)	1.284+ (0.062)	1.651** (0.002)	1.701** (0.000)	1.759**(0.000)
Cases	172	172	172	172	172

Note: Cells report odds ratios from ordered logit models with p-values in parentheses. +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.



Note: For three different values of the standardized TPA total score variable (two standard deviations below the mean, at the mean, and two standard deviations above the mean), this figure displays predicted probabilities of rating as developing or accomplished on Standards 1-5 of the NCEES.

Fig. 1. Predicted probabilities of being rated developing and accomplished on the NCEES.

Table 8
Teacher EVAAS estimates in 2012–13.

	All EVAAS estimates	MRM EVAAS estimates	URM EVAAS estimates
Factor 1: Planning and instruction	0.629 (0.383)	−0.003 (0.525)	1.388+ (0.724)
Factor 2: Analysis and feedback	−0.285 (0.394)	0.223 (0.492)	−0.748 (0.594)
Factor 3: Academic language	−0.124 (0.412)	−0.280 (0.527)	−0.075 (0.682)
Std. Total score	0.155 (0.269)	−0.054 (0.476)	0.335 (0.327)
Cases	114	61	53

Note: Cells report coefficients from regression models with cluster-adjusted standard errors in parentheses. +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.

possible sorting of higher scoring teachers into higher performing schools, this coefficient shrinks to 0.8 and is no longer statistically significant. While this may suggest that the *Planning and Instruction* factor is not a valid predictor of teacher value-added in URM eligible-courses, the small sample in this analysis—53 observations from 41 unique teachers—warrants caution when interpreting results. In the overall and MRM analyses, neither the TPA factors nor the standardized total score significantly predict teacher value-added.

5. Discussion

Overall, this study indicates that locally-scored teacher candidate performance assessments can have a reasonable degree of construct and predictive validity—measuring the main and cross-cutting TPA constructs and significantly predicting aspects of first-year teacher performance. While the EFA did not fully reproduce the TPA construct blueprint, we found factors that corresponded to expected constructs using the rater mean-centered TPA data. This result is comparable to Duckor and colleagues' analysis of locally-scored PACT portfolios and indicates that when scoring, faculty members can identify key domains of teacher candidate performance assessments (Duckor et al., 2014). Perhaps most importantly, the *Planning and Instruction* construct and the standardized total score significantly predicted teachers' evaluation ratings; the *Planning and Instruction* construct was also

positively associated with one teacher value-added measure. This validates the scoring by local faculty—they can identify higher quality instructional practices for student teachers that translate into performance outcomes for first-year teachers. Conversely, the locally-scored performance assessments were systematically higher than official scores and did not reliably identify high and low scoring candidates (Youngs & Bird, 2010). Taken together, these predictive validity and reliability results are comparable to findings from in-service teacher evaluation research which indicate that evaluation ratings are significantly associated with other teacher performance outcomes (e.g. value-added, student surveys) and that principals' ratings of teachers are often higher than those from external observers (Jacob & Lefgren, 2008; Kane & Staiger, 2012; Sartain, Stoelinga, & Brown, 2011). Given that local scorers had only nine hours of training from officially-calibrated faculty before they independently scored candidates' portfolios, we believe these mixed findings suggest the promise of locally-scored performance assessments to provide TPPs with data upon which they can engage faculty in program improvement and create a culture of evidence. Further work needs to assess the extent to which more extensive scoring training and continued experience scoring portfolios benefits the reliability of local scoring.

Regarding these results, we note a limitation of this study: it is based on a single cohort of graduates from a public university in North Carolina implementing a field test instrument. This has two important implications when interpreting findings. First, from a statistical/measurement standpoint, we may benefit from a longer time window in which to assess outcomes (e.g. multiple years of value-added or teacher retention data) and a larger sample (e.g. data from multiple institutions or cohorts) with which to estimate models. Second, from a generalizability standpoint, results may differ for other institutions or for other graduating cohorts. These limitations call for continued research and replication studies but should not slow the progress of research or the efforts of TPPs to become more evidence-based.

These results fit into the current policy environment that values evidence and continuous improvement for TPPs. In this context TPPs must decipher what sources of data best inform program improvement efforts. While measures of graduates' performance as classroom teachers can play a key role in program reform, TPPs also need sources of data that are more proximate to teacher preparation experiences, that provide a common framework for teacher educator conversations (Grossman et al., 2001), and that allow TPPs to gauge multiple aspects of teaching practice. Furthermore, these data must be valid, reliable, and predictive of later outcomes for classroom teachers. Our research shows that locally-scored teacher candidate performance assessments have the potential to fulfill this role for TPPs.

