# EVALUATING A PILOT OF THE TEACHER PERFORMANCE ASSESSMENT:

## *The Construct Validity, Reliability, and Predictive Validity of Local Scores*

Kevin C. Bastian, UNC-Chapel Hill
Gary T. Henry, Vanderbilt University
Yi Pan, UNC-Chapel Hill
Diana Lys, East Carolina University

EPIC EDUCATION POLICY INITIATIVE *at* CAROLINA

# Evaluating a Pilot of the Teacher Performance Assessment:

# The Construct Validity, Reliability, and Predictive Validity of Local Scores

Kevin C. Bastian
*University of North Carolina at Chapel Hill*
*Education Policy Initiative at Carolina*

Gary T. Henry
*Vanderbilt University*
*Peabody College*

Yi Pan
*University of North Carolina at Chapel Hill*
*Frank Porter Graham Child Development Institute*

Diana Lys
*East Carolina University*
*College of Education*

# Contents

Acknowledgements ..... i
Abstract ..... ii
Introduction ..... 1
Background ..... 3
TPA Constructs ..... 3
East Carolina University and TPA ..... 4
Measures, Data, and Samples ..... 5
TPA Scores ..... 5
Teacher Outcome Measures ..... 7
Analytical Methods ..... 8
Factor Analysis—Construct Validity ..... 8
Comparing Local and Official Scores—Reliability ..... 9
Teacher Outcome Analyses—Predictive Validity ..... 9
Results ..... 10
Factor Analysis ..... 10
Comparing Locally-Scored and Officially-Scored TPA ..... 12
Teacher Outcomes ..... 13
Discussion ..... 19
References ..... 21
Appendix ..... 24

**Acknowledgements**

We are grateful to the faculty and staff at East Carolina University, especially Ken Luterbach, 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.

Education Policy Initiative at Carolina (EPIC)
University of North Carolina at Chapel Hill
Abernethy Hall, CB #3435, Chapel Hill, NC 27599-3435
919-962-0668 publicpolicy.unc.edu

**Evaluating a Pilot of the Teacher Performance Assessment:**
**The Construct Validity, Reliability, and Predictive Validity of Local Scores**

Kevin C. Bastian, UNC-Chapel Hill
Gary T. Henry, Vanderbilt University
Yi Pan, UNC-Chapel Hill
Diana Lys, East Carolina University

January 2015

## Abstract

Locally-scored teacher performance assessment (TPA) portfolios offer teacher preparation programs (TPP) formative data on program performance, a common language of practice, and opportunities to enact evidence-based reforms. Before instituting changes based on locally-scored portfolios, however, TPP need to know how well faculty score candidates' portfolios and whether local scores predict candidates' performance as teachers-of-record. To address these questions we partnered with East Carolina University to examine the construct validity, reliability, and predictive validity of their 2011-12 graduates' locally-scored TPA portfolios. Results indicate that locally-scored portfolios identify three factors partially-aligned with the three constructs of TPA and more closely-aligned with the TPA cross-cutting themes. While local TPA scores were significantly higher than official (Pearson) scores, local score constructs significantly predicted teachers' evaluation ratings, and in some models, teachers' value-added. Overall, these analyses provide a framework for continued TPA research and highlight the potential utility of locally-scored portfolios for evidence-based reform.

## Introduction

In recent years public concern for the quality of teachers and teacher education has pushed policymakers and accreditation agencies to hold teacher preparation programs (TPP) accountable for the effectiveness of their graduates (Crowe, 2011). For example, shortly after the implementation of No Child Left Behind, states such as Louisiana, North Carolina, and Tennessee initiated efforts to link teachers' value-added scores to the TPP from which they graduated (Noell & Burns, 2006; Noell, Porter, Patt, & Dahir 2008; Gansle, Noell, & Burns, 2012; Henry, Thompson, Fortner, Zulli, & Kershaw, 2010; Henry, Thompson, Bastian, Fortner, Kershaw, Marcus, & Zulli, 2011; Henry, Patterson, Campbell, & Pan, 2013; TSBE, 2012, 2013). More recently, the federal Race to the Top competition mandated that states seeking funds commit to publicly reporting TPP's effectiveness and closing (expanding) low (high) performing TPP (Crowe, 2011; Henry, Smith, Kershaw, & Zulli, 2012). Likewise, the Council for the Accreditation of Educator Preparation (CAEP) requires TPP to demonstrate the impact of their graduates on student learning, classroom instruction, and employer satisfaction (CAEP, 2013).

