Documenting Teacher Candidates' Professional Growth through Performance Evaluation

Elizabeth Levine Brown*
Jennifer Suh
Seth A. Parsons
Audra K. Parker
Erin M. Ramirez

George Mason University

* Correspondence concerning this article should be addressed to Elizabeth L. Brown, George Mason University, College of Education and Human Development, 4400 University Avenue, MS 4B3, Thompson Hall 1804, Fairfax, VA 20030. E-mail: ebrown11@gmu.edu

Abstract

In the United States, colleges of education are responding to demands for increased accountability. The purpose of this article is to describe one teacher education program's implementation of a performance evaluation tool during final internship that measures teacher candidates' development across four domains: Planning and Preparation, Instruction and Management, Assessment, and Personal and Professional Development. Researchers examined data collected via midpoint and final internship evaluations across three program tracks using a measure created by the authors entitled the Profile for Evaluation of Intern (PEI). Although this measure is still in its preliminary phases, data analyses indicated positive, statistically significant differences across three tracks on 16 criteria on the performance evaluation tool and on candidates' overall 'grand' averages at the midpoint evaluation; however, no statistically significant differences remained at the final evaluation point. Benefits and challenges involved in employing a performance evaluation tool in teacher preparation are discussed. Implications for teacher educators, including recommendations for programmatic design and suggestions for how such tools inform program development and the field of teacher education, are discussed.

Keywords: Performance evaluation, teacher preparation, professional development schools, teacher candidates, assessment

In 2010, the National Council for the Accreditation of Teacher Education (NCATE) released its *Report of the Blue Ribbon Panel on Clinical Preparation and Partnerships for Improved Student Learning*. In it, they recommended, "turning the education of teachers 'upside down'" (NCATE, 2010, p. 2) via significant changes to how colleges of education "deliver, monitor, evaluate, oversee, and staff clinically based preparation" (NCATE, 2010, p. iii). In addition to recommendations for clinically based teacher preparation, partnership development with K-12 schools, and expanded research efforts, the Blue Ribbon Panel asserted the need for establishing high standards for and rigorous accountability of teacher education programs. Specifically, they called for the use of multiple sources of data to continuously evaluate candidates' and programs' effectiveness.

In response to this call, many states and colleges of education worked to design meaningful teacher evaluation systems. For example, after 1998 legislation required California teacher preparation programs to use performance evaluations in credentialing decisions, California teacher education programs became leaders in designing and using such evaluations (Pecheone & Chung, 2006). Stanford University led these efforts, creating the Performance Assessment for California Teachers (PACT) (Sandholtz & Shea, 2012) and later partnering with the American Association of Colleges for Teacher Education (AACTE) to develop edTPA. The edTPA provides a framework for teacher candidates to document their development in planning, instruction, assessment, and reflection (http://edtpa.aacte.org/). In-service teachers and teacher educators evaluate teacher candidates' edTPA submissions to ascertain their readiness for teaching. This system provides colleges of education with candidate performance data for program evaluation.

Using this foundational research, teacher educators at a large research university in the mid-Atlantic region of the United States developed, implemented, and tested a performance evaluation tool to assess their teacher candidates' professional growth. This elementary education program is organized around reciprocal partnerships with K-6 schools in three surrounding schools districts. Using a Professional Development Schools (PDS) model, the program and its K-6 partners collaborate to positively impact elementary teacher education, K-6 student learning, and in-service teacher professional development (Holmes Group, 1990; Neapolitan, 2011).

Teacher candidates in this program are enrolled in one of three different tracks, which are described in detail below. Using the program evaluation tool to assess candidates' maturation, the teacher educators examined areas where candidates excel and struggle as well as comparisons within and across the three tracks. Additionally, the authors explored how the data contribute to ongoing conversations about teacher education accountability and the evaluation of teacher candidates.

Background

This section begins by exploring the complexity in learning to teach. It then reviews the literature on field-based teacher preparation. Last, the literature on using performance assessments in teacher preparation programs is reviewed.

Complexities of Learning to Teach

Learning to teach is challenging. Teachers must simultaneously develop understandings of content, pedagogy, and child development and implement these understandings in a multifaceted K-12 context (Lampert et al., 2013). In their historical review, Hammerness, Darling-Hammond, Grossman, Rust, and Shulman (2005) identified three challenges to learning to teach and presented principles of learning for each to support candidates' development. First, candidates must begin thinking about teaching and learning from the perspective of teacher, which can often be quite different from their previous experience. Complicating this process is the enduring power of the "apprenticeship of observation," which is the phenomenon of individuals entering teacher preparation programs having spent numerous hours, through their K-12 education, observing teaching; they, therefore, believe they have a strong understanding of effective practice (Lortie, 1975). In this apprenticeship of observation, which does not often occur in other professions, future teachers form powerful ideas—and often misconceptions—of teaching and learning that shape their subsequent professional development.

Second, candidates must develop both the skill of thinking like a teacher and the ability to put their knowledge into action (Hammerness et al., 2005). This situation presents the "problem of enactment" where teachers have an understanding of content and pedagogy, but are unable to retrieve this information in the moment to put it into action (Kennedy, 1999). Enactment is facilitated when candidates have rich factual knowledge, an understanding of how this knowledge fits in the bigger picture, and the ability to organize this knowledge in a manner that supports quick retrieval and action (Hammerness et al., 2005). Finally, teacher candidates face the "problem of complexity," which requires them to make numerous decisions about students' academic, social, emotional, and behavioral needs simultaneously (Hammerness et al., 2005). Hammerness et al. assert that developing teacher candidates' metacognitive skills can enable them to better manage the complexities of K-12 classroom decision-making.

