Student Teachers' Assessments Involving Three Role Groups: Challenges and Possibilities

Ruchi Bhatnagar
Georgia State University, USA

Abstract: Evidence and use of standards have become buzz words in teacher education. In order to satisfy the policy requirements, meet accreditation standards, and respond to the critics of traditional routes of teacher preparation, teacher educators are attempting to balance their program philosophy with state and national standards in designing their assessment systems. Using a mixed methods design, this study examined the use of an assessment instrument by three role groups in the student teaching semester and the purposes these assessment data fulfilled for the student teaching triad, the teacher education program, and the policy makers. The findings of this study highlight the difficulties involved in creating standards for assessment in teacher education such that they inform the practice of teacher education, are valid indicators of student teachers' knowledge, performance, and dispositions, and reflect the effectiveness of teacher education programs.

Introduction

Schools of education have long been attacked by the general public as well as policy makers for their apparent failure in supplying well qualified teachers into the American classrooms (Labaree, 2010). Educators too have critiqued teacher preparation programs for not having a strong knowledge base, for being out-of-date and out-of-touch in preparing teachers for an outcomes-based education system, and lacking empirical evidence of factors that constitute an effective teacher education program (Ravitch, 2013). More recently, politically charged reformers, have called for privatizing education, proposing competition via choice and alternative routes to teacher certification (Ravitch, 2013; Zeichner, Payne, & Brako, 2015).

Thus, education reform initiatives have aimed at the improvement of teacher education and pushed for a standards-based curricula and standardized assessments (Cochran-Smith, et al., 2016). Simultaneously, teacher educators have faced pressure for increased transparency and accountability focused on teacher preparation in a rapidly changing 21st century global economy (Cochran-Smith, et al., 2016). To collect evidence for effectiveness of teacher education programs and the quality of teacher candidates, teacher educators and policy makers across the nation are attempting to integrate standards in their assessment systems (Bastian, Lys & Pan, 2018; Wei, Pecheone, & Wilzac, 2014). In particular, the use of edTPA has gained wide-spread popularity in multiple states around the country, with 18 states already having a policy for edTPA passing to gain initial teacher certification (SCALE 2018). One of the major arguments for the use of edTPA in various states has been that it is a valid and reliable measure of teacher assessment, examines the teacher candidates' ability to plan, instruct and assess in an authentic manner, and was developed by teacher educators, drawing from the knowledge base of teacher education (SCALE 2018; Wei, Pecheone, & Wilzac, 2014). However, critics of edTPA point out that high-stakes use of edTPA results in decrease of faculty autonomy, coopting of teacher education by for-profit private

entities, and takes focus away from context and critical pedagogy (Donovan & Cannon, 2018; Sato, 2014). However, there are many other states and teacher education colleges that are using edTPA without a policy in place or using traditionally used in-house developed assessments for assessing the preparedness of their teacher candidates. This study describes the use of standards called The Domains of Professional Learning to assess the preparedness of teacher candidates at one university in the Midwest, which does not have an edTPA related policy. Specifically, the goal of this study was to understand if the assessments fulfilled the purposes described by the program when three role groups: student teachers, cooperating teachers, and the field instructors (also called the student teaching triad) were involved in using standards in the assessment process.

Focus on Evidence, Assessment, and Accountability

Researchers have long argued that teacher quality matters and is related to children's learning in schools (Hill, Rowan, & Ball, 2005; Wayne & Youngs, 2003; Ziechner, 2007). Policy makers claim that teachers are not adequately prepared to face the challenges of the profession, especially in the first few years of their career, and consequently leave teaching, which reflects poorly on the quality of their preparation (Liston, Whitcomb, & Borko, 2006). According to the Alliance for Excellent Education (2014), the estimated cost of teacher turnover is $2.2 billion annually. Thus, it is argued that teacher education programs can and should play an important role in ensuring that well prepared and high quality teachers enter the K-12 system (Sykes, Bird, & Kennedy, 2010). With the issue of teacher quality being redefined as

effectiveness, teacher education programs are now under increased pressure to generate evidence that their graduates possess the attributes of effectiveness, are more likely to remain in the profession, and contribute to increased achievement of their students in schools (CAEP, 2018; Duncan, 2010; Wallace, 2009).

To address the issue of quality in teachers, policy makers and teacher educators have tried to offer several suggestions. This debate has raised questions regarding the importance of content matter knowledge, the merit of certification, ways of assessing teacher quality, and outcomes of teacher education programs (Darling-Hammond, Hammerness, Grossman, Rust, & Shulman, 2005; Wilson & Youngs, 2005). Teacher educators and policy makers agree that to prepare quality teachers, it is important to require reliable assessments and standardize what are thought to be effective teacher education processes (Cochran-Smith & Boston College Evidence Team, 2009; Cochran-Smith & Fries, 2005; Darling-Hammond, et al., 2005; Wilson & Youngs, 2005). Policy makers have also called for linking teacher quality to student outcomes, which is the ideal measure of teacher effectiveness (CAEP, 2018; Cochran-Smith et al., 2009). As a result, the endeavor of collecting evidence about the preparedness of their graduates and program's outcomes has gained much prominence in the design and functioning of most teacher education institutions over the last ten years (Goldhaber, 2015).

Amidst compelling demands from policy, the scholarship in teacher education has yet to come to a consensus about what we mean by quality, and the way we should assess it in beginning teachers. The demand

for evidence is being responded to in the absence of shared consensus about how quality should be measured and the utility of this information is uncertain (Cochran-Smith, 2006; Wilson & Youngs, 2005; Wineburg, 2006). Most teacher education programs have developed in-house systems of assessments and protocols in an attempt to collect evidence of their student teachers' quality and the effectiveness of their program (Wineburg, 2006). A survey conducted by the American Association of State Colleges and Universities (AASCU, 2005) found that teacher education programs spent a lot of money, time, and resources in collecting assessment data to meet policy requirements, but most of this data was unusable, because of poor reliability and validity. Furthermore, data requirements and definitions of quality used by state, federal, and national accreditation agencies are often different, which makes evidence collection more complex (Wineburg, 2006; Ziechner, 2007).