In response to these policies and the desire of teacher educators to prepare more effective beginning teachers, TPP have begun to reform their preparation practices. Ultimately, the success of these reforms will be judged, at least in part, by the value-added scores of TPP graduates. However, such measures of teacher effectiveness are insufficient, by themselves, to provide TPP the necessary evidence to initiate reforms for two reasons. First, value-added scores come too late to TPP—there are often several years between the time a teacher candidate enters a TPP and the time they enter the workforce and impact student test scores. Second, while indicating teachers' effectiveness, value-added scores do not provide information about graduates' teaching practices that would allow TPP faculty and staff to identify systemic strengths and weaknesses. While some states and districts use multiple measures—value-added, classroom observations, evaluation ratings—to examine teacher performance, these still suffer from the first problem—they come too late to formatively guide TPP reform.

Like measures of in-service teacher performance, traditional measures of teacher candidate performance do not provide TPP with evidence to drive program improvement. Specifically, research indicates that available measures of teacher candidate performance, such as GPA, Praxis I scores, disposition ratings, student teaching ratings, and course taking do not predict the effectiveness of graduates, as measured by value-added scores (Henry, Campbell, Thompson, Patriarca, Luterbach, Lys, & Covington, 2013). Instead, for TPP that want to make evidence-based program reforms, the most promising measures are teacher candidate performance assessments, usually labeled Teacher Performance Assessments (TPA). As argued by Peck and colleagues, locally-scored TPA 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). While there are advantages of greater standardization and independence in the TPA scoring process—e.g. using the scores for credentialing teachers—local TPA scoring offers TPP formative data to drive program improvement efforts.

Regardless of the scoring process, TPA scores must: (1) measure the constructs that they were designed to measure (construct validity) (Duckor, Castellano, Tellez, Wihardini, & Wilson, 2014); (2) be reliably scored by different individuals (reliability); and (3) predict the teacher candidates' performance as classroom teachers (predictive validity). Extant research suggests that TPA can be the fulcrum that leverages an evidence-based culture, however, without valid and reliable data, the TPA evidence may not guide TPP to adopt or adapt more effective preparation practices (Peck & McDonald, 2014; Peck Gallucci, Sloan, & Lippincott, 2009). Therefore, in this study, we evaluate the validity and reliability of a locally-scored TPA pilot in the College of Education at East Carolina University (ECU). Three research questions guide this study:

(1) Do locally scored TPA validly measure the constructs that the TPA was designed to measure?
(2) How do local TPA scores compare with official (Pearson) TPA scores?
(3) Do the measures extracted from the local TPA scores predict entry into or exit from teaching, evaluation ratings of teachers' performance, or value-added scores?

The instrument ECU piloted for this study is a widely used TPA that was developed by Stanford University and aligned with Common Core, CAEP, and Interstate Teacher Assessment and Support Consortium (InTASC) standards, in which pre-service teachers create a portfolio designed to demonstrate their readiness to enter the teaching profession. This portfolio, based on a period of three to five days of instruction during the student teaching experience, 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. While the TPA used in this pilot has been replaced by the edTPA—recently field-tested by its developers (Stanford Center for Assessment, Learning, and Equity (SCALE), 2013)—this study makes three contributions to the research literature. First, this study provides an independent evaluation of a pilot TPA with a focus on external teacher outcomes—entry into and exit from the profession, teacher evaluation ratings, and teacher value-added. Second, this study compares local TPA scores to official (Pearson) scores and subjects the locally-scored measures to tests that will allow TPP to judge the adequacy and utility of TPA scores as a guide for program reforms. This is especially important given the centrality of local scoring in the current research on TPP reform and establishing a culture of evidence within TPP (Peck & McDonald, 2014; Peck et al., 2014; Peck et al., 2009). Finally, this study serves as a "proof of concept" for the type of study that individual TPP or collections of programs can undertake to establish the utility of local TPA or edTPA scoring to guide program improvement efforts.