Field Experiences in Teacher Candidate Preparation: A Focus on PDSs

Field experiences are a critical component of teacher preparation as they provide candidates real-world contexts where they can navigate the previously described challenges of learning to teach (e.g., Cohen, Hoz, & Kaplan, 2013; Hollins, 2015; Zeichner, 2010). Often these field-based partnerships are fostered in Professional Development School (PDS) sites. PDS teacher education experiences typically include extensive field experience, high-quality supervision and mentoring, rich engagement with school faculty in planning and instruction, opportunities for participation in inquiry, and strong theory-to-practice connections (Damore, Kapustka, & McDevitt, 2011; NAPDS, 2008; Sandholtz & Wasserman, 2001). Teacher candidates who graduate from PDS programs demonstrate more effective instructional techniques, management, and assessment than candidates who graduated from traditional teacher education programs (Castle, Fox, & Souder, 2006; Ridley, Hurwitz, Hackett, & Miller, 2005).

The Use of Performance Evaluation Tools in Teacher Education

Defining and evaluating teacher effectiveness is a perpetual difficulty in the field of education (Margolis & Doring, 2013; Mascarenhas, Parsons, & Burrowbridge, 2010; Sandholtz & Shea, 2012). Paper and pencil tests of teachers' content or pedagogical knowledge, like other distal measures (e.g., SAT scores, GPAs, etc.), have proven to be ineffective in capturing teacher quality (Sandholtz & Shea, 2012). These evaluations "serve to trivialize and undermine our understanding of the complexity of teachers' work and diminish the critical role of teacher education in preparing teachers" (Pecheone & Chung, 2006, p. 33). Further, research demonstrates that evaluation tools are more effective than indirect tests in predicting teacher candidates' future classroom performance (Uhlenbeck, Verloop, & Beijaard, 2002).

Accordingly, performance evaluations have become increasingly popular in assessing teacher candidates' holistic development (Margolis & Doring, 2013). Performance evaluations include evidence of teachers' practice in the classroom while valuing the contextualized and unpredictable nature of classroom instruction (Darling-Hammond & Snyder, 2000), which provides a practically valid evaluation of teachers' work. Further, performance evaluations provide an integrated view of teacher knowledge and practice, which addresses a common critique of teacher education assessment as piecemeal and disconnected from actual practice (Darling-Hammond & Snyder, 2000; NCATE, 2010) as well as serve as powerful professional learning experiences for teacher candidates, mentor teachers, and university supervisors (Pecheone & Chung, 2006). For these reasons and in extension of NCATE's (2010) *Report of the Blue Ribbon Panel*, this study defines program evaluation tools as developmental measures that assess teacher candidates' knowledge, understanding, and execution of key teaching and learning mechanisms, specifically in the areas of planning, instruction, assessment, and reflection. Typically, these are summative measures occurring at the *end point* of a teacher candidates teacher education program, which may limit programs overall understanding of candidates' longitudinal development.

One such example of a summative assessment is edTPA, currently the most prominent performance evaluation, as it serves to evaluate teacher candidates' overall basic teaching skills and subject matter knowledge *prior* to entry into the profession (About edTPA, 2015). Proponents note several strengths of edTPA including its standards-based approach and its potential for unifying teacher preparation behind a common definition core of knowledge/pedagogy. In contrast, critics assert edTPA, is an unnecessary and unwelcome corporate influence in teacher preparation that may strive to standardized teacher preparation (Sawchuck, 2013). Further, this summative assessment does not address teacher candidates' performance over time. Not all states utilize edTPA as an assessment tool, providing opportunity for some teacher education programs, like the one involved in the current study, to develop their own performance evaluation tool grounded within the framework of their program (i.e., PDS) and assessing development over time.

The primary purpose of this study was to explore teacher candidates' performance over time and across different program tracks. In addition, we sought to understand how our performance evaluation tool might support our understanding of teacher candidates' professional growth and inform programmatic design. Unlike many performance evaluation tools, this study investigated performance at two data points: the midterm and final evaluations of teacher candidates across three program tracks. This analysis explored the usefulness of and issues within this evaluation tool for documenting these teacher candidates' professional development as well as illuminated how performance evaluations tools generally support teacher education. By analyzing trends in the criteria both within and across cohorts in different program tracks and across two data points, we highlighted areas in which teacher candidates develop quickly and less quickly. The research questions below guided this study:

1 . In what areas of the evaluation tool do teacher candidates receive high ratings and in what areas do they receive low ratings?
2 . How do teacher candidates' scores on the evaluation tool change from the first placement evaluation (midpoint) to the second placement evaluation (final)?
3. What are the differences among teacher candidates from different tracks within this elementary education PDS program?

Method

In this section, we first describe the context and sample for this research. Next, we describe the performance evaluation tool. Last, we describe how we analyzed the data to answer our research questions.

Context and Sample

The context for this research is a pre-service, graduate elementary education teacher preparation program housed in a college of education in a large, public university in the mid- Atlantic region of the United States. The sample consisted of 97 pre-service teacher candidates across three different program tracks. Candidate demographics across program tracks resemble national representation in teacher education programs, with the vast majority being white females from middle class backgrounds (Zumwalt & Craig, 2008). No significant demographic disparities exist across these three program tracks. All teacher candidates are required to complete 39 credit hours of coursework in content area methods, literacy methods, foundations, child development, differentiation, management, instructional planning, and technology.