Examples of Teacher Performance Assessments:

This discussion of assessment systems will be incomplete without a review of the available options within the field of teacher education, some of which offer a strong case for validity and reliability. California universities developed the first standardized TPAs which led to the development of nationally available teacher performance assessments. The PACT Consortium of 2001, comprised of 12 institutions, developed the PACT which became the precursor to nationally available and other state teacher performance assessments (Pecheone & Chung 2006). It is a subject-specific pedagogy portfolio covering 17 credential areas and focused on a

Teaching Event (TE) (Pecheone & Chung, 2006). The portfolio is comprised of five tasks: (1) Planning, (2) Instructing, (3) Assessment of student learning, (4) Reflection, and (5) Academic language. Evaluators use three rubrics per task for scoring. Data from the Teaching Event is used to meet standards for the high-stakes credentialing in California.

The edTPA is a subject-specific, teacher performance-based assessment created by educators and owned by Stanford University (SCALE, 2017). SCALE formally launched the edTPA in 2013, following two years of field testing with 12,000 candidates across 450 institutions of higher education and 29 states (SCALE, 2017). edTPA's structural design incorporates 80% general pedagogy (i.e. planning, teaching, and assessing) and 20% subject-specific pedagogy constructs across 27 content areas.

Content Development Team in conjunction with Educational Testing Services developed the PPAT which was approved in 2015, as a second nationally available TPA (Educational Testing Services, 2016). It has been explored by educators in 17 states and is designed after the InTASC Model Core Teaching Standards. It requires that the teacher candidate create a standards-based portfolio with embedded content rather than one that emphasizes subject-specific pedagogy like the PACT and the edTPA. It includes four tasks with the first one scored locally and the subsequent tasks scored externally: (1) Knowledge of students and their learning environment, (2) Assessment and data collection to measure and inform student learning, (3) Designing instruction for student learning, and (4) Implementing and analyzing instruction to promote student learning (Educational Testing Services, 2016).

Thus, considering advancements in the realm of performance assessment rubrics and portfolios, it is often preferable for states to utilize one of these already developed assessments and satisfy the CAEP requirements for validity and reliability. However, this choice is not a simple one since these assessments are often critiqued for being removed from the context, not involving feedback from the role groups vested in student teaching, and overwhelm the programs by their time, skill, and cost intensive nature (Dover & Schultz, 2016; Lachuk & Koeller, 2015; Sato, 2014). Teacher education is facing a narrowing of the curriculum due to standardized assessments, leading to a culture of teaching to the test (Gorlewski & Gorlewski, 2015).

Role of the Student Teaching Triad in Assessments:

The process of collecting evidence gets further complicated by the debate on the stakeholders that are best suited to evaluate the teacher candidates' proficiencies. Within the edTPA for instance, an objective outsider, is the evaluator (SCALE, 2018). However, critics have objected that if the assessments are not embedded within the context, and are done without a deep understanding of the school, pupils, and also the teacher candidate, they will not provide the candidates a deep and meaningful insight on their performance and the ways to grow as a teacher (Margolis & Doring, 2013). Since a critical aspect of teacher preparation is linked to field experiences and student teaching, it is argued that cooperating teachers be included in the assessment process (Tilemma, 2009; Valencia, Martin, Place, & Grossman, 2009). Research has shown that cooperating teachers play a critical role in enabling the teacher candidates bridge the gap between theory and practice, shape their perspectives on teaching and their role as future teachers, and equip them with important tools of the profession (Anderson, 2009). The cooperating teachers however, are far removed from the ivory tower of higher education and power and status issues may impede these P-12 partners' full involvement in the assessment of teacher candidates (Whitney, Golez, Nagel, & Nieto, 2002).

The other critical role in the student teaching semester is played by the university supervisors who are the contact persons for the implementation of the student teaching process, they are the providers of key information about the progress of the student teachers to the university. Ironically, research in teacher education has focused little attention on the work of university supervisors even though they are seen as participants in the student-teaching experience, an experience that, over time, has consistently been cited by classroom teachers as the single most influential component of their teacher education programs (Wilson, Floden, & Ferrini-Mundy, 2002). The work of supervision, often comes with vague rationale and expectations and tends to receive little attention in terms of orientation or professional development on program goals and pedagogical practices (Beck & Kosnik, 2002).

Teacher educators have also called for including self-assessments by the teacher candidates to allow them to incorporate the language of standards in their professional vocabulary and to get an opportunity to reflect on their practices as beginning teachers (Cochran-Smith, et al., 2009; Grossman & McDonald, 2008). However, to have teacher candidates rate themselves, especially if it counts for a grade on a

course, is challenging; since they are vested in the successful completion of the program and are developmentally not at a point where they can make sophisticated self-critical reflections on their practice, especially when it tarnishes their grades (Anderson, 2009; Margolis & Doring, 2013). Thus, when three role groups from diverse backgrounds come together to evaluate the student teachers, it can be expected that their positioning in the triad will add valuable insight to the student teachers' progress over the semester (Tilemma, 2009; Valencia, et al., 2009). At the same time, assessment data from role groups with different orientations to the purposes of the assessment system can possibly make it difficult to document evidence for teacher candidate quality and accomplish the purposes the programs expect their assessment systems to fulfill (Author, 2011).

The complexity increases further when the same standards are to be used across the college, for elementary as well as secondary student teachers. It is challenging to create assessment instruments across grade and subject areas when utilizing the same set of standards and to ensure that assessments will generate meaningful data for the student teaching triad as well as the program faculty (Valencia et al., 2009). Researchers have argued that the full potential of learning during the student teaching semester is not achieved because often there are missed opportunities during the semester to have substantial discussions that are based on assessment data to enable the teacher candidate to become a motivated and a life-long learner (Cochran-Smith, et al., 2009; He & Levin, 2008; Tillema, 2009).