In the sections that follow, we first provide further background on TPA and the pilot undertaken at ECU. Next, we detail the TPA and other data used in analyses and our research methods to address each research question. Finally, we present the results of our analyses and close with a discussion of the implications of our work for TPP, policy, and further research.

## Background

### *TPA Constructs*

As shown in the construct blueprint in Table 1, in 2011-12 the TPA consisted of 12 standards—planning for content understanding, knowledge of students for planning, planning for assessment, engaging students, deepening student learning, analysis of student learning, feedback, using assessment results, analysis of teaching, language demands, language supports, and language use—each of which aligned with one of the three main TPA constructs of planning, instruction, and assessment. In addition, five of the TPA standards were cross-listed with a TPA cross-cutting theme: analysis of teaching or academic language. Based on the construct blueprint, it is possible that the TPA standards may have five underlying factors, with five standards contributing to the planning construct, three standards contributing to the instruction construct, four standards contributing to the assessment construct, two standards contributing to the analysis of teaching cross-cutting theme (one of which contributes to instruction and the other to assessment), and three standards contributing to the academic language cross-cutting theme (two of which contributing to planning and the other to assessment). 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 standards. For example, using locally-scored portfolios from the Performance Assessment for California Teachers (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, Castellano, Tellez, Wihardini, & Wilson, 2014). Given the complexity of the TPA constructs and cross-cutting themes, which makes it difficult to predict the underlying factor structure of the 2011-12 TPA portfolio scores, we used exploratory factor analysis (discussed in the analytical methods section) to reveal the actual structure of the standards. Below, we further detail ECU's history with TPA, particularly its training of local scorers and local scoring processes in 2011-12.

**Table 1: Construct Blueprint for Teacher Performance Assessment Standards: Main and Cross-Cutting Constructs**

| 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<br>Knowledge of Students for Planning<br>Planning for Assessment | --- | Language Demands<br>Language Supports | 5 |
| Instruction | Engaging Students<br>Deepening Student Learning | Analysis of Teaching | --- | 3 |
| Assessment | Analysis of Student Learning<br>Feedback | Using Assessment Results | Language Use | 4 |
| Count of Main Only and Cross-cutting Standards | 7 | 2 | 3 | 12 |

*Note: This table places each of the standards into the main construct and, when applicable, into the cross-cutting theme as designated in the TPA blueprint.*

### *East Carolina University and TPA*

As a way to establish a culture of evidence and make formative improvements to teacher preparation practices, ECU began piloting the TPA during the 2009-10 academic year in the handbook areas of middle childhood (mathematics, English, science, and history-social studies) and secondary English and history-social studies. In 2011-12, ECU greatly expanded the pilot to include the handbook areas of elementary and special education, with local evaluators scoring candidates' portfolios. The majority of these local TPA evaluators were part-time faculty employed by ECU as university supervisors; additional TPA evaluators included tenured and tenure-track faculty, non-teaching faculty with public school teaching experience, and clinical teachers.

To prepare for local TPA scoring, ECU required each local evaluator to participate in nine hours of TPA training facilitated by SCALE-calibrated faculty (Dobson, 2013). In these sessions the TPA trainers first provided participants with a thorough description of each TPA standard and ECU's goals for TPA implementation. To calibrate the scoring of local evaluators, the TPA trainers provided the participants 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 standard. After the groups reached a consensus for each standard score, the TPA trainers revealed the official score and participants engaged in further discussion regarding the portfolio evidence. Finally, the TPA trainers offered session participants opportunities to practice the TPA scoring process using TaskStream™, ECU's electronic portfolio system.