With teacher candidate performance assessments that meet these criteria, TPPs can initiate actions to turn candidates' performance assessments into more effective practices and graduates. For instance, TPPs could conduct latent class analyses to group teaching candidates together based on their performance assessment scores and then use this classification structure to (1) inform targeted remediation of candidates prior to program completion and into their early-career period and (2) predict candidates' assignment to latent classes with other sources of program data (e.g. entry characteristics, coursework performance, participation in programmatic components) so that targeted intervention can occur more quickly for future cohorts (Halpin & Kieffer, 2015). Additionally, results from predictive validity analyses can help TPPs prioritize improvement efforts focused on performance assessment measures that are significantly associated with teacher performance. For example, if a particular performance assessment rubric strongly predicts teacher performance, then TPPs can identify the teaching knowledge and skills underlying that rubric, pinpoint when and by whom those knowledge and skills are taught to candidates, and design interventions to raise candidates' scores on that performance assessment rubric. Local performance assessment scores that are valid, reliable, and predictive of graduate outcomes will not directly lead to program improvement, however, meeting these criteria is vital to programs' efforts to turn performance assessment scores into evidence-based reforms.

Moving forward, researchers should conduct similar evaluations with locally-scored edTPA data. If the local edTPA scores return positive predictive validity results, TPPs can feel more confident in embracing edTPA as a data source around which to build a culture of evidence and teacher educator learning community, diagnose concerns with the current curriculum, and evaluate the effectiveness of program improvements. Here, it is important for researchers, to the extent possible, to assess local edTPA scores against multiple measures of classroom teacher performance—rather than value-added only. These multiple measures may facilitate a larger sample for analysis and allow TPPs to assess whether particular aspects of edTPA, such as the instruction task, predict well-aligned teacher performance outcomes. Regarding other purposes for teacher candidate performance assessments, such as high-stakes teacher certification decisions, these results suggest that it may be inappropriate for states or TPPs to base such decisions on locally-scored performance assessment portfolios. Rather, it is advisable to employ local scoring to provide a language, context, and forum for evidence-based program improvement and official scoring, if research supports its predictive validity, as a potential requirement for teacher certification.

Taken together, this study makes an important contribution to TPPs and considerations of data use for evidence-based program improvement. We show that with limited training, local faculty can score performance assessments with a reasonable degree of construct and predictive validity. This means TPPs can use these data—which are available prior to program completion and identify multiple domains of teaching practice—as a basis for program reforms. We believe this research should encourage continued examinations of teaching candidate performance assessments and help support the establishment of an evidence-based culture within TPPs that respects the criteria of construct validity, reliability, and predictive validity.

Acknowledgements

We are grateful to the faculty and staff at our collaborating public university for providing their TPA data and being such enthusiastic and receptive research partners. We wish to thank Alisa Chapman with the University of North Carolina General Administration (UNCGA) for her support and feedback and acknowledge funding for this research as part of the UNCGA Teacher Quality Research Initiative.

Appendix

Table 1
Between-rater variance in local TPA standard scores.

TPA rubric	Intra-class correlation	Estimated variance of random intercept
Planning for content understanding	0.316	0.133**
Knowledge of students for planning	0.278	0.107**
Planning for assessment	0.173	0.081*
Engaging students	0.174	0.080*
Deepening student learning	0.180	0.079*
Analysis of student learning	0.036	0.016
Feedback	0.058	0.030
Using assessment results	0.096	0.049
Analysis of teaching	0.117	0.057+
Language demands	0.117	0.048+
Language supports	0.157	0.064*
Language use	0.009	0.005

Note: +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.

Table 2
Predicted probabilities from ordered logit models (Std. Total score).