Teacher candidates select one of three program tracks: Year-Long (YL; *n* = 18), Semester-Long (SL; *n* = 31) or Intensive (IN; *n* = 48), all of which are nested within a PDS model. Table 1 provides further information regarding each program track. The YL track involves coursework across six academic semesters. Teacher candidates complete 15-30 field hours in each of the first four semesters, as well as a two semester-long final internship. The SL track encompasses seven consecutive academic semesters of coursework. Teacher candidates complete 15-30 field hours in each of the first six semesters, as well as two eight-week placements during one semester for final internship. Finally, the IN track comprises five consecutive academic semesters of coursework. In the first three semesters, teacher candidates complete 15-30 field hours per semester. In their fourth semester of "heavy fieldwork," IN teacher candidates are placed in the field for three days a week. In their last semester, IN candidates complete two eight-week placements during one semester for final internship.

Although engaged in the same course content, internship experiences differ across program tracks. YL teacher candidates participate in a one-year internship incorporating two placements (one upper- and one lower-grade level) for a semester each; whereas the SL and IN tracks embark on a semester-long internship with two placements (one upper- and one lower-grade level) for eight weeks each. Across all internship experiences, teacher candidates engage in an increased amount of responsibility, shifting from a co-teaching model to independent teaching in each internship placement. All candidates are placed within a PDS site and with classroom teachers who have been trained as clinical faculty via a required university course. Incorporating performance evaluation as a means to assess teacher candidates' comprehensive professional development during internship experiences is a key component of the elementary education PDS program.

In our PDS model two key individuals evaluate teacher candidates: the clinical faculty member (CF) and the university facilitator (UF). As a classroom teacher, the CF works directly with the teacher candidate throughout the internship. The UF is a faculty member or university representative affiliated with the elementary education program with expertise in teacher education, pedagogy, and practice in elementary content. All UFs spend one day a week at their designated PDS site, observing and supervising teaching candidates, attending school functions, meeting with school leaders, and hosting professional development seminars for teacher candidates, which often include school-based teachers, leaders, and administrators. As a collaborative unit, the CF, UF, and teacher candidate each uses the program evaluation tool to evaluate the teacher candidate's overall classroom performance across four domains.

The Profile for Evaluation of Intern (PEI) Tool

Teacher candidates, CFs, and UFs evaluate teacher candidates' performance across four domains: Preparation and Planning, Instruction and Classroom Management, Assessment, and Professional Development. The Preparation and Planning domain has nine criteria for evaluation; the Instruction and Management domain has 15 criteria; the Assessment domain has eight criteria; and the Professional Development domain has eight criteria (see Appendix). Using the program's PEI tool, each teacher candidate, CF, and UF completes the profile independently and then discusses the evaluation as a trio. First, the three parties rate teacher candidates' performance for items that measure each domain using a 1 to 5 point scale (1 = Performance needs significant improvement, 5 = Performance is of notable excellence). Ratings of 1 or 2 indicate skills that require scaffolding and support on the part of the CF and UF in order for the teacher candidate to develop the appropriate level of expertise. Ratings of 4 or 5 suggest that the candidate's performance regarding a skill or disposition is exceptional. For state licensure, our program benchmark rating for successful completion of internship is an average 3.0 score across both placements.

When piloting this instrument, we collected inter-rater reliability information for three collaborative units (CF, UF, intern). The three coders individually rated the intern's performance across the four domains of evaluation and then met to discuss their ratings in an effort to triangulate the individual and final scores for each of the teacher candidates. When scores were exact (i.e., all three coders provided the same score), that score was entered as the final score; however, when scores were exact-adjacent or adjacent (i.e., at least one of the coders provided a score one point higher or lower than the other coders), the final score is aggregated from all three individual scores. Although rare, if specific scores were not exact, exact-adjacent, or adjacent, the collaborative team returned to review other documentation of intern's performance (e.g., biweekly reports, observation reports by CF and UF, lesson plans) and reexamined and rescored collectively to ensure that all three raters completely agreed on the final scoring. These procedures follow accepted inter-rater reliability approaches (as noted in Auerbach, La Porte, & Caputo, 2004). Last, the trio calculated an aggregated score for each domain.

Development of the PEI. Program faculty designed the PEI tool to meet the Association for Childhood Education International (ACEI) and state standards as well as to establish reliability across teacher candidate, CF, and UF ratings for all three program tracks. We are currently in the initial reliability-testing phase, with assessments showing an 80% agreement for randomized trios sampled across program tracks and using final internship scores. Further validation of the instrument will occur as we continue to evaluate candidates' development using the tool. We assessed the validity of the PEI tool across several key principles. Foremost, the tool addresses face validity by grounding evaluation items in literature and former empirical evidence to ensure that the PEI measures its intended constructs (Rubin & Babbie, 2007). To further establish content validity (Crocker & Algina, 1986), faculty across the program as well as key school-based stakeholders (e.g., principals, teacher leaders) contributed feedback on the tool items, domains, and scoring, which provided evidence on the relevance and representation of the items for each sub-scale. Because the sample investigated here represents the pilot testing of this tool, we are unable to present quantitative ecological validity of the measure; however, as we continue to test this tool, we plan to develop a large enough sample to run an exploratory factor analysis (EFA) on the measure and calculate both validity and reliability of the instrument.

Data Analysis

Data analyzed here used mid-point and final evaluation ratings for 97 teacher candidates across three program tracks during internship placements one and two. To highlight how the PEI tool assessed these candidates' development across the four domains of development (research question one) and in respect to the three program tracks, we first ran descriptive statistics to identify mean scores for the first and second placements of the four domains of development overall and for each program track. Next, we ran paired sample t-tests to determine whether differences existed between participants' placement one and two evaluation scores to address research question two.