Once the local evaluation training was complete, ECU assigned teacher candidates' TPA portfolios to the newly trained local scorers. To control for bias, ECU 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, ECU assigned no more than five TPA portfolios to any local scorer. During the initial portfolio scoring, if the local evaluator rated any standard in a candidate's portfolio as a 1 or 2 (on a scale from 1-5, with 1 indicating an unacceptable level of performance and 5 indicating exceptional performance), ECU assigned the candidate's portfolio to a secondary evaluator. Lead faculty in each content-specific program area conducted these secondary TPA evaluations and adjusted standard scores if evidence warranted a higher rating. After this secondary evaluation if any standard scores of 1 or 2 remained in a portfolio, lead faculty remediated the candidates and ECU allowed the candidates to consult with their university supervisor and/or clinical teacher and revise their portfolio for a third round of scoring. This revision process served as a valuable learning opportunity for candidates; furthermore, since candidates typically complete their TPA portfolio one-half to two-thirds of the way through student teaching, it also allowed candidates to demonstrate their ability to successfully complete teaching tasks at a later point in the student teaching experience. While ECU faculty examine both the initial and final TPA scores (and what led to improvement in scores between the scoring periods) in their program improvement efforts, this analysis only includes candidates' initial TPA scores. We believe these initial scores best capture candidates' own knowledge and ability to effectively carry out key teaching tasks.

Finally, in the spring of 2012, SCALE offered ECU the opportunity to submit TPA portfolios to Pearson as part of the national TPA field test. This offer came after ECU had completed all the 2012 local evaluations and after many teacher candidates had graduated. In response to this offer, ECU sent an email to teacher candidates who had completed a portfolio and invited them to submit their portfolio for official (Pearson) scoring. The timing of this offer limited teacher candidate participation, and of the 249 candidates with locally-scored portfolios, only 64 submitted and received official TPA scores.

## Measures, Data, and Samples

### *TPA Scores*

This evaluation of a TPA pilot relied on two sets of portfolio scores provided by ECU: (1) 249 locally-scored portfolios from the 2011-12 graduating cohort and (2) 64 Pearson-scored portfolios for a subset of 2011-12 graduates who also have locally-scored portfolios. The 249 portfolios locally-scored in 2011-12 measure the 12 TPA standards in effect during the 2011-12 academic year as displayed in Table 2. The TPA scores for this study include those from eight

different handbook areas—elementary literacy, special education, secondary English and history-social studies, and middle childhood mathematics, English, science, and history-social studies—which were scored by 75 different raters at ECU (an average of 3.32 portfolios per local rater). The 64 Pearson-scored portfolios from 2011-12 cover the 12 TPA standards in effect during the 2011-12 academic year[1] and come from seven different handbook areas—all those from the locally-scored portfolios except special education. Table 2 presents descriptive statistics—means and standard deviations—for the sample (n=249) of local scores from 2011-12; we present comparable data on the 2011-12 official (Pearson) scores in Table 4 (below).

**Table 2: Descriptive Statistics from 2011-12 Local TPA Scores (n=249)**

| Standard | TPA Construct (Cross-Cutting Theme) | Mean & Standard Deviation |
|---|---|---|
| Planning for Content Understanding | Planning | 3.29 (0.74) |
| Knowledge of Students for Planning | Planning | 3.19 (0.70) |
| Planning for Assessment | Planning | 3.35 (0.74) |
| Engaging Students | Instruction | 3.32 (0.73) |
| Deepening Student Learning | Instruction | 3.25 (0.71) |
| Analysis of Student Learning | Assessment | 3.26 (0.68) |
| Feedback | Assessment | 3.27 (0.74) |
| Using Assessment Results | Assessment (Analysis of Teaching) | 3.26 (0.74) |
| Analysis of Teaching | Instruction (Analysis of Teaching | 3.21 (0.73) |
| Language Demands | Planning (Academic Language) | 3.06 (0.67) |
| Language Supports | Planning (Academic Language) | 3.24 (0.69) |
| Language Use | Assessment (Academic Language) | 3.18 (0.73) |

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

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

***Teacher Outcome Measures***

We include four teacher outcome measures in this study: (1) entry into teaching; (2) teacher attrition; (3) teacher evaluation ratings, and (4) teacher value-added scores. The full analysis sample includes all 249 graduates of ECU 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.

*Entry into Teaching:* 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 ECU graduates with local TPA scores, of which 181 taught in NCPS in 2012-13.