Total score value	Standard 1 leadership		Standard 2 classroom environment		Standard 3 content knowledge		Standard 4 facilitating student learning		Standard 5 reflecting on practice	
	Developing	Accomp.	Developing	Accomp.	Developing	Accomp.	Developing	Accomp.	Developing	Accomp.
2 SD below mean	29.20	5.75	16.01	15.98	23.78	3.64	29.53	5.62	32.73	4.38
1 SD below mean	19.62	9.34	12.94	19.63	16.08	5.88	19.77	9.20	21.67	7.46
At mean	12.62	14.83	10.37	23.87	10.47	9.35	12.65	14.70	13.59	12.41
1 SD above mean	7.88	22.73	8.27	28.69	6.64	14.56	7.85	22.66	8.21	19.96
2 SD above mean	4.82	33.20	6.56	34.06	4.14	21.96	4.77	33.26	4.83	30.48

Note: For five different values of the standardized TPA total score variable (2 SD below the mean to 2 SD above the mean) cells report predicted probabilities of rating as developing or accomplished (Accomp) on Standards 1–5 of the NCEES.

Table 3
Teacher evaluation ratings (Controlling for school covariates).

	Standard 1 leadership	Standard 2 classroom environment	Standard 3 content knowledge	Standard 4 facilitating student learning	Standard 5 reflecting on practice
Factor 1: Planning and Instruction	1.595* (0.022)	1.185 (0.387)	1.311 (0.256)	1.508* (0.036)	1.456+ (0.084)
Factor 2: Analysis and feedback	0.929 (0.696)	0.941 (0.761)	0.914 (0.700)	0.926 (0.704)	0.883 (0.559)
Factor 3: Academic language	0.968 (0.856)	1.039 (0.848)	1.132 (0.549)	1.104 (0.647)	1.106 (0.615)
Std. Total score	1.636** (0.001)	1.225 (0.149)	1.571** (0.007)	1.645** (0.001)	1.653** (0.003)
Cases	172	172	172	172	172

Note: Cells report odds ratios from ordered logit models with p-values in parentheses. Models control for the percentage of minority and free and reduced-price lunch students at the school. +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.

References

Admiraal, W., Hoeksma, M., van de Kamp, M., & van Duin, G. (2011). Assessment of teacher competence using video portfolios: Reliability, construct validity, and consequential validity. *Teaching and Teacher Education, 27*(6), 1019–1028.

Bastian, K. C., Patterson, K. M., & Yi, P. (2015). *UNC teacher quality research: Teacher preparation program effectiveness report*. Available from https://publicpolicy.unc.edu/files/2015/07/2015-Teacher-Preparation-Program-Effectiveness-Report.pdf.

Brown, J. B. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter, 13*(3), 20–25.

Council for the Accreditation of Educator Preparation. (2013). *CAEP accreditation standards*. Available from: http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf.

Courtney, M. G. R. (2013). Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2. 0 to make more judicious estimations. *Practical Assessment, Research & Evaluation, 18*(8), 1–14.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Crowe, E. (2011). *Race to the Top and teacher preparation: Analyzing state strategies for ensuring real accountability and fostering program innovation*. Available from: http://files.eric.ed.gov/fulltext/ED518517.pdf.

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.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5–6), 523–545.

Dinno, A. (2012). *Paran: Horn's test of principal components/factors*. Retrieved from: http://cran.r-project.org/web/packages/paran/.

Dobson, E. E. (2013). *Examining the impact of early field experiences on teacher candidate readiness*. Order No. 3610779. East Carolina University. ProQuest Dissertations and Theses, 186. Available from: http://search.proquest.com.jproxy.lib.ecu.edu/docview/1500831604?accountid=10639, 1500831604.

Duckor, B., Castellano, K. E., Tellez, K., Wihardini, D., & Wilson, M. (2014). Examining

the internal structure evidence for the performance assessment for California teachers: A validation study of the elementary literacy teaching event for Tier I teacher licensure. *Journal of Teacher Education, 65*(5), 402–420.

edTPA. (2015). *Educative assessment and meaningful support: 2014 edTPA administrative report*. Available from: https://secure.aacte.org/apps/rl/resource.php?resid=558&ref=edtpa.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.

Federal Register. (2014). *Teacher preparation issues*. Available from: https://www.federalregister.gov/articles/2014/12/03/2014-28218/teacher-preparation-issues.

Gansle, K. A., Noell, G. H., & Burns, J. M. (2012). Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education, 63*(5), 304–317.

Grossman, P., Wineburg, S. S., & Woolworth, S. (2001). Toward a theory of teacher community. *Teachers College Record, 103*(6), 942–1012.

Halpin, P. F., & Kieffer, M. J. (2015). Describing profiles of instructional practice: A new approach to analyzing classroom observation data. *Educational Researcher, 44*(5), 263–277.