Finally, to examine research question three, we conducted a one-way ANOVA with the independent variables being the tracks (i.e., YL, SL, and IN) and the dependent variable being the results for each of the 40 items across the four domains. This analysis compared the three tracks' scores across the 40 items as well as the overall average scores for each domain. An alpha level of .05 was used for all analyses. In line with quantitative analyses, all effect sizes for the paired sample t-test data are reported using Cohen's d, whereas all effect sizes for the one-way ANOVAs are reported using eta squared (η^2) (Lakens, 2013). When testing for assumption, Levene's Homogeneity of Variance statistic was violated, $p<.05$, for some of the criteria and thus a Welch test and Brown-Forsythe test were run on both placement one and two scores. Because the Welch test is more conservative, it was the statistic reported.

Results

In this section, we present results by research question. To review, those research questions are: 1) In what areas of the evaluation tool do teacher candidates receive high ratings and in what areas do they receive low ratings? 2) How do teacher candidates' scores on the evaluation tool change from the first placement evaluation (midpoint) to the second placement evaluation (final)? and 3) What are the differences among teacher candidates from different tracks within this elementary education PDS program?

Research Question 1 – Teacher Candidates' Strengths and Weaknesses

A detailed examination of the descriptive statistics for each of the evaluation criteria revealed commonalities among the tracks. When looking at the average across three tracks for each category, we selected the criteria that averaged less than a program satisfactory scoring of 3.0. Table 2 represents all mean differences for each domain.

For domain one, *Preparation and Planning*, in the first placement, three of the nine criteria were below 3.0. By the second placement, all teacher candidates from the three tracks averaged a mean above the satisfactory score of 3.0 on these criteria; however, these means were still lower than the other six criteria averages. The highest criterion was 7, *gathers creates and organizes materials and equipment in advance*, for both the first (*M* = 3.31) and second (*M* = 3.86) placement.

In the second domain, *Instruction and Management*, there were five criteria, out of 15, below the satisfactory score of 3.0 for the first placement. The two highest criteria were Criterion 20, *demonstrates courtesy and caring in relationships with students (M* = 3.55) and Criterion 23, *works toward developing a positive classroom community* (*M* = 3.46). For the third domain, *Assessment*, candidates in the three tracks scored below the satisfactory score of 3.0 for four out of eight criteria. Criterion 25*, uses assessment that matches the objective* (*M* = 3.15) represented the highest mean.

For domain four, *Personal and Professional Development*, all of the teacher candidates across all three tracks averaged above the satisfactory score of 3.0 during their first and second placements. The criteria with the highest and lowest means for this domain were Criterion 36, *welcomes assistance for improvement* (*M* = 3.69) and Criterion 38, *can develop and explain professional judgments* (*M* = 3.18) respectively. When examining the mean scores between placement one and two for all the tracks, Domain 3 had the lowest aggregated mean at first placement (*M* = 2.96) and second placement (*M* = 3.39). For all cohorts, Domain 4 had the highest aggregated mean for both placements.

Research Question 2 – Candidates' Change Scores from Placement 1 to 2 and Differences Among Tracks

For the overall sample (*n*=97), there were significant mean differences between first placement and second placement (M=-.52, SD=.46), *t*(96) = -11.09, *p*<.001; *d* = 1.08). A paired t-test for each cohort (i.e., YL, SL, and IN) found statistically significant differences between placement one and placement two scores (YL cohort (*n*=18), (M=-.75, SD=.50), *t*(17)=6.32, *p*<.001; *d* = 1.75); SL cohort(*n*=31) (M=-.46, SD=.46), *t*(30)=-5.51, *p*<.01; *d* = 1.08); and IN cohort (*n*=46), (M=-.47, SD=.43), *t*(47)=-7.67, *p*<.001; *d*= 0.96)), illustrating the increase of difference is almost even among the three groups. All measures of Cohen's *d* demonstrate a large effect size (e.g., large = .8 according to Levine & Hullett, 2002).

Research Question 3 – Differences among Teacher Candidates from Different Tracks

Results of the one-way ANOVA, assessing differences across the tracks by criterion, showed 15 of the 40 items, as well as the overall average score (Overall), were statistically different following placement one. Results reported medium to large effect sizes (i.e., medium = .06, large = .14) (see Table 3).

To determine differences among the three tracks after placement one, we ran a Games-Howell post-hoc test in accordance with violating Levene's Homogeneity of Variance statistic (p<.05). The YL track was lower for each of the 16 criteria on which there were significant differences. Typically, both SL and IN tracks scored higher, including the overall placement average score; however, for three of the 16 items, only one of the tracks scored significantly higher (see Table 4). We ran the same analyses on the second placement profile scores and found that there were no significant differences among tracks for any of the 40 items on the performance evaluation tool or for the overall average score across the four domain average ratings.

Discussion and Implications for Practice

Using the PEI tool, this study evaluated teacher candidates' development across four domains at the midpoint and final evaluation points of internship. Findings highlighted statistically significant growth from the first to the second placements across all three tracks. More telling, perhaps, are the trends that emerge from the data analyses as they reveal particular areas of strength and weakness in our teacher candidates' professional development. Each finding holds important implications for teacher education broadly.

First, candidates excelled in dispositional areas such as Criterion 20, *demonstrates courtesy and caring relationships with students*, and Criterion 23, *works towards developing a positive classroom community*, in Domain 4, *Professional Development*. This finding suggests that programmatic efforts to recruit and select applicants with professional dispositions are important selection activities. Additionally, growth in this area elucidates that, during internship, candidates learn about the dispositions of a professional educator.