*Teacher Attrition:* Contingent on entering the state's public school teacher workforce in 2012-13, this analysis examines the relationship 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 did not return to NCPS in 2013-14.

*Teacher Evaluation Ratings:* The dependent variable for this analysis comes 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 formal classroom observations and paper-based evidences to document key teaching behaviors and rate teachers as either: not demonstrated, developing, proficient, advanced, 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 with local TPA scores who both taught in NCPS in 2012-13 and were evaluated by a school administrator.[2]

*Teacher Value-Added Scores:* To examine whether local TPA scores predict teacher value-added, we relied on teachers' EVAAS (Education Value-Added Assessment System) estimates 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

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

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 teachers' EVAAS estimates are the dependent variable and the sample includes 114 EVAAS estimates—61 MRM estimates and 53 URM estimates—for 76 unique teachers with local TPA data who taught a tested-grade/subject in 2012-13.

## Analytical Methods

### *Factor Analysis—Construct Validity*

Due to the complexity of the construct blueprint, we implemented exploratory factor analysis (EFA) to examine the underlying factor structure of 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 and somewhat outdated 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 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, ECU'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 standards (six at the 0.05 level and two at the 0.10 level). Given this evidence of non-ignorable between-rater variation in TPA standard scores, we followed the suggestion of Reise and colleagues and group mean-centered TPA scores and then conducted EFA on this 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).

### *Comparing Local and Official Scores—Reliability*

To compare the local versus official (Pearson) TPA scoring, for each TPA standard we examined the correlations between the local and official scores. We also used paired t-tests to assess whether there were significant differences in the mean values for the two sets of scores. The sample of 64 TPA portfolios with official scores was too small to conduct an EFA.

### *Teacher Outcome Analyses—Predictive Validity*

To understand the relationship 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 standards. 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 specification—to assess the multi-variate relationship between teacher outcomes and local TPA scores. In these regression models the focal variables were either the TPA constructs identified by factor analysis or the standardized total score across all 12 locally-scored TPA standards. Below, we detail our regression approaches to address each research outcome.

*Entry into the Teaching and Teacher Attrition:* 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 how a one standard deviation increase in a factor or the total score impact the odds of workforce entry or exit.

*Evaluation Ratings:* To determine whether local TPA scores predict teachers' evaluation ratings, we specified separate ordered logistic regression models for each of the five NCEES standards, 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 how a one

standard deviation increase in a factor or the total score impact the odds of rating one level higher on the NCEES.

*Teacher Value-Added:* 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.[3] We then performed separate analyses for the MRM and URM data. In this way we estimate the relationship 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 on 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 how a one standard deviation increase in a factor or the total score predict teachers' value-added to student achievement.

## Results

### *Factor Analysis*

Following Reise and colleagues, we began by using EFA to examine the factor structure of the group mean-centered local TPA scores (Reise, Ventura, Nuechterlein, & Kim, 2005). Parallel analysis on this group mean-centered data revealed a three factor structure. Table 3 shows that TPA standards 1-5 loaded onto the first factor, TPA standards 6-9 loaded onto the second factor, and TPA standards 10-12 loaded onto the third factor. Comparing the TPA constructs with the 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, Castellano, Tellez, Wihardini, & Wilson, 2014).

The first factor contains three standards in the planning construct—planning for content understanding, knowledge of students for planning, and planning for assessment—and two standards 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 standards from the assessment construct—analysis of student learning, feedback, and using assessment results—and one standard from the instruction construct—analysis of teaching. We refer to this second factor as *Analysis and Feedback* and note that two of the standards, 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* standard also loaded with the assessment construct rather than the instruction construct (SCALE, 2013). Finally, the third factor contains two standards from the planning construct—language demands and language supports—and one standard from the assessment construct—language use. Given that

[3] 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.

these three TPA standards comprise the *Academic Language* cross-cutting theme, we refer to the third factor as *Academic Language.*

**Table 3: Factor Loadings with the 2011-12 Local TPA Scores**

| | | Factor Loadings with Group Mean-Centered 2011-12 Local TPA Scores | | |
|---|---|---|---|---|
| **TPA Standard** | **TPA Construct (Cross-Cutting Theme)** | **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.*

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 studies detailed below, we examine the relationship between these three latent factors from the locally-scored TPA data and teacher outcomes.