Henry, G. T., Campbell, S. L., Thompson, C. L., Patriarca, L. A., Luterbach, K. J., Lys, D. B., et al. (2013). The predictive validity of measures of teacher candidate programs and performance: Toward an evidence-based approach to teacher preparation. *Journal of Teacher Education, 64*(5), 439–453.

Henry, G. T., Kershaw, D. C., Zulli, R. A., & Smith, A. A. (2012). Incorporating teacher effectiveness into teacher preparation program evaluation. *Journal of Teacher Education, 63*(5), 335–355.

Henry, G. T., Thompson, C. L., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Marcus, J. V., et al. (2011). *UNC teacher preparation program effectiveness report*. https://publicpolicy.unc.edu/files/2015/07/UNC-Teacher-Preparation-Program-Effectiveness-Report-July-2011.pdf.

Henry, G. T., Thompson, C. L., Fortner, C. K., Zulli, R. A., & Kershaw, D. C. (2010). *The impact of teacher preparation on student learning in North Carolina public schools*. Available from https://publicpolicy.unc.edu/files/2015/07/The-Impact-of-Teacher-Preparation-on-Student-Learning-in-NC-Public-Schools-Jan-2010.pdf.

Horn, J. L. (1965). A rationale and test for the number of factors in factor-analysis. *Psychometrika, 30*(2), 179–185.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101–136.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2*(2), 130–144.

Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. MET Project*. Available from: http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf.

Longford, N. T., & Muthen, B. O. (1992). Factor-analysis for clustered observations. *Psychometrika, 57*(4), 581–597.

Miller, M., Carroll, D., Jancic, M., & Markworth, K. (2015). Developing a culture of learning around the edTPA: One university's journey. *The New Educator, 11*(1), 37–59.

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Washington, DC: National Academy Press.

Noell, G. H., & Burns, J. L. (2006). Value-added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57*(1), 37–50.

Noell, G. H., Porter, B. A., Patt, R. M., & Dahir, A. (2008). *Value added assessment of teacher preparation in Louisiana: 2004-2005 to 2006-2007*. Available from: http://www.laregentsarchive.com/Academic/TE/2008/Final%20Value-Added%20Report%20(12.02.08).pdf.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California teachers (PACT). *Journal of Teacher Education, 57*(1), 22–36.

Peck, C. A., Gallucci, C., Sloan, T., & Lippincott, A. (2009). Organizational learning the program renewal in teacher education: A socio-cultural theory of learning, innovation, and change. *Educational Research Review, 4*(1), 16–25.

Peck, C. A., & McDonald, M. A. (2014). What is a culture of evidence? How do you get one? And…should you want one? *Teachers College Record, 116*, 1–27.

Peck, C. A., Singer-Gabella, M., Sloan, T., & Lin, S. (2014). Driving blind: Why we need standardized performance assessment in teacher education. *Journal of Curriculum and Instruction, 8*(1), 8–30.

R Core Team. (2014). *R: A language and environment for statistical computing*. Retrieved from: http://www.R-project.org/.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84*(2), 126–136.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Available from: http://evaluation.ccsd59.org/D59_Evaluation_Page/Resources_for_the_Plan_files/Rethinking%20Teacher%20Evaluation%20in%20Chicago.pdf.

SAS Institute. (2011). *SAS 9.3*. Cary, NC: SAS Institute.

SCALE.. (2013). *2013 edTPA field test: Summary report*. Available from: http://edtpa.aacte.org/news-area/announcements/edtpa-summary-report-is-now-available.html.

Tennessee State Board of Education. (2012). *2012 report card on the effectiveness of teacher training programs*. Available from: http://www.tn.gov/thec/Divisions/fttt/12report_card/PDF%202012%20Reports/2012%20Report%20Card%20on%20the%20Effectiveness%20of%20Teacher%20Training%20Programs.pdf.

Tennessee State Board of Education. (2013). *2013 report card on the effectiveness of teacher training programs*. Available from: http://www.tn.gov/thec/Divisions/fttt/13report_card/1_Report%20Card%20on%20the%20Effectiveness%20of%20Teacher%20Training%20Programs.pdf.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Youngs, P., & Bird, T. (2010). Using embedded assessments to promote pedagogical reasoning among secondary teaching candidates. *Teaching and Teacher Education, 26*(2), 185–198.