Results show also that candidates tended to score high in the first half of their internship on practice-based skills that are routine or related to organizational skills and logistics. For example, Criterion 7, *gathers, creates and organizes materials and equipment in advance*, was an area where candidates showed more proficiency earlier in their internship compared to other criteria. Being organized is important to teaching, but it is not necessarily related to pedagogical ability. Another example of a logistical skill is Criterion 1, *uses curriculum guidelines and learning standards during planning to meet the needs of learners*. It is an expectation that teacher candidates use existing pedagogical resources, such as curriculum guidelines and standards, to plan learning activities.

Conversely, teacher candidates scored particularly low in areas related to diversity and culturally responsive teaching (e.g., Criteria 2, 4, 5, and 18), which speaks to the national challenges candidates face when working with diverse student populations (Hollins & Guzman, 2005). These candidates across all program tracks reflect the national demographics of the profession; that is, they tend to be white, middle class women (Zumwalt & Craig, 2008), which contrasts with an increasingly diverse K-12 student population. Given this divide, many candidates struggle to understand and relate to their students. Further, as teacher preparation approaches remain inconsistent and outcome measures are poorly prepared with few investigating the longitudinal effects of these approaches (Cochran-Smith & Fries, 2005), results here show that our candidates struggled to quickly impart the knowledge and tools for effectively meeting diverse learners' needs. As such, these findings reinforce former literature findings but also provide evidence for preparation programs to better incorporate culturally responsive teaching tenets in coursework, field experiences, and formative evaluation measures.

Findings revealed that candidates scored lower in criteria that measured practice-based skills requiring adaptive and responsive teaching and reflective practice. In fact, the mean scores for criteria related to differentiated instruction and higher-order thinking were among the lowest. In Domain 2, Instruction and Management, Criterion 15, *encourages critical thinking and problem solving*, and in Domain 4, Criterion 38, *can develop and explain professional judgments*, proved to be areas of weakness for teacher candidates. These findings reflect some of the challenges associated with learning to teach as they require adaptive expertise (Darling-Hammond & Bransford, 2005) or the ability to make pedagogical judgments about what to do in specific situations (Allen, Matthews, & Parsons, 2013; Parsons, 2012). As a result, this elementary program asserts that candidates not only learn about students and how they engage with content but also require learning through situated practice (Mascarenhas et al., 2010). This finding provides support for calls in teacher education for robust, systematic, course-embedded field experiences.

The PEI as an evaluation tool assists teacher candidates in improving their teaching during their internship experiences. Educators agree that teaching is a complex task that cannot be reduced to simple routines (Hammerness et al., 2005; Kennedy, 1999). In many ways, the criteria listed in the PEI have been used to address important practice-based skills, often referred to as high-leverage practices. High leverage practices, according to Ball, Sleep, Boerst, and Bass (2009), include activities of "teaching that are essential to the work and that are used frequently, ones that have significant power for teachers' effectiveness with pupils" (p. 461). In the PDS model, teacher candidates analyze how expert teachers navigate this complexity of teaching and begin to develop knowledge about when, why, and how aspects of their competency are relevant. This conditional knowledge guides teacher candidates to become more adaptive and responsive in unanticipated situations (Duffy, Miller, Parsons, & Meloth, 2009). Hence, performance evaluation tools have implications for teacher candidates' professional development in becoming high-quality teachers.

Further, the PEI serves as a tool for improvement, reflection, and course building within a teacher education program. Through annual reviews of aggregated data, preparation programs can recognize areas where teacher candidates require additional support and maturation to achieve quality success in the classroom. For instance, analyzing data in Domain 2, *Instruction and Management*, revealed that these candidates were not prepared to teach diverse learners and to differentiate instruction. Moreover, candidates struggled to incorporate higher-order thinking into their lessons and instruction. As a result, this program adjusted coursework in mathematics and science methods courses to introduce problem-based learning and inquiry activities to immerse candidates in these approaches. Additionally, during internship reflection exercises, the program focused on equipping candidates with questioning techniques to elicit student thinking.

As well, Domain 3, *Assessment*, surfaced as another struggle for candidates. Many scored low on the overall aggregated averages in this domain. Further investigation of the data, however, revealed that these candidates had little exposure to areas of assessment measured. To address these candidate needs, the program developed and incorporated an action research component into the second placement internship where candidates participate in inquiry-based research surrounding a relevant need for their assigned classroom.

Performance evaluation tools, as evidenced by our PEI tool here, can bolster teacher education programs' formative and summative evaluation mechanisms and highlight the trajectory of teacher candidates' knowledge, skills, and dispositions over time. As the PEI measure shows consistent results over time, future validation measures (such as an exploratory factor analysis) will examine the statistical soundness of the instrument. Nonetheless, by exploring the efficacy and results of the PEI, the tool currently serves not only to examine the growth and development of our candidates across expected skills and knowledge but also to facilitate programmatic development for purposes of enhancing teacher quality. In the future, performance evaluation instruments, such as the PEI, might be administered earlier and at multiple intervals to better inform teacher educators' understanding of candidates' professional development. In addition, strategies for incorporating evidence into the evaluation process should be considered. Finally, ongoing review and evaluation of the tool with all stakeholders is essential to maintain currency and relevancy.

Conclusion

This paper explored the relationship between how a performance evaluation tool informed our understanding of teacher candidates' development and how results were used to tailor specific program improvements to support candidates' instructional practices. Results documented how one elementary program used a performance evaluation tool grounded in a PDS framework to evaluate teacher candidates' professional development across four domains of practice as well as inform program improvement.