### *Comparing Locally-Scored and Officially-Scored TPA*

To address our second research question we used bivariate correlation and paired t-tests to compare the TPA standard scores of the 64 ECU teacher candidates whose portfolios were both locally and officially (Pearson) scored. The left panel of Table 4 displays the correlations between each local and Pearson-scored standard. 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. Here, for all 12 TPA standards, the local scores were significantly higher than the official Pearson scores. Combined with the factor analyses, results indicate that local scoring is aligned with some of the key constructs/themes of TPA, however, local scores are significantly higher than official scores from Pearson.

**Table 4: Comparing Locally and Pearson-Scored Portfolios from 2011-12**

| **Correlations Between Locally-Scored and Pearson-Scored Portfolios** | | **Mean Standard Scores and Standard Deviations** | |
|---|---|---|---|
| **TPA Standard** | **Correlation** | **Locally-Scored** | **Pearson-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 Pearson-scored TPA portfolios. The right panel of this table displays the average TPA scores and standard deviations from the ECU portfolios that were both locally and Pearson-scored (n=64). +, *, and ** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively.*

### *Teacher Outcomes*

*Bivariate Correlations:* To begin, we examined the bivariate correlations between our TPA measures and the teacher outcomes (see 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. 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 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 End-of-Course exams, final exams, and $5^{th}$ and $8^{th}$ grade science exams.

**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.*

*Entry into the Teacher Workforce:* Based on multivariate logistic regression, the *Academic Language* factor significantly predicts entry into the NCPS teacher workforce (see Table 6). Holding factors 1 and 2 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 are 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.

**Table 6: Workforce Entry and Attrition Outcomes**

| | **Becomes a Teacher in NCPS** | **Exits NCPS** |
|---|---|---|
| Factor 1: Planning and Instruction | 0.729<br>(0.105) | 0.931<br>(0.816) |
| Factor 2: Analysis and Feedback | 1.047<br>(0.798) | 0.808<br>(0.416) |
| Factor 3: Academic Language | **1.666***<br>(0.021) | 0.844<br>(0.604) |
| | | |
| Std. Total Score | 1.232<br>(0.155) | 1.138<br>(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.*

*Teacher Attrition:* 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.

*Teacher Evaluation Ratings:* As shown in Table 7, the ordered logistic regression results 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 standards loading onto the *Planning and Instruction* factor.

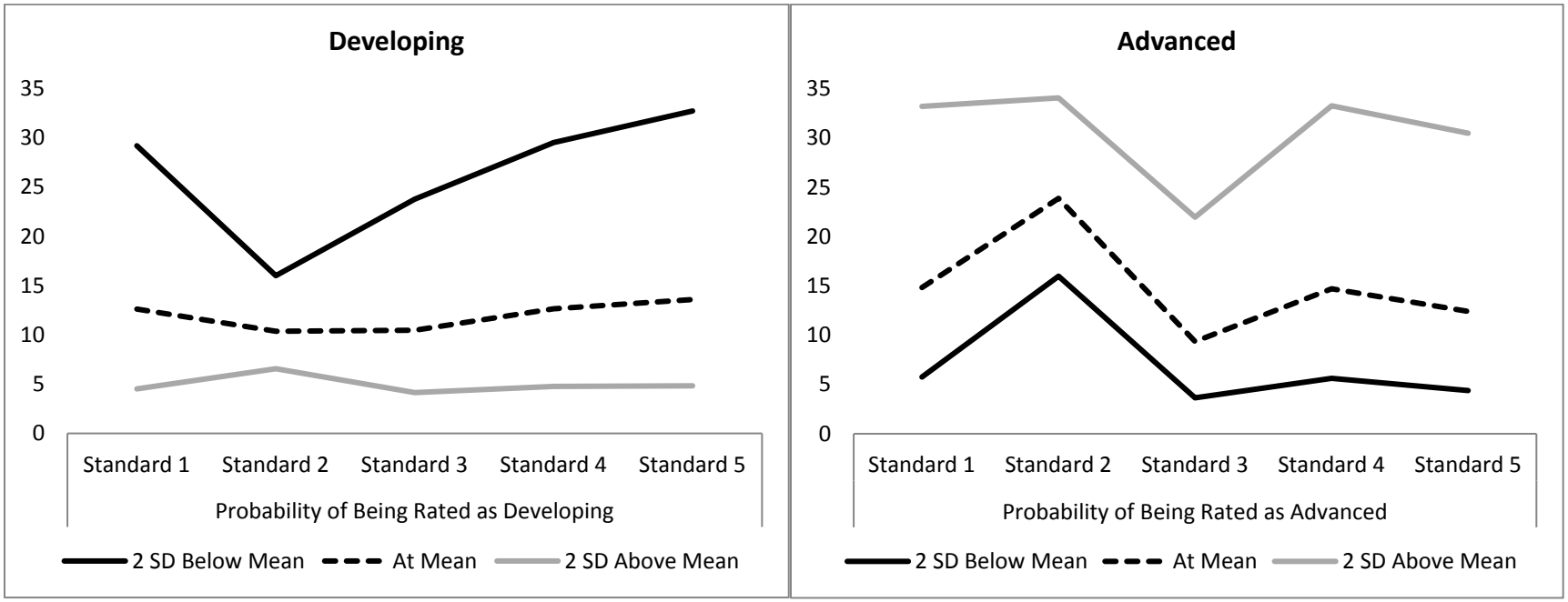
**Table 7: Teacher Evaluation Ratings in 2012-13**

| | **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.*

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, Figure 1 displays predicted probabilities for receiving an evaluation rating of developing or advanced 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 probability of receiving a rating of developing on Standard 4 and only a six percent probability of receiving an advanced rating; at the other end of the distribution, teachers with a total score two standard deviations above the mean have a five percent probability of rating at developing and a 33 percent probability of rating as advanced.

**Figure 1: Predicted Probabilities of Being Rated Developing and Advanced on the NCEES**




*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 advanced on Standards 1-5 of the NCEES.*

Finally, to investigate the extent to which these significant evaluation results may be due to teacher skill (as measured by TPA scores and evaluation ratings) rather than the sorting of teacher candidates with higher TPA scores into schools with more advantaged students, 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 can predict evaluation ratings when the candidates become teachers regardless of the school in which they teach.

*Teacher Value-Added:* The *Planning and Instruction* factor is significantly associated with teacher value-added in analyses limited to the URM estimates (see Table 8). Specifically, a one standard deviation increase in the *Planning and Instruction* factor is associated with students gaining an additional 1.4 normal curve equivalent points on their End-of-Grade, End-of-Course, and final exams in URM-eligible courses. When including control variables for the percentage of minority and free and reduced-price lunch students at the school, 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. Finally, in the overall and MRM analyses, neither the TPA factors nor the standardized total score significantly predict teacher effectiveness.

**Table 8: Teacher EVAAS Estimates in 2012-13**

| | **All EVAAS Estimates** | **MRM EVAAS Estimates** | **URM EVAAS Estimates** |
|---|---|---|---|
| Factor 1: Planning and Instruction | 0.629<br>(0.383) | -0.003<br>(0.525) | **1.388**$^{+}$<br>(0.724) |
| Factor 2: Analysis and Feedback | -0.285<br>(0.394) | 0.223<br>(0.492) | -0.748<br>(0.594) |
| Factor 3: Academic Language | -0.124<br>(0.412) | -0.280<br>(0.527) | -0.075<br>(0.682) |
| | | | |
| Std. Total Score | 0.155<br>(0.269) | -0.054<br>(0.476) | 0.335<br>(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.*

## Discussion

This study indicates that locally-scored TPA have a reasonable degree of construct validity, reliability, and predictive validity. While the TPA construct blueprint was not fully reproduced in the EFA, we found factors that corresponded to expected constructs using the rater mean-centered TPA data. In addition, and perhaps most importantly, constructs from the locally-scored TPA portfolios predicted teachers' entry into the teaching profession, evaluation ratings, and value-added scores. We believe these findings underscore the potential of locally-scored TPA to provide TPP with data upon which they can create a culture of evidence and adopt/adapt reforms to preparation practices.