The program in this study drew from research on teacher education assessments and made attempts to maximize the value of assessments by using standards in assessment, gave voice to the three stakeholders in the student teaching semester, and expected that they would use the domains assessments at three points during the semester, followed by a conference guided by the assessments on the domains.

Purpose of the Study

Three aspects of teacher education programs that are most likely related to teacher effectiveness in schools are: content knowledge, field experience, and quality of teacher candidates (National Research Council, 2010). This study includes an investigation of all of these aspects related to teacher effectiveness by examining the assessment process in the student teaching semester used for evaluating the quality of the student teachers both at the elementary and secondary levels at one university. In this university, content knowledge in addition to professional skills and dispositions were used for defining quality in student teachers, and were called the Domains of Professional Learning (refer Table 1). The program assessed the quality of its student teachers by inviting the field instructors, cooperating teachers, and the student teachers to provide their assessments using the domains of learning on a four-point rubric. This rubric was adapted from Danielson's (1996) measures of teacher quality. While the Danielson framework had 4 domains, further dived into 22 components, the adaptation of this rubric at our university used 5 domains, with a total of 14 components. The decision to adapt the Danielson rubric was guided by the need to keep the rubric components limited and user friendly, while capturing the elements of effective teaching deemed important by the faculty. Each domain was given a score on a 4-point scale, with 4

being the highest score. The domain characteristics were to be taken into account while scoring the domain on a 4-point scale.

Table 1

Five Domains of Professional Learning

Domain name	Domain Characteristics
Domain 1: Planning, Assessing, Evaluating	☐ Choose and articulate worthwhile purposes of lesson plans ☐ Assessing prior knowledge of students ☐ Plan and carry out assessment and evaluation
Domain 2: Knowing and Representing Subject Matters	☐ Demonstrate knowledge of key concepts and foundational methods ☐ Adapt or design curricula and employ instructional strategies ☐ Attend to concerns and ways of knowing that span disciplines.
Domain 3: Knowing, Motivating and Engaging Students	☐ Attend to individual students' interests, strengths, prior knowledge and skills ☐ Attend to the diversity of culture and experience that students bring into the classroom
Domain 4: Building Classroom Community	☐ Promote a climate in which learning is valued and on-going, in which all students are able to share in and contribute to social and intellectual life ☐ Uphold fair and equitable standards for conduct that encourage responsibility
Domain 5: Becoming a Member of a Profession	☐ Act ethically and with integrity ☐ Demonstrate respect for families and cultures ☐ Be reflective, seek professional development opportunities. ☐ Be open to constructive feedback from others

The domains specifically focused on student teachers' ability to plan, assess, evaluate, demonstrate knowledge of subject matter, ability to motivate students, build a classroom community, and become a member of the teaching profession. The assessments and discussions of the triads around the domains were expected to keep their work focused around these important facets of learning to teach during the semester. The function of these assessments was primarily to provide formative feedback to the student teachers. At the end of the student teaching semester, these assessments had a summative purpose of recommending the student teachers for certification. Since the program invested time and resources in collecting assessment data, it was important to find out how the assessments provided by the three role groups

addressed the program's purposes of assessment. Specifically, the following questions were addressed in the study:

(1) How did the ratings and feedback on the domains compare across;
   • Role groups?
   • Time point (baseline, mid-point, and end of semester)?
   • Type of domain?
   • Cohort (elementary and secondary)?

(2) What purposes did these assessments fulfill for the student teaching triad? How did their perspectives compare with the purposes of assessment laid out by the teacher education program?

(3) What factors affected the use of the domains as intended by the teacher education program?

Methods

This study was conducted in a teacher education program situated in a large research university in the Midwest. The domains assessment system was used for both elementary and secondary candidates.

Participants

The participants in this study included members of 51 elementary triads (including 51 student teachers, 51 cooperating teachers, and 26 field instructors) and 28 secondary triads (including 28 student teachers, 28 cooperating teachers, and 28 field instructors) involved in student teaching in one semester. The ratings and the written feedback provided by this set of participants on the domains rubric were used for comparing their assessments over time, and across role groups.

Six triads (three elementary and three secondary) were randomly chosen from the participants mentioned above, for collecting the observation and interview data. These 6 triads were observed while they used the domains of assessment rubric in their three-way conferences. The 3-way conferences typically lasted 30 minutes. After observations, each triad member was interviewed individually to understand their perception of the assessment rubric and the purposes that the assessment fulfilled for them. Each interview lasted 30-40 minutes, which sought to probe observation notes during their 3-way conferences and generally, their feedback about the rubric and its components, and the purposes it specifically served for them as a stakeholder. The undergraduate elementary and secondary teacher education program coordinators were interviewed to get the program's perspectives on the design and use of the assessment. All interviews and observations were audiotaped and transcribed, and transcripts of interviews were sent to participants to ensure consistency and provide them an opportunity to approve.

Measures

The triads used the domains rubric to assess the student teachers at the baseline, midterm, and end of student teaching. The quantitative data in this study was drawn from the ratings provided by the student teaching triad on the five domains at baseline, midterm, and end of student teaching, using a four-point scale. The qualitative in this study included: write-ups provided by triad members on the five domains, observations and interview data collected

from six triads, and interviews of the elementary and secondary program coordinators.

Procedures

The data were analyzed in this study using a mixed methods design (Creswell, Plano Clark, Gutmann, & Hanson, 2003). The design used the triangulation framework (Denzin, 1978), including triangulation of data sources (ratings, assessment narratives, interviews, and observations), and triangulation of assessments provided by the three role groups. The quantitative analyses included Multivariate Analysis of Variance (MANOVA) to analyze differences in ratings across the three within-group variables: role group, type of domain, and time of assessment, using cohorts (elementary and secondary) as the between group variable. The role group variable had three levels (student teachers, cooperating teachers, and field instructor), type of domain had five levels (domains 1 through 5), and time of assessment had three levels (baseline, midpoint, and endpoint). The MANOVA test helped to understand how differences in role groups, type of domain, and time point affected the outcome variable- ratings for the elementary and secondary cohorts (Tabachnick & Fidell, 1996).

The qualitative analyses examined the use and understanding of the rubric. Although, assessments from three role groups added richness to the data, inclusion of diverse perspectives also made the data hard to interpret (Author, 2008; Tilemma, 2009). At the same time, examining inconsistencies that appeared in the data obtained from multiple sources was critical to draw plausible explanations for these differences (Mathison, 1988). The interviews conducted with the members of the triad observed during the post-assessment conferences included questions to probe triads' understanding of the language of the domains, their perceptions about how these were useful in informing their practice during the student teaching semester, and the aspects that needed improvement to integrate these standards with their practice.

A content analysis of the assessment narratives, interviews, and observations of student teaching triads was undertaken to investigate the thematic content of the transcripts (Cohen & Manion, 2003; Miles & Huberman, 1994). Member checking was employed after initial analysis of individual interview transcripts and observations. A coding scheme was established after the first level of analysis. To determine the inter-rater reliability for the coding procedure, a team of three researchers concurrently analyzed the assessment narratives from the files of five student teachers, two observation narratives, and three interviews. The researchers met on an ongoing basis to compare responses and further revise and resolve differences in coding until a coding scheme with a high inter-rater reliability was finalized.

Results

MANOVA test showed that for both cohorts, there existed a significant interaction between role group, time of assessment, and type of domain (refer Table 2) indicating ratings given by the three role groups on the different domains differed significantly across the three time intervals (p<.001). Post-hoc analyses using t-tests were conducted to analyze the mean differences in ratings for the different categories within the independent variables. Post-hoc analysis showed the assessments were used in a similar manner by both elementary and secondary cohorts. Overall, ratings

given across the three time intervals increased significantly (p<.001) from one point of assessment to the next (refer Table 3), ratings for the domains 4 and 5 were significantly (p<.05) higher than the rest of the domains (refer Table 4), and the cooperating teachers gave significantly (p<.05) higher ratings than the student teachers and the field instructors (refer Table 5).

Table 2

MANOVA test: Between Group Analysis for Elementary and Secondary Cohorts

Source	DV	df	F	p	η
Corrected Model	Elementary	44	124.11	0.00	0.82
	Secondary	44	161.19	0.00	0.85
Intercept	Elementary	1	185100.69	0.00	0.99
	Secondary	1	187496.99	0.00	0.99
Role	Elementary	2	34.01	0.00	0.05
	Secondary	2	31.01	0.00	0.05
Time	Elementary	2	1939.76	0.00	0.76
	Secondary	2	2308.87	0.00	0.79
Domain	Elementary	4	281.07	0.00	0.48
	Secondary	4	510.35	0.00	0.63
Role * Time	Elementary	4	2.64	0.03	0.01
	Secondary	4	0.52	0.72	0.00
Role * Domain	Elementary	8	7.68	0.00	0.05
	Secondary	8	5.34	0.00	0.03
Time * Domain	Elementary	8	33.10	0.00	0.18
	Secondary	8	33.77	0.00	0.18
Role * Time * Domain	Elementary	16	3.28	0.00	0.04
	Secondary	16	3.52	0.00	0.04

Table 3

Mean Difference Test: Ratings across Three Time Periods

Time Period	Cohort	Mean Rating	S.D	p
Baseline	Elementary	2.70	0.39	0.00
	Secondary	2.58	0.44	0.00
Midterm	Elementary	3.27	0.25	0.00
	Secondary	3.24	0.37	0.00
Endpoint	Elementary	3.87	0.22	0.00
	Secondary	3.81	0.31	0.00

Table 4

Mean Difference Test: Ratings across Five Domains

Time Period	Cohort	Mean Rating	S.D	p
Domain 1	Elementary	2.94	0.26	0.05
	Secondary	2.73	0.31	0.05
Domain2	Elementary	3.09	0.34	0.05
	Secondary	2.95	0.27	0.05
Domain 3	Elementary	3.23	0.30	0.05
	Secondary	3.23	0.23	0.05
Domain 4	Elementary	3.55	0.29	0.05
	Secondary	3.52	0.13	0.05
Domain 5	Elementary	3.61	0.32	0.05
	Secondary	3.61	0.20	0.05

Table 5

Mean Difference Test: Ratings Provided by the Members of the Triad

Role group	Cohort	Mean rating	S.D	p
Student Teachers	Elementary	3.30	0.14	0.05
	Secondary	3.30	0.23	0.05
Cooperating Teachers	Elementary	3.34	0.11	0.05
	Secondary	3.19	0.26	0.05
Field Instructors	Elementary	3.19	0.20	0.05
	Secondary	3.16	0.25	0.05

As shown by effect sizes as well as comparisons across cohorts, the time of assessment seemed to play the most important role in determining ratings on the domains. Domains also differed significantly in terms of the ratings they received and domain differences determined 50% variance in elementary and 60% in secondary triads. Although role group differences were significant, they explained only 5% variance in the ratings.

The narratives provided by the members of the triad in student teachers' assessment files also reflected patterns observed in the ratings. Although the variability in the narratives on the domains was most visible during the baseline assessment, the narratives did clarify why the triads tended to give lower ratings on some domains and higher ratings on other domains. On domain 1(planning, assessing, and evaluating) all triads gave low ratings during the baseline and also expressed challenges in performance on this domain, "I do not feel strong in this domain. With my cooperating teacher's help I

hope to develop purposeful lesson plans" (secondary student teacher). Similarly, on domain 2 (knowing and representing subject matter) student teachers expressed concerns such as, "I find it hard to make effective use of teaching materials and making interdisciplinary connections".

The domains 3 (knowing, engaging and motivating students), 4 (creating a classroom community, and 5 (becoming a member of the profession) were rated higher than domains 1 and 2 by both cohorts. The narratives on these domains included a frequent reference to the personality of the student teachers. For example in domain 3, one elementary cooperating teacher wrote, "Her personality and her enthusiasm are catchy". In domain 4 a field instructor wrote, "She loves teaching, works hard to create a positive environment in her classroom" (Elementary). Domain 5 a student teacher wrote, "I enjoy interacting with my colleagues. I am comfortably slipping into the shoes of a teacher" (Secondary). The narratives at the midpoint and endpoint from all role groups were positive and full of praise for the student teachers, and tended to be general rather than highlighting any aspects of student teachers' weaknesses.

The observations and interviews of triad members provided further insight into the interpretation of the language of domains by the triad and the purposes the assessments served for them. The qualitative findings were organized around broad themes: purposes that the domains fulfilled or did not fulfill and factors that affected the usefulness of the domains.

Purposes Fulfilled/Unfulfilled by the Domains

The need to show growth. From the perspective of the triad, a major purpose served by the domains was that they used the assessments to reflect growth in the student teachers' competency during different times in the semester. The triads used both the ratings and narratives to show growth. In all of the three-way meetings observed, one element was common, talk around how much growth had happened since the previous assessment interval.

It is encouraging for student teachers to see growth on ratings and a higher rating than the previous one shows it better than anything else. But for me, the narratives are just as important to show growth. You can say much more in your write-up, and explain why you gave the rating you gave. (Sandy, field instructor)

Providing formative feedback to the student teachers. Another purpose that some of the field instructors identified was that the domains provided the three role groups a medium to provide formative feedback to the student teachers and brought structure to their discussions during the three way meetings. For example, a field instructor said:

There is so much going on during student teaching, that there is seldom an opportunity for the student teacher to sit down and have a formal feedback session with the cooperating teacher and the field instructor; of course other than the five-minute casual talks on "how did it go?" But in order to form specific goals and time lines, we created a parallel

system and we used it in addition to the domains. (Robin, field instructor)

Like Robin, other field instructors interviewed also felt that the feedback on the rubric tended to be too general to be useful. Most cooperating teachers interviewed also did not like this system of giving feedback. One of them said in her interview that she gave feedback to her student teachers on regular basis, but not along the domains, "feedback is important, but it also has to be contextualized, and related to what [my student teacher] is doing on a daily basis. I can't do that on the domains. I fill out the paper work when it is required, other than that I give feedback as and when needed" (Sandy, cooperating teacher).

Most student teachers interviewed expressed that the importance of the assessments became apparent to them late in the student teaching semester as the following quote suggests, "In the baseline I had no idea why I was filling out the domains and it was not useful, but after the midterm I got the idea about how everything I was doing related to the domains" (Emily, student teacher).

The program's perspective on the purposes of the assessment system is reflected below:

It is a documentation of all the work done by the triad and gives them a system to refer to when they meet. But this is still 'soft assessment', it is not objective, it requires interpretation, and is based on the judgment of each person on how to use it. Language of domains is a starting point, framing of what it would mean to assess knowledge of a beginning teacher.

(Paula, coordinator, elementary teacher education program)

As Paula said, the domains were "soft" assessments thus, it is not surprising that each role group differed in their interpretation of the purposes and use of this document. This issue was compounded by the triads' lack of awareness of the utility of the domains and connection to their day-to-day practice. In their narratives, the importance of showing growth on the assessments from baseline to the end of term took precedence over all other purposes for all three role groups. It is possible that the act of providing a rating at each assessment interval made the assessments appear high-stakes, which added to the discrepancy between the intended and perceived purposes of the assessment.

Factors Affecting the Usefulness of the Domains

Limiting nature of the four-point scale. Showing growth was an important purpose for the triads however; the four-point scale was seen by many as limiting and created difficulty in reflecting growth, as the following two quotes suggest: "It is a small scale. You can't go too far off from where you started. [Another teacher education program] uses a 10-point scale, and had enough categories for me to place the student teacher in" (Ross, cooperating teacher). "If a student teacher comes into my class I expect them to be at least a 2 in terms of their skills. And I am a veteran teacher and I am not a 4, so basically I am only left with two choices, 2 and 3; it does not provide a fair picture of how much growth has taken place" (Darlene, cooperating teacher). Although 60% of the interviewees perceived the rating scale to be

limiting, the MANOVAs showed that their ratings were actually able to capture growth over time. The ratings given by the members of the triad on different domains were significantly different over the three assessment intervals.

Ambiguities in the language of domains. Ambiguities in definition of the domains posed a problem for the triads and they voiced concerns in their interviews. The most common issue with the language used on the rubric was that of "time management" and how it was described on the rubric in domain 3. All interviewees pointed out that the definition was vague and some even said that it was incorrect, "Time management is obscure. I think most of us see it as pacing the lesson, incorporating wait time, time for practice and review, and wrapping up. On the rubric it says very strange things like, "providing clear expectations", and cooperating teachers find it hard to understand" (Casey, field instructor). The assessment narratives indicated that the triads had used time management in the way Casey had described it. The triads thus modified aspects of the domains to mean something that they personally identified with.

On another note, all role groups voiced the concern that classroom management was a big issue for student teachers and initially, most of them focused on classroom management. However, they felt that there was very little on classroom management mentioned on the rubric, "[classroom management] comes up in very indirect ways on the rubric. It should be its own separate entity" (Charlotte, field instructor). 'Making interdisciplinary connections' in domain 2 was another area that the triads found

difficult to assess, as the following field instructor suggests, "I think a better way of putting this would be to say "connects to the real world", or "connects to children's lives". That's at least how I used it" (Casey, field instructor). Thus, the need for a clearer emphasis on the elements that were deemed important in learning to teach by the triads was another feedback for the program.

Descriptors used to define ratings on the rubric. A related issue in the design of the rubric was the way the scale defined the ratings 1 through 4. For instance, the top rating in the scale described expectations for student teachers' as: "maximizes, regularly, mostly, always". The student teaching triad felt that this description of the top category appeared to be unrealistic for student teachers, "now I have been teaching for some 30 plus years, I might not even be top-notch on all domains" (Darlene, cooperating teacher). Darlene's comment indicated that she saw the top rating to be perfect in every domain, which was an unrealistic expectation for student teachers. She pointed out that rating students on a scale that moved from "none- always" was impractical and even undesirable, "I wouldn't call doing all the things all the time good teaching!" Some also expressed that the final rating appeared to be a description of an expert in the field rather than a beginning teacher. Despite their dissonance with the description of the scale, they all still gave the highest rating on all domains at the end of the semester.

Role of the field instructor and cooperating teacher. The field instructors played a vital role in establishing the use of the domains within the triad as one noted, "I try my best to organize my feedback

around the domains, and when any issues arise, I often help them make connections to the domains and point out how it relates to such and such domain" (Charlotte). But in triads where the field instructors were not as explicit about the relevance of the domains, the use of the domains remained marginal. Among the six post-assessment conferences observed, two conferences did not include a reference to the domains rubric while the triads discussed their assessments.

The role of the cooperating teachers was a critical one in determining whether the language of the domains became integrated with everyday practice during student teaching. For example, during an observation, one cooperating teacher said, "it is like signing a mortgage! So much paper work! It is hard for me to keep track of all the items to be completed" (Joanne). Joanne was the cooperating teacher who had also mentioned that she did not use the domains to provide feedback on a regular basis. Thus, in the cases where the cooperating teachers did not find domains helpful as a framework, the domains were not utilized by the student teachers to organize their practice around.

Lack of prior exposure/orientation. Of the total sample of assessment files analyzed, 28% of the field instructors, 80% of the cooperating teachers, and 31% of student teachers wrote their assessment narratives without using the language of the domains. A major reason behind the irregular use of the language of domains in the assessment was inadequate training of the triad members to use the assessment system.

The domains seemed very foreign to the student teachers. They were provided a sheet in the practicum but this did not become significant. I [as a field instructor] was unclear on how to use it too. It was just one of the papers in a fat binder initially. I felt confused about using these categories the scale on the rubric, and my concern was that [cooperating teacher] and [student teachers] would also be confused. (Robin, field instructor)

Most field instructors felt unprepared and lacked confidence in the beginning of the semester, and said that they sometimes could not clarify the domains to the cooperating teacher and the student teachers. Student teachers' responses were also very similar to the field instructors', and four out of six student teachers interviewed said they had never seen the domains prior to student teaching.

Additionally, there was confusion among the triad members about the real purpose of the domains, as intended by the teacher education program. Most cooperating teachers viewed this assessment as high-stakes and going on the student teachers' permanent records. Therefore, their feedback tended to be positive and not very insightful. As one field instructor said, "I find the cooperating teachers are more likely to say NA [not applicable] than give a low rating or saying anything remotely negative" (Casey, field instructor).

In summary, a combination of factors such as: ambiguities in the design and language of the rubric, the absence of a shared understanding of the purpose of the domains, and limited experience in using the domains prior to student teaching, affected the extent

to which the triad members could use the assessments to inform their practices.

Discussion

Current conversations around assessments in pre-service education list multiple expectations for the assessments to fulfill in order to be beneficial for informing the practice of teacher education and policy. Teacher education programs are under pressure to provide reliable and valid appraisals of their teacher candidates' quality (CAEP, 2013; Cochran-Smith, 2006; Wilson & Youngs, 2005). To be informative for the design of teacher preparation, preservice assessments should ideally be based on program goals, be context specific, while reflecting state and national standards (Cochran-Smith, et al., 2009). Assessments should also provide formative and summative feedback to the student teachers, and be valid and reliable indicators of their quality (Cochran-Smith, et al., 2009; Darling-Hammond, 2006). Researchers have pointed out that it is the responsibility of teacher education programs to acquaint their student teachers with the standards that are used for assessing beginning teachers (Darling-Hammond, et al., 2005).

This study investigated the use of the domains assessment system to understand the purposes it served for the student teaching triad. The program had also hoped that these assessments would generate accountability related evidence in response to teacher education policies. However, the results suggested that the assessments on the domains were non-specific and failed to offer substantive feedback to the student teachers. Little interpretation about student teachers' quality could be made from these assessments because they were understood and used

inconsistently by the three role groups. There was a lack of clarity about the real purposes of the domains among all role groups. Among the triad, the ratings given by the cooperating teachers were significantly higher than the field instructors and student. The emphasis of ratings and narratives provided on the rubric shifted to making the student teacher appear "growing well" over the semester. It was not surprising therefore, that the only purpose that was clearly fulfilled by the domains was that of showing growth in student teachers' knowledge and skills over the semester.

Ratings on all domains increased significantly from the baseline to the endpoint. The MANOVAs showed that the triads in both cohorts interpreted the domains 1 and 2 as being more difficult and as a result, domains 1 and 2 received the lowest ratings from all role groups during the three assessments. However, from the program's perspective, all domains were equally important in student teachers' preparation and they had not intended for any one domain to be easier than another.

The program had developed this instrument based on its vision of quality in student teachers, adapting the Danielson (1996) framework. However, the users of the assessment system did not always agree with the program's conceptions of quality. The cooperating teachers were the group that expressed the most disagreement with the way the domains were framed. For them, the use of the domains remained peripheral in informing the mentoring process. Another complication that arose in the use of the domains was that all role groups had different training levels which made the use of the domains a mere formality in the first half of the student teaching semester. The large

gap between the intended and interpreted meaning of these standards was an important feedback for the program. Moreover, the program coordinators mentioned that this was a "soft" assessment and did not provide a definite set of expectations for the assessors. Most tools of assessment in teacher education that are developed in-house, as was true for the domains assessment rubric, are medium to high inference, making it difficult to attain high reliability in assessment data when judgments are sought from a wide range of perspectives such as: the student teachers, cooperating teachers and field instructors (Author, 2008; Darling-Hammond, 2006). Despite critiques, edTPA does exhibit the features of an assessment that can be used as a benchmark for programs across the country. edTPA's use of content-specific handbooks and detailed rubrics that are content specific, but sharing a parallel structure give an opportunity to the programs to see the aspects within and across contents that are strong and the ones needing improvement. edTPA certainly is being recognized as a marked improvement over the home-grown assessment systems traditionally used in teacher education programs, such as the one shared in this study.

Conclusion

Teacher education assessments, whether created at the program level or externally, like the edTPA or the PPAT, have their respective trade-offs, specifically about familiarity with the content, purpose and use of the assessment when the student teaching triads are involved in assessments. The cooperating teachers are perhaps the most removed from this process and are usually the ones with the least amount of say in the evaluation process. Like this study showed, usually well-meaning, the triad concurred with the

program requirements and filled out the paper work, however, the extent to which this really went into informing the student teaching process and enhancing teacher quality is questionable. The involvement of the 3 role groups, although ideal, is not without challenges, especially when a grade at the end of the semester and teacher candidate's licensure are at stake.

What is even more complicated is to establish a predictive validity of the teacher education assessments for future effectiveness of the teacher candidates in P-12 classsrooms. Henry, Thompson, Patriarca, Luterbach, Lys, and Covington (2013) analyzed their own program assessments to examine teacher candidates' progress/performance, based on five indicators, on program assessments to see if they predicted student learning as beginning teachers. They found that their program assessments (i.e., student teaching evaluations, summative portfolios, and dispositions) did not measure multiple constructs, as intended to inform the teacher education program, but instead provided a global rating. Furthermore, none of the instruments produced measures that predicted candidates' effectiveness on student achievement in reading or math. As the ratings on the domains also indicated how a global score from the student teaching triad on the rubric reflected very little about the teacher candidate's actual preparedness. The focus remained on showing growth on the rubric and giving a higher rating than the previous assessment point.

Thus in conclusion, although teacher educators recognize that strong assessments are needed to understand teacher candidates' preparedness, gauge

program quality, and satisfy the accountability and accreditation requirements, the purposes the assessment system is expected to fulfill are numerous and, different states and programs are opting for different solutions to the assessment problem.

References

Alliance for Excellent Education. (2014). *On path to equity: Improving effectiveness of beginning teachers*. Retrieved on May 20, 2018, from http://www.all4ed.org/publications.

American Association of State Colleges and Universities. (2005). *Democracy and civic engagement*: A guide for higher education. Washington, DC: Author.

Anderson, D. (2009). The impact of cooperating teachers on the teaching perspectives of student teachers. *The International Journal of Learning, 16,* 120-133.

Bastian, K. C., Lys, D., & Pan, Yi. (2018). A framework for improvement: Analyzing performance assessment scores for evidence-based teacher preparation program reforms. *Journal of Teacher Education, 69*, 1-15.

Beck, C., & Kosnik, C. (2002). Professors and the practicum: Involvement of university faculty in preservice practicum supervision. *Journal of Teacher Education, 53*, 6-19.

Berry, B. (2005). The future of teacher education. *Journal of Teacher Education*. *56*, 272-278.

Bhatnagar, R. (2008). *Assessing student teachers' preparedness: Use of domains in the undergraduate program (unpublished doctoral dissertation).* University of Michigan, Ann Arbor.

Bhatnagar, R. (2011). *Use of standards in assessment of student teachers: Purposes fulfilled, untapped potential*s.

Paper presented at the annual conference of American Educational Research Association, New Orleans, La.

Borko, H., Liston, D., & Whitcomb, J. (2007). Genres of empirical research in teacher education. *Journal of Teacher Education*, *58*, 3-11.

Cochran-Smith, M. (2006). Taking stock in 2006: Evidence, evidence everywhere. *Journal of Teacher Education, 57*, 6-12.

Cochran-Smith, M., & Boston College Evidence Team. (2009). "Re-Culturing" teacher education: Inquiry, evidence, and action. *Journal of Teacher Education*, *60*, 458-468.

Cochran-Smith, M., & Fries, M. K. (2005). The AERA Panel on Teacher Education research: Context and Goals. In M. Cochran-Smith & K. Zeichner, eds. *Report of the American Educational Research Association panel on research on teacher education*. Washington, DC: American Educational Research Association.

Cochran-Smith, M., Villegas, A. M., Abrams, L. W., Chavez-Moreno, L. C., & Mills, T. (2016). Research on teacher preparation: Charting the landscape of a sprawling field. In D. H. Gitomer & C. A. Bell (Eds). *Handbook of research on teaching* (5th ed., pp. 439-547). Washington, D. C.: American Educational Research Association.

Cohen, L., & Manion, L. (2003). *Research methods in education*. London. UK: Routledge.

Council for the Accreditation of Educator Preparation (CAEP) (2018). Standard 5: Provider quality, continuous improvement, and capacity. Retrieved May 9, 2018 from http://www.ncate.org/standards/standard-5

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advance mixed methods research designs. In A.Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209-240). Thousand Oaks, CA: Sage.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education*, *57*, 120-138.

Darling-Hammond, L., Hammerness, K., 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* (pp. 390-441). San Francisco: John Wiley & Sons, Inc.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2nd ed). New York: McGraw Hill.

Donovan, M. K., & Cannon, S. O. (2018). The university supervisor, edTPA, and the new making of the teacher. *Educational Policy Analysis Archives, 26 (28)*, 1-30.

Dover, A. G., & Schultz, B. D. (2016). Troubling the edTPA: Illusions of objectivity and rigor. *The Educational Forum, 80*, 95-106

Duncan, A. (2010). Teacher preparation: Reforming the uncertain profession. *Education Digest: Essential Readings Condensed for Quick Review*, *75*(5), 13-22. Retrieved November, 2010 from ERIC database.

Educational Testing Services. (2016). ETS: A rich history of educator performance assessment Innovation. Author. Retrieved from https://www.ets.org/s/ppa/pdf/ets-performance-assessment-history.pdf

He, Ye., & Levin, B. B. (2008). Match or Mismatch? How congruent are the beliefs of teacher candidates, cooperating teachers, and university-based teacher educators? *Teacher Education Quarterly, 35(4),* 37-55.

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

Hill, H.C., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2),371- 406.

Goldhaber, D. (2015). *Teacher effectiveness research and the evolution of U.S. teacher policy* (The Productivity for Results Series No. 5). George W. Bush Institute, Education Reform Initiative. Retrieved from http://gwbcenter.imgix.net/Resources/gwbi-teacher-effectiveness-research.pdf Google Scholar.

Gorlewski, D. A., & Gorlewski, J. A. (2015). Producing professionals: Analyzing what counts for edTPA. In K. A. O'Hara (Ed.), *Teacher evaluation: The charge and the challenges.* New York, NY: Peter Lang, pp. 19-37.

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal*, *45*, 184-205.

Labaree, D. (2010). Teach for America and teacher ed: Heads they win tails we lose. *Journal of Teacher Education, 61*, 48-55.

Lachuk, A. J., & Koellner, K. (2015). Performance-based assessment for certification: Insights from edTPA implementation. *Language Arts, 93*(2), 84-95.

Levine, A. (2006). *Educating school teachers. The education schools project*. Retrieved from http://www.edschools.org/teacher_report.htm

Liston, D., Whitcomb, J., & Borko, H. (2006). Too little or too much: Teacher preparation and the first years of teaching. *Journal of Teacher Education*, *57*, 351-358.

Lit, I., & Lotan, R. (2013). A balancing act: Dilemmas of implementing a high-stakes performance assessment. *The New Educator, 9*(1), 54-76.

Margolis, J., & Doring, A. (2013). National assessments for student teachers: Documenting teaching readiness to the tipping point. *Action in Teacher Education, 35*(4), 272-285.

Mathison, S. (1988). Why triangulate? *Educational Researcher*, 17 (2), 13-17.

Miles, B. M., & Huberman, M. (1994) *Qualitative data analysis: An expanded sourcebook* (2nd Edition). New York: Sage Publications.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy*. Washington, DC: The National Academies Press.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The performance assessment for California teachers (PACT). *Journal of Teacher Education*, *57*, 120-138.

Ravitch, D. (2013). *Reign of error: The hoax of the privatization movement and the danger to America's public schools.* New York, NY: Alfred A. Knopf.

Sandholtz, J., & Shea, L. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. *Journal of Teacher Education*, *63*(1), 39-50.

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education, 65,* 421–434.

Stanford Center for Assessment, Learning and Equity (2017). *edTPA annual administrative report.* Stanford, Ca.

Stanford Center for Assessment, Learning, and Equity (2018). *edTPA Participation Map*. Retrieved May 9, 2018 from http://edtpa.aacte.org/state-policy.

Sykes, G., Bird, T., & Kennedy, M. (2010). Teacher education: Its problems and some prospects. *Journal of Teacher Education, 61 (5),* 461-476.

Tabachnick, B.G., & Fidell, L. S. (1996). *Using Multivariate Statistics*. Harper Collins College Publishers: New York.

Tillema, H. H. (2009). Assessment for learning to teach: Appraisal of practice teaching lessons by mentors, supervisors, and student teachers. *Journal of Teacher Education, 60*, 155-167.

Wallace, M. R. (2009). Making sense of the links: Professional development, teacher practices, and student achievement. *Teachers College Record, 11*, 573-596.

Wayne, A., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*, 89-122

Whitney, L., Golez, F., Nagel, G., & Nieto, C. (2002). Listening to voices of practicing teachers to examine the effectiveness of a teacher education program. *Action in Teacher education, (23)4*, 69-76.

Wilson, S. F. & Youngs, P. (2005) Research on accountability processes in teacher education. M. Cochran-Smith, M. & K. Zeichner, (Eds). *Report of the American Educational Research Association panel on the research on teacher education*. Washington, DC: American Educational Research Association.

Wei, R.C., Pecheone, R.L., & Wilczak, K.L. (2014). *Performance Assessment 2.0: Lessons from Large-Scale Policy & Practice.* Stanford Center for Assessment, Learning, and Equity.

Whittaker, A., Pecheone, R., & Stansbury, K. (2018). Fulfilling our educative mission: A response to edTPA critique. *Education Policy Analysis Archives, 26*(30).

Wineburg, S. M. (2006). Evidence in teacher preparation: Establishing a framework for accountability. *Journal of Teacher Education*, *57*, 51-64.

Zeichner, K. (2007). Accumulating knowledge across self-studies in teacher education. *Journal of Teacher Education, 58*, 34-46.

Zeichner, K., Payne, K. A., & Brayko, K. (2015). Democratizing teacher education. *Journal of Teacher Education, 66*(2), 122-135.

Corresponding Author Contact Information:

Author Name: Ruchi Bhatnagar

Georgia State University, USA

Email: rbhatnagar@gsu.edu