Stemming from results that candidates scored particularly low in areas related to diversity and culturally responsive teaching, the program revised coursework and field experiences to bolster candidate awareness of culturally relevant pedagogical practices (Ladson-Billings, 1995). Specifically, to increase sociocultural consciousness, the program incorporated teacher candidates' reflection and simulated activities centered on challenging their own views, biases, and perceptions of culture, and facilitated discussions about families, including participation in a home visit assignment, all of which informed how these sociocultural attributes influence teaching, student development, and student learning.

Additionally, with candidates scoring lower in the criteria that measured practice-based skills requiring adaptive and responsive teaching and reflective practice, the program focused on instructional practices that would help teacher candidates engage in productive discussions during problem solving in teaching mathematics. The mathematics education faculty incorporated the five practices for orchestrating mathematics discussions by Smith and Stein (2011) as part of the mathematics instruction. These five practices prepared candidates for enhanced opportunities for critical thinking and problem solving. Teacher candidates anticipated students' responses, problem solved with their colleagues prior to teaching a lesson, and reflected on their responses to students' thinking following their lessons.

Beyond program improvement, the PEI provided an opportunity for teacher candidates to self-assess their progress as beginning teachers, becoming autonomous and reflective in identifying their own maturation in practice, while receiving ongoing feedback from clinical faculty and university faculty on their teaching. Through this collaborative evaluation tool, this study introduces a nuanced approach to evaluating candidates and presents an evaluative tool that contributes to teacher education's current directions of research and practice for developing high-quality teachers.

References

About EdTPA. (n.d.). Retrieved June 29, 2015, from http://edtpa.aacte.org/about-edtpa

Allen, M., Matthews, C. E., & Parsons, S. A. (2013). A second-grade teacher's adaptive teaching during an integrated science-literacy unit. *Teaching and Teacher Education, 35*, 114-125. doi:10.1016/j.tate.2013.06.002

Auerbach, C., La Porte, H. H., & Caputo, R. K. (2004). Statistical methods for estimates of interrater reliability. In Roberts, A.R. & Yeager, K.R. *Evidence Based Practice Manual: Research and Outcome Measures in Health and Human Services*, pp 444-448, New York, NY: Oxford University Press.

Ball, D., Sleep, L., Boerst, T., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. *Elementary School Journal, 109*, 458-474. doi:10.1177/0022487109347321

Castle, S., Fox, R. K., & Souder, K. O. (2006). Do professional development schools (PDSs) make a difference? A Comparative study of PDS and non-PDS teacher candidates. *Journal of Teacher Education, 57*, 65-80. doi:10.1177/0022487105284211

Cochran-Smith, M., & Fries, K. (2005). Researching teacher education in changing times: Politics and paradigms. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying teacher education* (pp. 69-110). Mahwah, NJ: Erlbaum.

Cohen, E., Hoz, R., & Kaplan, H. (2013). The practicum in preservice teacher education: A review of empirical studies. *Teaching Education, 24*, 345-380. doi:10.1080/10476210.2012.711815

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart, and Winston.

Damore, S. J., Kapustka, K. M., & McDevitt, P. (2011). The urban professional development school network: Assessing the partnership's impact on initial teacher education. *The Teacher Educator, 46*, 182-207. doi:10.1080/08878730.2011.582929

Darling-Hammond, L., & Bransford, J. (Eds.). (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Journal of Teacher Education, 16*, 523-545. doi:10.1016/S0742-051X(00)00015-9

Duffy, G. G., Miller, S. D., Parsons, S. A., & Meloth, M. (2009). Teachers as metacognitive professionals. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 240-256). Mahwah, NJ: Lawrence Erlbaum.

Hammerness, K., Darling-Hammond, L., Grossman, P., Rust, F., & Shulman, L. (2005). The design of teacher education programs. In L. Darling-Hammond, & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 390-441). San Francisco, CA: Jossey-Bass.

Hollins, E. R. (Ed.) (2015). *Rethinking field experiences in preservice teacher preparation: Meeting new challenges for accountability*. New York, NY: Routledge.

Hollins, E. R., & Guzman, M. (2005). Research on preparing teachers for diverse populations. In M. Cochran-Smith & K. M. Zeichner (Eds.) *Studying teacher education* (pp. 477-549). Mahwah, NJ: Erlbaum.

Holmes Group. (1990). *Tomorrow's schools: Principles for the design of professional development schools*. East Lansing, MI: Holmes Group.

Kennedy, M. (1999). The role of preservice teacher education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 54-85). San Francisco, CA: Jossey Bass.

Ladson-Billing, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into Practice, 34*(3), 159-165. doi:10.1080/00405849509543675

Lampert, M., Franke, M. L., Kazemi, E., Ghousseini, H., Turrou, A. C., Beasley, H., & Crowe, K. (2013). Keeping it complex using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education, 64*, 226–243. doi:10.1177/0022487112473837

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4 (863)*. 1-12. doi:10.3389/fpsyg.2013.00863

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*, 612-625. doi:10.1111/j.1468-2958.2002.tb00828.x

Lortie, D. (1975). *Schoolteacher: A sociological study*. Chicago, IL: University of Chicago.

Margolis, J., & Doring, A. (2013). National assessments for student teachers: Documenting teaching readiness to the tipping point. *Action in Teacher Education, 35*, 272-285. doi:10.1080/01626620.2013.827602

Mascarenhas, A., Parsons, S. A., & Burrowbridge, S. C. (2010). Preparing teachers for high- needs schools: A focus on thoughtfully adaptive teaching. *Bank Street Occasional Papers*, 25, 28-43.