Furthermore, these results suggest promise for the transition from TPA to edTPA. By eliminating TPA's two cross-cutting themes of *Analysis of Teaching* and *Academic Language*, edTPA has a simpler and clearer construct blueprint, now focused only on the three domains of planning, instruction, and assessment. Field testing of the edTPA standards suggests a three-factor solution that is consistent with the three construct blueprint or a one-factor solution that is consistent with the overall edTPA scoring procedure (SCALE, 2013). Given the significant relationships found in this pilot study between the *Planning and Instruction* factor and the TPA total score and teacher evaluation ratings, researchers should conduct similar evaluations with the edTPA official (Pearson) and local scores. If both the official and local edTPA scores return positive predictive validity results, TPP can embrace edTPA scores as a valid data source with which to help build a culture of evidence and formatively assess preparation reforms. Additionally, states and TPP can use official edTPA scores as a requirement for teacher certification. If either type of scoring does not return positive predictive validity results, efforts can be focused on improving the other method of scoring.

Regarding the use of local TPA scores for high-stakes teacher licensure decisions, these results suggest that it may be inappropriate to use locally-scored TPA portfolios. The local scores were systematically higher than the official scores for the same candidates. Furthermore, the local scores varied from rater to rater in ways that may be attributable to differences in teacher candidate quality and/or differences in the raters. Given this analysis, it seems advisable to employ local scoring to provide a language, context, and forum for evidence-based program reform and official scoring for credentialing (if states/TPP use TPA scores as a requirement for certification).

Even with our pilot study findings that support TPA, the limited correlations between the factors or the total score and value-added remain a concern. This study was limited to a total sample of 249 teacher candidates with even fewer teachers having value-added scores (n=76 with 114 value-added scores). With this sample size it seems promising that we found positive correlations between the *Planning and Instruction* factor and teacher value-added. The small sample, however, likely influenced the reliability of estimates and further research, with a larger number of teachers, is necessary.

While studies such as this one may be delayed to await more data—from either more graduating cohorts or more TPP—and the full implementation of edTPA (or other performance assessments designed to measure teacher candidates' readiness to enter the profession), we believe this study should encourage evaluating the validity of teacher candidate performance assessments as soon as possible. Such research does not condone making definitive conclusions regarding the utility of performance assessments with limited data, but it does support the establishment of an evidence-based culture within TPP that respects the criteria of construct validity, reliability of scoring, and predictive validity.

**References**


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.

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

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. In press, *Journal of Teacher Education*.

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.

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.

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. Carolina Institute for Public Policy. Available from: http://publicpolicy.unc.edu/files/2014/02/ImpactTeacherPrepPro_Jan2010_Final.pdf

Henry, G.T., Thompson, C.L., Bastian, K.C., Fortner, C.K., Kershaw, D.C., Marcus, J.V., & Zulli, R.A. (2011). UNC teacher preparation program effectiveness report. Carolina Institute for Public Policy. Available from: http://publicpolicy.unc.edu/files/2014/02/TeacherPrepEffectRpt_Final_2011.pdf

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., Campbell, S.L., Thompson, C.L., Patriarca, L.A., Luterbach, K.J., Lys, D.B., & Covington, V. (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., Patterson, K.M., Campbell, S.L., & Yi, P. (2013). UNC teacher quality research: 2013 teacher preparation program effectiveness report. Education Policy Inititative at Carolina. Available from: http://publicpolicy.unc.edu/files/2013/11/UNC_TQR_OverallProgramReport_Final.pdf

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

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

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

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.

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.

## Appendix

### Appendix Table 1: Between-Rater Variance in Local TPA Standard Scores

| TPA Standard | 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.*

**Appendix 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* | *Advanced* | *Developing* | *Advanced* | *Developing* | *Advanced* | *Developing* | *Advanced* | *Developing* | *Advanced* |
| 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 advanced on Standards 1-5 of the NCEES.*

**Appendix 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.*

UNC
COLLEGE OF
ARTS & SCIENCES

publicpolicy.unc.edu