National Association of Professional Development Schools (NAPDS). (2008). *Policy statement on professional development schools*. Retrieved from http://napds.org/

National Council for Accreditation of Teacher Education (NCATE). (2010). *Transforming teacher education through clinical practice: A National strategy to prepare effective teachers. Report of the Blue Ribbon Panel on Clinical Preparation and Partnerships for Improved Student Learning*. Washington, DC: Author.

Neapolitan, J. E. (2011). *Taking stock of professional development schools: What's needed now*. New York, NY: Teachers College Press

Parsons, S. A. (2012). Adaptive teaching in literacy instruction: Case studies of two teachers. *Journal of Literacy Research, 44*, 149-170. doi:10.1177/1086296X12440261

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education, 57*, 22-36. doi:10.1177/0022487105284045

Ridley, D. S., Hurwitz, S., Hackett, M. R. D., & Miller, K. K. (2005). Comparing PDS and campus-based preservice teacher preparation. *Journal of Teacher Education, 56*, 46-56. doi:10.1177/0022487104272098

Rubin, A., & Babbie, D. (2007). *Research methods for social work* (3rd ed.). Belmont, CA: Brooks/Cole.

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

Sandholtz, J. H., & Wasserman, K. (2001). Student and cooperating teachers: Contrasting experiences in teacher preparation programs. *Action in Teacher Education, 23*, 54-65. doi:10.1080/01626620.2001.10463075

Sawchuck, S. (2013, December). Performance-based test for teachers rolls out. *Education Week, 33*(13), pp. 1, 22.

Smith, M., & Stein, M. K. (2011). *Five practices for orchestrating productive mathematics discussions*. Thousand Oaks, CA: Corwin Press.

Uhlenbeck, A. M., Verloop, N., & Beijaard, D. (2002). Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*, 242-272. doi:10.1111/1467-9620.00162

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

Zumwalt, K., & Craig, E. (2008). Who is teaching? Does it matter? In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre, & K. E. Demers (Eds.), *Handbook of research on teacher education* (3rd Ed., pp. 404-423). New York, NY: Routledge.

Table 1

Description of program tracks

Cohort	Semesters of Academic Coursework	Hours of Field Work	Internship Type
Year-Long	6	15-30 (first 4 semesters)	2 semester-long placements *(1 upper and 1 lower grade)*
Semester-Long	7	15-30 (first 6 semesters)	1 semester; 2 8-week placements *(1 upper and 1 lower grade)*
Intensive	5	15-30 (first 3 semesters) 3 days/week (in 4th semester)	1 semester; 2 8-week placements *(1 upper and 1 lower grade)*

Table 2

Lowest mean scores for each PEI domain

Criterion	Mean
Domain One	
Criterion 4	2.98
Criterion 5	2.90
Criterion 6	2.85
Domain Two	
Criterion 12	2.98
Criterion 15	2.99
Criterion 21	2.93
Criterion 22	2.88
Criterion 24	2.95
Domain Three	
Criterion 27	2.88
Criterion 30	2.89
Criterion 25	2.87
Criterion 32	2.77

Table 3

F Values, Significance, and Effect Size (η^2) for Statistically Significant Criterion

Criterion	*F* Value	Sig.	Effect Size (η^2)
2	7.80	.001**	.13
3	6.38	.002**	.11
4	3.45	.04*	.06
5	3.19	.04*	.06
8	3.82	.03*	.07
11	3.39	.04*	.06
13	8.10	.001**	.14
15	5.29	.007**	.09
17	8.94	.000***	.15
25	5.01	.008**	.09
26	4.59	.01*	.08
27	3.97	.02*	.07
29	3.12	.04*	.06
30	3.15	.04*	.06
32	5.46	.006**	.10
Overall	3.33	.04*	.06

Note. Sig. = Significance. * = $p<.05$. ** = $p<.01$. *** = $p<.001$.

Table 4

One-way ANOVA of Differences of Cohorts by Evaluation Criterion on Midterm Evaluation

Dependent Variable	Group	Group	Mean Diff.	Sign.
Criteria	(I)	(J)	(I-J)	
C2: Develops unit and lesson plans to meet the developmental and academic needs of diverse learners	YL	IN	-.61	.001**
		SL	-.66	.002**
C3: Plans a sequence of engaging activities, focused on achievement of the instructional objective(s)	YL	IN	-.61	.001**
		SL	-.58	.005**
C4: Selects learning experiences, technology, and materials to acc. different styles/levels of learning	YL	IN	-.42	.014*
		SL	-.44	.022*
C5: Relates activities to students' culture, interests, knowledge, and experiences.	YL	IN	.16	.057
		SL	-.41	.043*
C8: Plans for using various methods to assess students' learning.	YL	IN	-.47	.017*
		SL	-.46	.044*
C10: Uses a variety of teaching methods, techniques, and strategies.	YL	IN	-.38	.047*
		SL	-.37	.100
C13: Provides opportunities for learners to participate actively and successfully at diff. levels.	YL	IN	-.55	.001**
		SL	-.58	.001**
C14: Provides opportunities for learners to work independently and in cooperative groups.	YL	IN	-.36	.037*
		SL	-.40	.042*
C15: Encourages critical thinking and problem solving.	YL	IN	-.52	.007**
		SL	-.53	.015*
C17: Motivates students through interesting and challenging activities.	YL	IN	-.62	.001**
		SL	-.64	.002**
C18. Communicates high expectations while respective ind. differences and cultural diversity.	YL	IN	-.39	.046*
		SL	-.45	.037*
C25: Uses assessment that matches the objective.	YL	IN	-.43	.008**
		SL	-.45	.012*
C26: Uses assessment to inform future instruction.	YL	IN	-.52	.009**
		SL	-.43	.062
C27: Adapts pacing, methods, and materials using feedback from students.	YL	IN	-.49	.008**
		SL	-.43	.047*
C29: Assesses for understanding and mastery through evaluation of student's work.	YL	IN	-.39	.031*
		SL	-.41	.043*
C32: Gathers, organizes, and analyzes student data to communicate progress to others.	YL	IN	-.53	.005**
		SL	-.53	.012*
FINAL	YL	IN	-.33	.023*
		SL	-.32	.061

Note. YL = Year Long. SL = Semester Long. IN = Intensive. Sig. = Significance. Ind.= Individual. *. = *p* <.05.
**. =*p*<.01.

Appendix

Profile for Evaluation of Intern

Intern: _________________________________ **Spring** _______ **Fall** _______ **Yr** __________
School: _________________________________ **Grade Level**: ____________________________
Evaluator: ______________________________ **UF or CF or Intern**
Recommended Interim Grade: ____________ **or Final Grade:**_________________________

This assessment of the intern's performance is to be completed by the clinical faculty/cooperating teacher, the university facilitator and the intern. The items reflect the important standards and competencies expected of professional educators, and the rating scale reflects their movement toward achieving proficiency over the course of the internship. This form may be used to record the interim AND final ratings.

- A rating of 3 indicates that the Intern has achieved consistent proficiency in a particular skill or disposition. An average of 3 or higher across all areas (Grand Average) represents a passing grade.
- Ratings of 1 or 2 indicate skills that require scaffolding and support on the part of the CF and UF in order for the Intern to develop the appropriate level of expertise. Please include comments that indicate a plan to address these skills and dispositions.
- Ratings of 4 or 5 suggest that the Intern's performance regarding a skill or disposition is exceptional. These ratings should be reserved for documentable excellence. Please include comments that indicate the ways in which the Intern has exceeded expectations.
- The interim or final grade is based on this profile, but may not be numerically correlated.
- Graduate Grading Scale: S=Satisfactory; NC=No Credit; IP=In Progress

Performance Rating Scale:

5	=	Performance is of notable excellence.
4	=	Performance often goes beyond expectations.
3	=	Performance is consistently proficient.
2	=	Performance needs some improvement.
1	=	Performance needs significant improvement.
NR	=	Performance on this item was not rated during this evaluation.

Summary of Scores:

Interim		**Final**	
Preparation & Planning	_______	Preparation & Planning	_______
Instruction and Management	_______	Instruction and Management	_______
Assessment	_______	Assessment	_______
Professional Development	_______	Professional Development	_______

Grand Average (average of scores) ________	Grand Average (average of scores) ________

I. Preparation and Planning	**Interim**	**Final**
1. Uses curriculum guidelines and learning standards during planning to meet the needs of learners.		
2. Develops unit and lesson plans to meet the developmental and academic needs of diverse learners.		
3. Plans a sequence of engaging activities, which are focused on achievement of the instructional objective(s).		
4. Selects learning experiences, technology and materials to accommodate different styles and levels of learning.		
5. Relates activities to students' culture, interests, knowledge, and experiences.		
6. Integrates materials and activities that are sensitive to culture, disabilities and gender.		
7. Gathers, creates and organizes materials and equipment in advance.		
8. Plans for using various methods to assess students' learning.		
9. Collaborates with other teachers and specialists in planning.		

<u>Preparation and Planning</u>

Average Rating (to 2 decimal places) ________

Interim Comments:

Final Comments:

II. Instruction and Management	**Interim**	**Final**
10. Uses a variety of teaching methods, techniques and strategies.		
11. Consistently presents accurate content.		
12. Consistently provides clear instruction		
13. Provides opportunities for learners to participate actively and successfully at different levels.		
14. Provides opportunities for learners to work independently and in cooperative groups.		
15. Encourages critical thinking and problem solving.		
16. Appropriately uses a variety of materials, technology and other media to achieve instructional objectives.		
17. Motivates students through interesting and challenging activities.		
18. Communicates high expectations while respecting individual differences and cultural diversity.		
19. Creates and/or uses established routines to provide an orderly and supportive environment.		
20. Demonstrates courtesy and caring in relationships with students.		
21. Manages time, space and materials to keep students productively involved in learning.		
22. Demonstrates ability to manage 2/+ classroom activities simultaneously, with evidence of attention to each		
23. Works toward developing a positive classroom community.		
24. Handles disruptive or destructive behavior firmly and fairly.		

Instruction and Management

Average Rating (to 2 decimal places) ____________

Interim Comments:

Final Comments:

III. Assessment	**Interim**	**Final**
25. Uses Assessment that matches the objective		
26. Uses assessment to inform future instruction.		
27. Adapts pacing, methods and materials using feedback from students		
28. Assesses for understanding and mastery through observation of students' performance.		
29. Assesses for understanding and mastery through evaluation of students' work.		
30. Keeps records of students' progress and problems.		
31. Communicates with students to inform them of their progress.		
32. Gathers, organizes, and analyzes student data to communicate progress to others.		

Assessment

Average Rating (to 2 decimal places) ______

Interim Comments:

Final Comments:

IV. Personal and Professional Development	Interim	Final
33. Possesses the basic skills and knowledge needed to guide students' learning.		
34. Demonstrates effort to continue learning both content and pedagogy.		
35. Reflects on his/her professional practice.		
36. Welcomes assistance for improvement.		
37. Implements suggestions and recommendations for improvement.		
38. Can develop and explain professional judgments.		
39. Engages in productive relationships with professional colleagues and support staff.		
40. Demonstrates stamina, flexibility and a positive attitude.		

Professional and Personal Development

Average Rating (to 2 decimal places) _______

Interim Comments:

Final Comments: