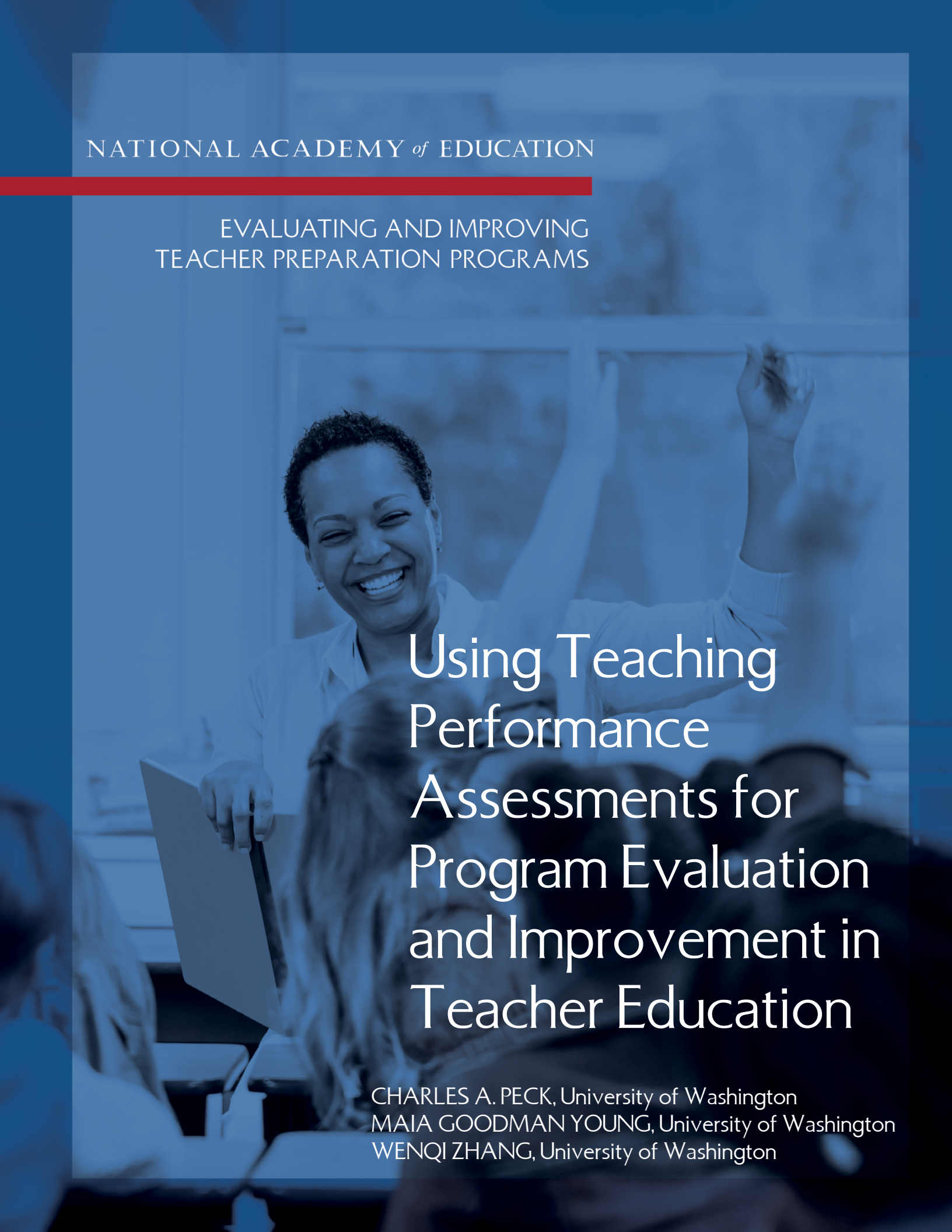


NATIONAL ACADEMY *of* EDUCATION

EVALUATING AND IMPROVING
TEACHER PREPARATION PROGRAMS

A photograph of a woman in a classroom, smiling and holding a folder, with her hand raised towards a whiteboard. The image is overlaid with a blue tint.

Using Teaching Performance Assessments for Program Evaluation and Improvement in Teacher Education

CHARLES A. PECK, University of Washington
MAIA GOODMAN YOUNG, University of Washington
WENQI ZHANG, University of Washington

NATIONAL ACADEMY OF EDUCATION 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project and research are supported by funding from the Bill & Melinda Gates Foundation. This paper was prepared for the National Academy of Education (NAEd) to inform and support the work of the steering committee for *Evaluating and Improving Teacher Preparation Programs*, including the consensus report. The opinions expressed are those of the authors and not necessarily those of the NAEd or the Bill & Melinda Gates Foundation.

Digital Object Identifier: 10.31094/2021/3/3

Copyright 2021 by the National Academy of Education. All rights reserved.

Suggested citation: Peck, C. A., Young, M. G., & Zhang, W. (2021). *Using teaching performance assessments for program evaluation and improvement in teacher education*. National Academy of Education Committee on Evaluating and Improving Teacher Preparation Programs. National Academy of Education.

National Academy of Education

Evaluating and Improving Teacher Preparation Programs

**Using Teaching Performance Assessments for Program Evaluation
and Improvement in Teacher Education¹**

Charles A. Peck, *University of Washington*
Maia Goodman Young, *University of Washington*
Wenqi Zhang, *University of Washington*

November 2021

CONTENTS

INTRODUCTION	2
CONCEPTUAL FRAMING AND RESEARCH QUESTIONS	2
TEACHING PERFORMANCE ASSESSMENTS: WHAT ARE THEY, AND WHAT ARE THEY USED FOR?	3
Design and Evaluation of TPAs.....	4
The Teaching Portfolio: Early Research and Development Work	6
Changing Policy Context: TPAs and the Press for Accountability.....	7
Locally Developed, Standardized TPAs.....	8
IMPLEMENTING STANDARDIZED TPAs AT SCALE: NEGOTIATING THE TENSIONS BETWEEN ACCOUNTABILITY AND IMPROVEMENT.....	9
What Are the Features and Affordances of Standardized TPAs?.....	9
What Questions Are Raised by the Uses of Standardized TPAs?	11
Equity and TPAs	13
TPAs and Decisions About Licensure.....	15
HOW ARE THE AFFORDANCES OF TPAs AS RESOURCES FOR PROGRAM IMPROVEMENT AFFECTED BY THE WAYS IN WHICH THEY ARE IMPLEMENTED?	15
State Policy Context.....	16
Program-Level Factors Affecting TPA Implementation.....	19
SUMMARY REFLECTIONS AND RECOMMENDATIONS	24
State Policy	25
Program Practices	26
Research Priorities	27
CONCLUSION	29
REFERENCES	30
AUTHOR BIOGRAPHIES	37

¹ We thank Eva Baker, Mary Sandy, Sheila Valencia, Dan Goldhaber, Tine Sloan, Kevin Bastian, and Mark Olofson for their intellectually generous and helpful reviews of our early drafts of this paper.

INTRODUCTION

In this paper we examine the uses of teaching performance assessments (TPAs) as resources for learning, program evaluation, and improvement in teacher education. In undertaking our review, we take note of our positionality as *practitioners* within the field of educator preparation. Our professional work as educators is carried out largely at the program level, inside the practical dilemmas of policy and practice that emerge as we attempt to *use* the TPA tools described in this paper. Our review is intended to be of value to policymakers, faculty, and academic leaders as they navigate the challenges and opportunities of using TPAs as resources for improving the work of teacher preparation.

Because of this, we conceptualize TPAs not simply as tests, but as historically situated forms of activity wherein teacher educators, preservice teachers, and policymakers navigate a complex interplay of policy mandates, conceptual and material resources, organizational conditions, and motivations related to diverse and often contested goals for assessment, accountability, and program improvement. This view is consonant with contemporary conceptualizations of assessment “validity” in which the consequences of *using* an assessment tool are evaluated in concert with its psychometric properties (Kane, 2013; Messick, 1994). Moss (2013) has elaborated on the idea of consequential validity to include attention to how assessment data are used in practice, arguing that “the focus of validity questions will need to shift again to the broader learning or organizational environment and the extent to which it is sufficiently well resourced to support an evidence-based professional practice that enhances student learning” (p. 96). In this paper, we are particularly interested in clarifying the conditions under which TPAs function (or do not function) as useful tools for decision-making related to issues of teacher licensure, program evaluation, and program improvement.

We begin by outlining our conceptual framing and related research questions about the uses of TPAs as resources for program evaluation and improvement. We describe some of the defining features and affordances of TPAs and consider how these measures engage traditional questions and concerns about the validity and generalizability of the findings they yield. Our discussion then proceeds with a brief history of teaching portfolios, locally developed performance assessments, and standardized measures of teaching performance, noting how the purposes of these assessments have been shaped by the increasing emphasis on external accountability in public policy over recent decades. Using this historical account as context, we focus the main body of our review on the contemporary research literature related to the uses of standardized TPAs in preservice teacher education. We conclude the paper with a set of recommendations for policy, practice, and research aimed at evaluating and improving the uses of TPAs as resources for the improvement of teacher education.

CONCEPTUAL FRAMING AND RESEARCH QUESTIONS

In this section we identify a set of working assumptions about the role of contextual factors in shaping how TPAs are used as tools for measuring and improving preservice teaching quality. First, as noted above, we assume that TPAs, as a specific form of professional activity, are embedded in larger historical and policy contexts that shape how they are understood and enacted. We assume that these policy contexts are themselves

affected by interest group advocacy, including the actions of professional organizations. The types and uses of assessment in teacher education have changed dramatically over the past two decades as the public policy zeitgeist has shifted toward accountability goals (Bales, 2006; Cochran-Smith et al., 2018).

We assume that the effects of macro-level changes in public perception and public policy related to teacher quality and teacher education are mediated by local conditions affecting how assessment data are used in practice (Davis & Peck, 2020; Spillane & Meile, 2007). We attend particularly to the ways in which the values, beliefs, and perceptions of the people involved in program-level work shape their approaches to TPA implementation. We also consider what may be gleaned from the literature about the different ways in which TPAs may be used as tools for decision-making. Our interests include the variety of ways in which TPAs are scored, and how the data are disaggregated and represented to support access, interpretation, and action by program faculty and staff. Finally, we review research related to organizational policies, practices, and routines that afford or constrain opportunities to use TPAs as a resource for supporting candidate learning and program improvement, including the role that leadership plays in orchestrating the local negotiation of practice related to TPA implementation. We use this conceptual framing to engage the following questions:

1. What are the social and political contexts in which TPAs have emerged as prominent and often controversial tools for policy and practice in teacher education?
2. What are the characteristic features and related affordances of TPAs as tools for decision-making, including decisions related to candidate licensure, support, and program improvement?
3. How is TPA implementation mediated by the values, beliefs, and motives of the faculty, staff, and academic leaders in a teacher education program, the ways that the assessment tools are used, and the organizational policies and practices of the institutions involved?

TEACHING PERFORMANCE ASSESSMENTS: WHAT ARE THEY, AND WHAT ARE THEY USED FOR?

The term “performance assessment” refers to approaches to description and evaluation of human competence and skill based on evidence collected in the contexts of an individual’s participation in “authentic” activities of practice. A contemporary definition of the term offered by the Educational Testing Service refers to a performance assessment as “a test in which the test taker actually demonstrates the skills the test is intended to measure by doing real-world tasks that require those skills, rather than by answering questions asking how to do them” (Educational Testing Service, 2020).

This is by no means a new idea. Proposals for new forms of educational assessment based on performing “real-world” tasks gained momentum in the early 1990s in the context of widespread critiques of standardized achievement testing (Shepard, 1991) including tests used to evaluate the qualifications of prospective teachers (Darling-Hammond, 1986; Haertel, 1991). Particularly at issue in many of these critiques were questions about the relevance of traditional teacher test data to the tasks of improv-

ing preservice teachers' instruction in classrooms. Indeed, Messick (1994) described improvement of learning outcomes as one of the primary goals of performance assessment: "Exposure to authentic assessment is expected to provide the student with a meaningful educational experience that facilitates learning and skill development as well as deeper understanding of the requirements and standards for good performance" (p. 17). In the case of teacher education, the hope was that performance assessments would not only provide a means of evaluating preservice teachers but would also support their development of classroom practice.

The term "performance assessment" can refer to a variety of test designs used to sample or represent real-world activity, including simulations, projects, essays, demonstrations, and work products (Davey et al., 2015). In the context of preservice teacher education, however, the term "teaching performance assessment" most often refers to work samples or "portfolios" that integrate the collection, analysis, and evaluation of artifacts and related products derived from actual classroom teaching practice. Portfolios may consist of relatively informal collections of artifacts gathered over several months of practicum work in the classroom, or they may be highly standardized processes in which specific artifacts of teaching are required to be collected and analyzed by the candidate over the course of several lessons. These typically include data from (P-12) student classroom assessments, lesson plans, video records of teaching, and samples of student work, accompanied by analytic and reflective commentaries.

Data from TPAs are used within preservice teacher education to evaluate individual candidate teaching practice, both as formative measures of candidate progress and as summative measures related to decisions about licensure. In addition, TPA data may be aggregated across candidates for the purpose of program evaluation, both with respect to evaluation of changes in local program learning outcomes over time and to enable comparisons across programs. Both program-level TPA data and artifacts from individual assessment portfolios can be used as resources for learning and program improvement.

Design and Evaluation of TPAs

TPAs differ from traditional forced-choice measurement methodologies in ways that are related to their focus on documentation and evaluation of performance in "real-world" contexts. For example, whereas traditional assessments of teacher knowledge attempt to equalize opportunities to demonstrate competence by standardizing the conditions under which tests are administered, TPAs typically rely on data from practice in actual classrooms, which introduces significant variation in the conditions under which these assessments are conducted. These include differences in students, cooperating teacher support, and school curriculum policies and practices. The focus of assessment is on the *process* of teaching, and equity in opportunity to perform is based on a standardized collection of planning documents, observational records of teaching, and samples of P-12 student work, with related analytic commentaries. In Table 1, below, we identify some of the design parameters and related evaluation questions that represent important considerations for evaluating TPAs. (For more detailed explication and discussion of design principles and evaluation criteria for performance assessments, see Davey et al. [2015], Khattri et al. [1998], and Moss [2013].)

TABLE 1 Design and Evaluation Considerations for Teaching Performance Assessments (TPAs)

Design Parameter	Key Question	Evaluation Focus	Resources
Content Validity	To what extent does the TPA measure important aspects of teaching practice?	Alignment of TPA with contemporary research on teacher effectiveness and/or professional standards of teaching practice (e.g., Interstate Teaching Assessment and Support Consortium)	Sato, 2014
Generalizability	To what extent are TPA scores consistent across raters?	Published studies of scorer training and inter-rater agreement	Bastian et al., 2016; Pecheone & Chung, 2006
	To what extent are TPA scores consistent across variation in students, curriculum domains, or school contexts?	Studies of effects of context variables on TPA scores	Adie & Wyatt-Smith, 2020; Bastian et al., 2020
Predictive Validity	To what extent are TPA scores correlated with measures of socially important educational outcomes?	Studies of the relationship of TPA scores and P-12 student achievement and teacher employment and retention	Bastian et al., 2016; Goldhaber et al., 2017
Consequential Validity	To what extent do TPAs screen ineffective teachers from entering the workforce?	Studies of the uses of TPAs in decision-making	Goldhaber et al., 2017; Ledwell & Oyler, 2016
	To what extent do TPAs lead to candidate learning?	Studies of preservice teacher learning	Chung, 2008; Lin, 2015
	To what extent do TPA data lead to program improvement?	Studies of program improvement process and outcomes	De Voto et al., 2020; Lys et al., 2016; Peck et al., 2010

The use of TPAs as resources for evaluation and improvement of instructional practice has been part of a larger vision and agenda for the professionalization of teaching, articulated most directly in the work of the National Board for Professional Teaching Standards (NBPTS; National Board for Professional Teaching Standards, 2021b). Within this view, a defining feature of a profession is its control and authority over standards of practice, membership, and evaluation within the profession itself, rather than external bureaucracies (Cochran-Smith et al., 2018; Darling-Hammond, 1986; Shulman, 1987a). The implicit theory of professionalization involves establishing a relationship between consensus on professionally defined standards of practice, localized self-study and peer-mediated assessment related to those standards, and commitment to learning in and from the contexts of actual practice as a resource for improvement (Darling-Hammond & Snyder, 2000). Within this framework, classroom-based performance assessments of teaching are envisioned at multiple points along a continuum of professional development, beginning with preservice teacher preparation, and continuing through

certification at the residency and advanced levels of professional practice (National Board for Professional Teaching Standards, 2021a).

While recognizing the potential of performance assessments as resources for improvement of teaching practice, Linn et al. (1991) also identified some of the tensions that accompany the assessment of complex, “real-world” activity:

If great weight is attached to the traditional criteria of efficiency, reliability, and comparability of assessments from year to year, the more complex and time-consuming performance-based measures will compare unfavorably with traditional standardized tests. (Linn et al., 1991, p. 16)

Tensions between accountability-driven assessment goals, which press for standardized control of assessment conditions to facilitate comparison of scores, and improvement-driven goals, which encourage more detailed description and analysis of performance in context, have proven to be thematic to the history of the development and implementation of TPAs over the past decades. Indeed, the design of TPAs has shifted over time as the field has responded to changes in policy priorities related to these tensions.

The Teaching Portfolio: Early Research and Development Work

Early development of the teaching portfolio began with a focus on improving the practice of classroom teachers through systematic collection and analysis of artifacts from a teacher’s actual work in the classroom and gained national prominence through the work undertaken by Lee Shulman et al. (1987a) in the Teaching Assessment Project at Stanford University in the late 1980s. Working with this group, Wolf (1991) defined a teaching portfolio as more than a “container” for records of a teacher’s practice, noting that “a portfolio also *embodies an attitude that assessment is dynamic* [emphasis added] and that the richest portrayals of teacher (and student) performance are based on multiple sources of evidence collected over time in authentic settings” (p. 130).

In one of the early empirical assessments of the teaching portfolio as a resource for learning and improvement of practice, Athanases (1994) reported on a yearlong process in which 24 elementary classroom teachers collected artifacts of their literacy-related work with their students. Almost all of the teachers reported improvements in their teaching as a result of their portfolio work, particularly with respect to their practices related to student assessment. Despite a number of methodological limitations noted by Athanases (1994), this report provided one of the first well-documented evaluations of the potential of “authentic” performance-based assessments of teaching as a resource for evaluation and improvement of teaching practice.

Subsequent studies of the use of teaching portfolios in teacher education programs suggested that many of the benefits of the process reported for experienced teachers by Athanases were also evident with preservice teachers (Borko et al., 1997; Darling-Hammond & Snyder, 2000; Zeichner & Wray, 2001). For example, Borko et al. (1997) studied the experiences of 21 teacher candidates who completed teaching portfolios during their student teaching practicum. Both written comments and interviews with the candidates identified opportunities for reflection and improvement of teaching as

the most frequently mentioned benefit of the portfolio process. Subsequent research in additional teacher education programs affirmed many of the findings of Borko et al. (1997), including the affordances of the teaching portfolio as an opportunity for candidate learning and improvement of practice, the importance of faculty and program support for the portfolio assessment process, and concerns about the time-intensive nature of the portfolio work (Anderson & DeMeulle, 1998; Snyder et al., 1998).

Two important themes are evident in these early studies of the teaching portfolio process in preservice teacher education. First, it is clear from multiple studies that the process of constructing and reflecting on a teaching portfolio represents a rich opportunity for learning and improvement of practice for preservice teachers (Anderson & DeMeulle, 1998; Borko et al., 1997; Snyder et al., 1998). Anderson and DeMuelle (1998) also noted the affordances of portfolio assessments as tools for *faculty* learning and *program* improvement but noted these were taken up less frequently in the programs they studied.

Second, as anticipated by Linn et al. (1991), almost all of the investigations of portfolio implementation reported significant tensions between the multiple uses of teaching portfolios as resources for candidate learning, for licensure decisions, and for program improvement (Borko et al., 1997; Delandshere & Arens, 2003; Snyder et al., 1998; Zeichner & Wray, 2001). These tensions have become more salient as public concerns about teacher quality and related policy pressures for teacher education program accountability and improvement have intensified over the past two decades (Cochran-Smith et al., 2018; Delandshere & Petrosky, 2010).

Changing Policy Context: TPAs and the Press for Accountability

Concerns about the quality and effectiveness of teacher education programs have existed for many decades (Conant, 1963; Goodlad, 1990; Levine, 2006; Sarason, 1993). However, these concerns have intensified as public confidence in the effectiveness and efficiency of government institutions has declined broadly. Policymakers have responded to these changes with mandates for increased external accountability for public institutions, increased investment in private-sector actors as alternatives to government programs, and increased reliance on market-based theories of program improvement (Boreham, 2004; Tröhler et al., 2014).

Congruent with the larger shift in the policy landscape, public policy interventions in the field of teacher education have intensified dramatically over the past two decades (Cochran-Smith et al., 2013). The new policies reflect a significant shift away from reliance on local professional judgment in decisions about teacher licensure toward policies emphasizing standardized measurement and external accountability (Cochran-Smith et al., 2018; Crowe, 2011). Many of the tensions identified in the early research and development work on teaching portfolios have become more salient as these types of “authentic” assessments have been appropriated within contemporary policy initiatives focused on accountability. For example, while early examples of portfolio assessment in preservice teacher education were included as *one of many* sources of data in local decisions about teacher licensure, many contemporary policies require passing scores on a TPA as a condition for licensure. So, while the tensions between the uses of performance assessments as resources for learning and their reliability and validity when

used as tools for consequential decisions about licensure have been recognized since some of the earliest research and development work (Gellman, 1993; Linn et al., 1991; Messick, 1994), policy changes affecting how these tools are used have made these tensions more visible and more problematic (Gitomer et al., 2019).

At the same time, the policy pressures for increased standardization and accountability in teacher licensure have offered the field opportunities to advance the professionalization of teaching by linking a national conversation aimed at building consensus around professional standards of teaching practice with the tools of authentic, practice-based assessment (e.g., NBPTS and the National Board Certification Process), and by introducing these tools into state and national teacher education policy initiatives (Darling-Hammond, 2010). A crucial design principle underlying this work is that both standards of practice and TPAs based on those standards should be developed by teachers (and teacher educators) themselves (Haertel, 1991). Indeed, the locus of power and control over the design of assessment tools used to evaluate candidate teaching performance has been one of the most significant focal points of controversy among teacher educators.

Locally Developed, Standardized TPAs

One strategic response to the tensions between increased external public policy pressures for accountability and local program values regarding authenticity, autonomy, and agency in assessment of preservice teacher quality is evident in program-level efforts to develop more *standardized* tools for assessing teaching performance. An early example of this approach was the Teacher Work Sample (TWS) methodology developed at Western Oregon University (Schalock, 1998; Schalock et al., 1997). The TWS assessment process provided for a standardized collection and evaluation of artifacts of preservice teachers' practice collected over a 3-5 week period, including a description of the classroom context, lesson plans, samples of student work, and plans for revision of instruction based on analysis of student learning data. Follow-up surveys and interviews with teacher candidates and faculty using the TWS indicated that the performance assessment process was a valuable resource for the improvement of teaching practice for both teacher candidates and faculty (Reusser et al., 2007). At the same time, data from the TWS provided an important evidentiary warrant for both state program certification and national accreditation (Schalock, 1998).

The TWS methodology was subsequently taken up by the Renaissance Group, a consortium of university-based teacher education programs, and adapted in ways that were congruent with differences in state policy contexts in which member institutions were situated. In one example of this strategy, faculty and academic leaders at California State University, Fresno (Fresno State), used the TWS as a foundation for developing the Fresno Assessment of Student Teachers (FAST) as a strategic response to new California state mandates for the use of a standardized performance assessment as part of the program certification process (Torgerson et al., 2009). In reports on both the Oregon and Fresno State work, authors noted the importance of local ownership of the assessment process as integral to faculty support and engagement with the challenging work of implementing a TPA. Torgerson et al. (2009) elaborated on the importance of this feature of the work at Fresno State:

In contrast to other university programs that had to select a performance assessment and secure faculty support and buy-in, the Fresno State faculty effort, expertise, and investment in the creation of FAST made its adoption a natural part of a multi-year process to improve programs and assessment. (p. 80)

The complex interplay between state policy mandates, the affordances and constraints of standardized TPAs as tools for improving teacher education programs, and the dynamics of how these tools are implemented at the program level is the focus of the remainder of our review.

IMPLEMENTING STANDARDIZED TPAs AT SCALE: NEGOTIATING THE TENSIONS BETWEEN ACCOUNTABILITY AND IMPROVEMENT

In undertaking our review and analysis of contemporary uses of standardized TPAs at the state and national level, we are interested specifically in how the implementation of the TPA process is mediated not simply by features of the tools themselves, but also by contextual variables related to state policy, faculty perceptions and motivation, and organizational supports affecting how they are implemented in practice (Cohen et al., 2020; Moss, 2013; Peck & Davis, 2019). We assume that standardized TPAs, particularly when used across programs in the context of state policy mandates, represent a negotiated response to the tensions between external pressures for accountability and local priorities related to authenticity and usefulness of the assessment. We do not imagine these tensions disappearing in the foreseeable future. We also understand these tensions as a potential impetus for learning, organizational change, and improvement (Engeström, 1987; Foot & Groleau, 2011). Consequently, our analysis is aimed at identifying what can be learned from the extant research literature about these tools and how they might be used in ways that afford equitable evaluation of teacher candidates as well as useful resources for candidate learning, program evaluation, and program improvement.

What Are the Features and Affordances of Standardized TPAs?

Standardized TPAs are designed to gather and evaluate data from actual teaching practice, including not only observational records of interactions between teachers and students but also artifacts of the kinds of cognitive processes that are involved in planning, enacting, and improving teaching. Based on their alignment with widely accepted national standards for teaching practice (NBPTS; Interstate Teaching Assessment and Support Consortium [InTASC]), standardized TPAs generally require descriptions of classroom context and student demographics, examples of lesson and unit planning, records of instruction (either direct observation or via video recording), samples of student work with accompanying analysis of student learning outcomes, and follow-up instructional plans based on student learning data. Standardized TPAs also differ from one another in significant ways, including the length of time during which artifacts of the teaching process are collected and analyzed, the specific prompts and evaluation rubrics used in the assessment process, and procedures for calibration of scorers. The practice-based and integrated view of teaching practice that emerges through this kind

of holistic assessment allows for a richer and more contextualized view of teaching competence than either standardized tests of content and pedagogical knowledge, or coursework assignments and projects, which focus on the specific aspects of teaching addressed in individual courses (Sloan, 2013; Snyder et al., 1998).

The holistic, integrated, and contextual view of teaching that TPAs can provide offers unique opportunities for evaluation and improvement of practice for both candidates and faculty. For candidates, some of the most significant challenges of learning to teach involve *integrating* what they have learned from various courses and fieldwork experiences. For example, the actual practice of teaching requires candidates to integrate and align what they have learned about culturally responsive teaching (Gay, 2001), equitable classroom management (Milner & Tenore, 2010), and the subject-specific pedagogies of academic disciplines such as literacy or mathematics (Shulman, 1987b). TPAs can provide concrete evidence of how candidates are managing these challenges in their classroom practice, thus creating opportunities for feedback, evaluation, and learning that may otherwise be lost in the silos of instruction and evaluation that are challenges to building and sustaining coherence in most teacher education programs (Grossman et al., 2008; Richmond et al., 2019).

Practice-based, integrative descriptions of teaching can be particularly valuable in the context of *collective* processes of program evaluation and improvement undertaken by faculty, field supervisors, mentor teachers, and program administrators, each of whom may otherwise have only a relatively narrow understanding of the program as a whole (Sloan, 2013; Snyder et al., 1998). These kinds of assessment data can provide concrete information about what candidates are taking up (or not) from their coursework and fieldwork experiences. This integrated, field-based view of candidate teaching practice can be particularly useful in identifying possible sources of program outcome problems (or virtues) that show up in broader, more decontextualized program evaluations such as graduate or employer satisfaction surveys or value-added assessments of P-12 student achievement in program graduates' classrooms (Davis & Peck, 2020).

A standardized TPA also affords an opportunity for faculty, field supervisors, and mentor teachers to develop a common and concrete language of practice, which in turn supports the kinds of communication and collaboration that are critical resources for professional learning and programmatic change (Bloxham et al., 2016; Nicolini et al., 2012; Nolen et al., 2011). The development of a common language of practice takes place as faculty negotiate a shared understanding of terms used to evaluate the practical activity of teaching as required for consistent scoring of teaching portfolios (Sloan, 2013; Whittaker & Nelson, 2013). It is important to note that having a common language does not require or imply agreement in decisions about practice, nor does it imply adherence to static philosophical views about teaching (Buchmann & Floden, 1992). Rather, the idea is that a common language of practice allows program members to share a relatively clear understanding of the questions and alternatives under consideration in decisions about practice (Hall & Horn, 2012). Indeed, without a common language as a tool for understanding and negotiating differences in values and practices across program participants, it is hard to imagine how programs may achieve or sustain intellectual or practical coherence, or even recognize instances of incoherence across courses and across coursework and fieldwork settings.

What Questions Are Raised by the Uses of Standardized TPAs?

Early and enthusiastic appraisals of the potential of performance assessments as a resource for improvement of instruction also consistently recognized the inherent tensions between highly contextualized descriptions of practice such as those offered by TPAs and the challenges of making equitable comparisons of practice across individuals and settings (Haertel, 1991; Linn et al., 1991; Messick, 1994). As we noted earlier, the practical implications of these tensions have become more significant as the uses of TPAs have expanded in the context of intensifying accountability policy. In this section we identify some of the recurring questions that have been raised about using TPAs for the purposes of comparison, particularly in the context of high stakes decisions about teacher licensure (Adie & Wyatt-Smith, 2020). In summarizing some of the research related to these concerns, we focus particularly on three well known examples of standardized TPAs: The TWS (Denner et al., 2004; McConney et al., 1998), the Performance Assessment for California Teachers (Pecheone & Chung, 2006), and the edTPA (Stanford Center for Assessment, Learning, and Equity, 2015). Our interests are not in critiquing these specific measures, but in clarifying some of the strengths and limitations they share. We organize our discussion in terms of some of the practical questions that policymakers and practitioners must consider in using these tools, and in interpreting the data they produce with respect to both accountability and improvement purposes.

Are TPAs Valid Measures of Teaching Practice?

TPAs are typically aligned with state and national standards for teaching practice, including those developed by InTASC and NBPTS. These standards of practice represent a decades-long effort by teachers and teacher educators to establish some national consensus on standards and expectations for practice to be used to guide the preparation and evaluation of teachers (Shulman & Sykes, 1986). A considerable body of evidence suggests that TPAs based on these standards do measure important aspects of teaching practice (Bastian et al., 2016; Campbell et al., 2016; Cooner et al., 2011; Denner et al., 2004; Pecheone & Chung, 2006; Sato, 2014; Stanford Center for Assessment, Learning, and Equity, 2015; Stewart et al., 2015). In addition, several studies have found that scores on these kinds of measures may predict the effectiveness of novice teachers after licensure, although these findings are not consistent across studies (Bastian et al., 2016; Chen et al., 2021; Darling-Hammond et al., 2013; Goldhaber et al., 2017; Stanford Center for Assessment, Learning, and Equity, 2015).

Some standardized TPAs have been criticized for not attending more substantively to issues of teaching practice related to social justice and equity (National Association for Multicultural Education, 2014; Stillman et al., 2013; Tuck & Gorlewski, 2016). Particularly at issue are relational dimensions of teaching that are difficult to measure (Behizadeh & Neely, 2019; Choppin & Meuwissen, 2017). For instance, Choppin and Meuwissen (2017) reported that candidates felt that the videos produced for the edTPA did not capture the relationship building that happened between teachers and students in informal interactions outside of class time. Others have argued that performance on TPAs is conflated with skills that are not essential for teaching, such as candidate writing skills (Behizadeh & Neely, 2019), technology skills (Choppin & Meuwissen, 2017), and “test-wiseness” (Clark-Gareca, 2015). These are important concerns that merit more

focused and rigorous research than has been reported to date. One example of related work under way is the recent revision of the CalTPA (https://www.csusm.edu/soe/currentstudents/caltpa_overview.pdf), with the accompanying direct investigation and evaluation of its features using criteria drawn from the literature on equity and racial justice in teacher education (Escalante et al., 2021).

Some recent research suggests that candidate performance on a TPA may also be affected by the quality of program preparation and support candidates receive in preparing for the assessment (Adie & Wyatt-Smith, 2020; Kim & Sato, 2019), including the characteristics of mentor teachers and student teaching placement schools (Bastian et al., 2020). Of course, programs *should* affect teaching practice—that is their mission, and the rationale for investing in program improvement. However, when achieving a certain TPA score is used as a requirement for licensure, there is a concern about the validity of comparative judgments about teacher quality in the context of differences in opportunities to learn and perform. From one perspective, the primary concern should be to protect the educational welfare of the P-12 student by ensuring that prospective teachers are competent (i.e., that they can pass rigorous licensure requirements such as a TPA). However, from a teacher candidate’s perspective, an additional concern is that decisions about licensure be made equitably, with appropriate consideration of the ways context—including both coursework and practicum work—may affect teaching performance. For example, Bastian et al. (2020) found that student teaching placement characteristics had a statistically significant impact on edTPA scores, a finding congruent with other research on the impact of placement characteristics on the learning and effectiveness of novice teachers (Goldhaber et al., 2017; Ronfeldt, 2015). While these studies raise questions about the ways that TPA scores—and indeed other measures of candidate learning and performance—should be interpreted in the context of significant differences in opportunities to learn in both university and field-based aspects of the teacher preparation program, we note that these tensions are by no means restricted to evaluations of teaching. They also exist in other fields of professional education such as law, medicine, and architecture, in which they are consistently resolved in favor of protecting the standards of the profession and the interests of the populations to be served.

Practical strategies for improving the validity of judgments about individual teaching in ways that protect the interests of both candidates and P-12 students include collecting broader samples of teaching practice, both across time and context, as these are likely to yield more robust findings about the quality of preservice teacher practice. For example, sampling preservice teaching performance in multiple field settings representing variation in student populations, mentor teacher characteristics, and school curriculum policy allows for some direct evaluation of the impacts of these contextual factors on the quality of candidate teaching practice. The costs of collecting and evaluating broader samples of teaching performance may be ameliorated by embedding these as formative assessments at multiple points within a program of teacher preparation (e.g., Alloway & Lesh, 2019).

Other concerns about the validity of TPAs may be grounded in differences in values and beliefs about teaching, including those that may exist between university-based teacher educators and local practitioners and community members (Zeichner et al., 2015). The recent and dramatic increase in public attention and engagement with issues

of racial equity and social justice also reminds us that educational values and priorities change over time. This suggests that questions about the validity of TPAs are likely, and appropriately, to be a focus of ongoing dialogue and deliberation between researchers, practitioners, teacher educators, and the communities they serve.²

Do Different Evaluators Rate TPAs Consistently?

The challenges of achieving consistent inter-rater agreement in scoring and evaluation of TPAs have been noted since the introduction of these assessment methodologies (Haertel, 1991; Messick, 1994). Psychometric studies carried out with several TPA instruments suggest that these challenges persist across differences in specific work sample content and protocols for rater training and scoring. For example, while satisfactory inter-rater agreement was achieved in scoring Renaissance TWSs when the assessments were evaluated by panels of three raters (Denner et al., 2009), a subsequent study of rater discourse during these evaluation processes showed that, while agreement was often achieved, it was not necessarily based on shared understanding and evaluation of the candidate's work, but on locally negotiated conventions for achieving agreement (e.g., "splitting the difference") (Bullough, 2010).

Reliability studies for both Performance Assessment for California Teachers (PACT) and edTPA have similarly suggested that while substantial inter-rater agreement can be achieved for these measures with strong training and audits that conduct back-reads (Pecheone & Chung, 2006), the procedures used to calculate agreement for these measures may mask meaningful differences in rater evaluations (Gitomer et al., 2019; Peterson & Lyness, 2015; Porter & Jelinek, 2011). For these reasons it is important to plan for multiple scorers near the cut point that determines passage of an assessment (e.g., Whittaker et al., 2018). Concerns about inter-rater agreement in the scoring of TPAs are not restricted to these three measures. Detailed psychometric studies of other TPAs (e.g., Riggs et al., 2009) reflect similar challenges in achieving consistency in both calibration and scoring procedures. In other cases, relevant psychometric data are not included in published descriptions of the instrument (e.g., Meyer et al., 2018).

The consistency with which difficulties in achieving agreement among scores across raters have been reported suggests the importance of careful consideration about the kinds of decisions for which these tools are used, as well as additional care in rescoring assessments near the cut score for passage. These findings additionally underscore the importance of rigorous scorer training, as well as ongoing re-calibration and assessment of inter-rater agreement. Finally, it is important that inter-rater agreement data be *used* to continuously evaluate the scoring process and to identify both the needs for general changes in scorer training, as well as interventions to remediate the performance of individual scorers (e.g., Meyer et al., 2018; Stanford Center for Assessment, Learning, and Equity, 2015).

Equity and TPAs

Concerns about racial equity related to educational assessment in higher education are pervasive and compelling (Stewart & Haynes, 2015). Petchauer et al. (2018) trace

²The CalTPA work cited above is an example of what this might look like.

the history of standardized assessment in teacher education, noting how racial group differences in performance on teacher tests have been an ongoing focus of concern, particularly with respect to the gatekeeping functions of these tests (Gitomer et al., 2011; Goldhaber & Hansen, 2010). Although the focus of these concerns historically had been on multiple choice tests of academic skills and subject-matter knowledge, where racial disparities in performance have been more pronounced, research on standardized TPAs has shown that some historically marginalized cultural and/or racial groups may not perform as well as their White counterparts on these assessments. For example, Goldhaber et al. reported that, although failure rates were relatively low for both groups, Hispanic teacher candidates in Washington State were three times more likely to fail the edTPA than White candidates (Goldhaber et al., 2017). Petchauer et al. (2018) note that studies comparing the performance of candidates from different racial groups on PACT found lower test scores for Black candidates (Stanford Center for Assessment, Learning, and Equity, 2015), but did not replicate the findings of Goldhaber et al. (2017) regarding Hispanic candidates. Taken together, these results suggest that equity concerns related to standardized TPAs warrant careful and conservative approaches to interpretation of test score data, particularly in the context of high stakes decisions about licensure. However, we note Haertel's (1991) early questions about the sources of these risks:

The fact that, on average, some racial/ethnic, gender, or other identifiable groups outperform others on an examination does not in itself imply that the examination is biased, but should nonetheless trigger a careful scrutiny of the test itself, the conditions of its use, and the prior preparation of examinees to ensure, to the extent humanly possible, that test bias is not present. (p. 24)

While a facile response to equity concerns about TPAs might be to eliminate the tests, it is entirely possible that racial and/or cultural differences in TPA pass rates signal problems with the quality and effectiveness of program support and preparation of candidates to succeed on these assessments. For example, in one survey study of the experiences of candidates of color with the edTPA (Williams et al., 2019), candidates of color perceived themselves to be more ready for the assessment than White candidates perceived themselves to be, and yet they failed more often than White candidates. While the study does not provide enough detail to allow for unambiguous interpretation of the interactions that took place between these candidates and their instructors, the authors suggest the possibility that the quality of faculty feedback to candidates may have contributed to poor alignment of candidates of color's self-assessment of their readiness with their subsequent performance on the TPA. Eliminating the TPA as a measure of candidate readiness to teach may erase a signal of inequitable practices in teacher preparation without engaging their sources.

Several recent studies have documented significant differences in both the quality and quantity of faculty and program-level supports for candidates as they prepare for high stakes TPAs (Cohen et al., 2020; De Voto et al., 2020; Ratner & Kolman, 2016). The extent to which these differences affect TPA performance is still unclear. In the Williams et al. (2019) study, candidates' edTPA scores were positively related to the number of support activities that they attended. The Cohen et al. (2020), Denner et al. (2009), and Ledwell and Oyler (2016) reports all show that significant variation in these kinds of

supports may exist even across programs within a single institution. As mentioned above, findings from Bastian et al. (2020) suggest that at least one program variable, the quality of mentor teachers selected for supervision of practicum experiences, had a substantial impact on TPA performance. Taken collectively, these studies suggest the quality of program preparation and support, including the selection and support of mentor teachers, is indeed likely to affect TPA scores. The findings amplify concerns about equity related to the uses of standardized TPAs as high stakes measures of teaching quality, but also suggest that the locus and burden of accountability policies may be disproportionately placed on candidates rather than programs.

TPAs and Decisions About Licensure

The considerations involved in using TPAs as high stakes measures of readiness to teach are particularly acute in cases where candidate scores are close to the established “cut scores” for licensure (Goldhaber et al., 2017). Following advice from assessment researchers and test developers, augmentative procedures such as double and triple reading of portfolios scoring at or near cut scores for licensure have been developed and implemented for several TPAs, such as those used with National Board Certification, the edTPA, and the TWS. We might reasonably anticipate that further research and development work could improve both the instruments and procedures for scoring them. However, Davey et al. (2015) also argue that “lower levels of score comparability may simply need to be accepted as the price for measuring otherwise inaccessible constructs or for measuring in more direct ways,” further noting that “less than fully comparable scores may also be perfectly acceptable for use in lower stakes circumstances” (p. 52).

This conclusion is consistent with commentaries on performance assessment from other fields of human service that underscore both the importance and the psychometric challenges of considering issues of social context in evaluating professional practice (Govaerts et al., 2007; Truijens et al., 2019). Contemporary views of learning, however, suggest that professional practice is not simply embedded in social context but *constructed* in large measure from the social and material resources those contexts provide (Lave & Wenger, 1991). These considerations suggest that the process of evaluating preservice teacher *practice*, particularly for the purpose of making equitable and appropriate decisions about individual teacher licensure, is inescapably *interpretive* in nature. These decisions are strengthened not only by reliance on multiple sources of relevant evidence (including TPA data) but also by careful and deliberative dialogue among local teacher educators, including supervisors and mentor teachers whose work is situated in the P-12 classroom (Moss, 1992; Moss et al., 1998).

HOW ARE THE AFFORDANCES OF TPAs AS RESOURCES FOR PROGRAM IMPROVEMENT AFFECTED BY THE WAYS IN WHICH THEY ARE IMPLEMENTED?

More than two decades of research on TPAs, including both informal records of teaching collected in portfolios and more standardized samples of teaching practice used in large scale assessments of preservice teachers, demonstrate that TPAs *can* provide meaningful opportunities for candidate and faculty learning and improvement of

practice (Borko et al., 1997; Bunch et al., 2009; Chung, 2008; Kohler et al., 2008; Lin, 2015). It is also clear that TPA data can be used effectively as a resource for program improvement (Cuthrell et al., 2019; Peck et al., 2010; Pointer Mace & Luebke, 2021; Reusser et al., 2007). However, it is equally clear that these outcomes are not always achieved and that the extent to which TPAs are actually used as resources for program evaluation and improvement depends very much on how they are implemented (De Voto et al., 2020). For instance, an implementation process that is rushed or lacks adequate resources can lead to increased faculty alienation and resistance to the TPA. In the following section, we describe some of the policy conditions and implementation processes that may affect how TPAs are utilized (or not) as resources for program improvement.

State Policy Context

A number of studies have investigated state policy conditions that may affect TPA implementation. These include the “stakes” of the assessment for candidates and programs, the speed of implementation, and the opportunities for inter-organizational collaboration in the implementation process.

High Stakes/Low Stakes

One of the most common concerns about the use of standardized TPAs is that the use of these assessments as high stakes requirements for teacher licensure induces candidates to engage them as a compliance task, overlooking their affordances as an opportunity for learning and improvement of practice (Behizadeh & Neely, 2019; Dover, 2018; Rennert-Ariev, 2008). This clearly happens in many cases. However, there is also evidence that some candidates may adopt a “compliance” stance even when completing *low* stakes TPAs, or even non-standardized teaching portfolios (Borko et al., 1997; Chye et al., 2019; Cronenberg et al., 2016; Darling-Hammond, 2010). There is also evidence that some preservice teachers engage TPAs with an inquiry and learning orientation, even when the stakes for licensure are high (Bacon & Blachman, 2017; Bunch et al., 2009; Lin, 2015; Okhremtchouk et al., 2013).

Programmatic responses to high stakes state policies requiring passing TPA scores for state licensure also vary dramatically (De Voto et al., 2020; Ledwell & Oyler, 2016). Some teacher educators interpret TPAs as a valuable opportunity for learning and improvement of both individual and collective practice and describe the value of these classroom-based data as a new and compelling source of insight and motivation for change. For example, one program director described this kind of proactive inquiry and improvement-oriented faculty responses to PACT data in this way:

The persuasive piece was once they saw the student work. I mean, where a few people kind of went, “Whoa.” I teach this in my class and I’m not seeing it ... looking at the student work from the mock scoring there was that “ah hah” moment where [it was clear that] our candidates didn’t know much about (academic language)... [One professor] changed her entire series of assignments ... to better reflect what the holes in the data [showed]—also to incorporate more clearly the notion of academic language and mathematics. She literally rewrote everything related to that assignment because

it was so compelling to her, the data ...and seeing the student work. (cited in Darling-Hammond, 2006, p. 19)

In other cases, faculty may engage state mandates related to accountability with clear intention to deflect what is interpreted as an unwarranted intrusion of state control into matters that have historically been the domain of local faculty discretion and decision-making (Kornfeld et al., 2007). However, while faculty and academic leaders may share concerns about loss of local control over licensure decisions, some nevertheless elect to organize local implementation of mandated TPAs with the goal of using the data as an opportunity for program-level learning and improvement (De Voto et al., 2020; Lys et al., 2014; Peck et al., 2010).

These studies suggest that, while high stakes uses of standardized TPAs may be problematic for many reasons, their effects on both candidate and faculty learning and improvement of practice are inconsistent and are strongly mediated by factors such as local leadership action (we discuss this in more detail later in the paper). In view of this finding, we consider additional contextual factors related to state policy that may affect how TPAs are perceived and used.

Pace/Speed of Implementation

There have been notable differences in the pace of mandated TPA implementation by policymakers in different states. In two of the earliest statewide TPA implementation efforts (in California and Washington State), state policymakers took a relatively gradual approach to implementation. These states required a passing score on the standardized TPA as a high stakes requirement for teacher licensure only after a substantial period of planning and preparation during which the implementation of a TPA was a program requirement, but not used directly for licensure decisions (Chung, 2008; Peck et al., 2012). It is worth noting, however, that even with what in retrospect might be considered a relatively gradual schedule for full implementation of the new assessment policies, teacher educators in both California and Washington reported considerable stress and confusion as the implementation process unfolded at the state level (Lit & Lotan, 2013; Peck et al., 2012). We note that while stress may be an inevitable side effect of TPA implementation—the *meanings* of the process appear to be strongly mediated by how its purposes are understood by program members. Change efforts oriented around goals for inquiry and program improvement do not avoid stress but often do yield programmatic changes that are highly valued by teacher educators (Darling-Hammond, 2006; Peck et al., 2010; Whittaker & Nelson, 2013). In contrast, the work (and stress) required for TPA implementation when it is oriented toward compliance goals appears more likely to be experienced as alienating and “subtractive” in its effects on program integrity and faculty autonomy (DeVoto et al., 2020).

In some states, most notably New York, policymakers have elected to require implementation and high stakes use of a standardized TPA quite rapidly (Clayton, 2018a; Ledwell & Oyler, 2016; Reagan et al., 2016). In a comparative review of edTPA implementation reports from six early adopter states, Reagan et al. (2016) concluded:

rapid implementation of the edTPA in New York may have resulted in pushback and adjustments to the full-scale consequential implementation of the edTPA...

and that

...the timeline for the edTPA in New York may have limited meaningful dialogue among the multiple actors involved in the implementation of the assessment. (Reagan et al., 2016, p. 16)

We conclude that while a slower pace of implementation does not necessarily avoid stress and confusion, it is likely to allow better communication and more effective planning and problem solving between state policymakers, teacher educators, and P-12 partners. This, in turn, can lead to more meaningful use of TPAs as tools for achieving program improvement goals that matter to program faculty, staff, and candidates.

Inter-Organizational Collaboration and Support

Several studies have highlighted the value of inter-organizational collaboration as a support for TPA implementation (Olson & Rao, 2017; Peck et al., 2012; Warner et al., 2020). In some cases, collaborative relationships between teacher education and state policy agencies have been reported to facilitate program-to-program communication and problem solving as implementation challenges have arisen. One example of this kind of collaboration between state policymakers and teacher educators was reported by Meyer et al. (2018), who described the development and implementation of a standardized, high stakes TPA in the State of Kansas. The authors noted that strong state administrative support for the assessment, coupled with local program policy of incorporating scoring costs into student tuition, allowed the TPA to be implemented at a relatively low direct cost to students.

In another state-level study of TPA implementation, Peck et al. (2012) described the collaborative effort between the American Association of Colleges for Teacher Education (AACTE)-affiliated network of teacher education programs and state policy administrators in Washington State during the pilot phases of edTPA implementation. Based on follow-up interviews with 26 program administrators, faculty, and field supervisors as well as documents collected from 8 programs participating in a field test of the edTPA, Peck et al. concluded that

it appears to us that the most potent resource available for supporting positive implementation outcomes may be the collaborative relationship that has been forged between the PESB (the state Professional Education Standards Board), edTPA (SCALE) and WAACTE programs. (p. 22)

The value of building state-level supports for inter-organizational collaboration related to TPA implementation was also underscored in an account of edTPA implementation in the State of Illinois (Olson & Rao, 2017). Using an analysis of state and program-level documents describing edTPA implementation efforts in the state between 2010 and 2017, Olson and Rao describe a formal collaboration between state policymakers, SCALE, and the state network of teacher education program affiliated

with AACTE: the Illinois Teaching Performance Assessment Consortium. Olson and Rao found that this inter-organizational structure functioned as an important support for both programs and policymakers as they navigated the practical challenges of edTPA implementation, including development of a shared understanding of the TPA, and cross-program sharing of specific strategies and resources for faculty and candidate support.

Program-Level Factors Affecting TPA Implementation

Local program characteristics, policies, and practices clearly affect TPA implementation (De Voto et al., 2020; Ledwell & Oyler, 2016). These include the values, beliefs, and motivations of the people undertaking the work of implementation, the ways in which the TPA itself is scored and used, and the organizational supports available to candidates, faculty, and program staff related to the work of implementation. We consider each of these program characteristics separately, and then discuss the critical function of program leadership in shaping how they interact with one another in making TPAs function as useful resources for program improvement.

People: Local Values, Perceptions, and Interpretations of the TPA

Research from the P-12 sector has shown that teachers' responses to reform policy initiatives are deeply rooted in interactions between teachers' prior values, beliefs, and perceptions about the purposes of their work and their perceptions of the purposes underlying the proposed reform (Coburn et al., 2009; McLaughlin, 1987; Spillane, 2000). Similarly, several studies of TPA implementation make it clear that what has been referred to as the "stance" of program faculty and staff can affect TPA implementation, including the quality and amount of support provided to teacher candidates in preparing for the assessment (Cohen et al., 2020; De Voto et al., 2020; Ledwell & Oyler, 2016; Ratner & Kolman, 2016). For example, in a study of faculty working in teacher education programs situated in public colleges of education in New York City, Ratner and Kolman (2016) collected open-ended survey responses from nine faculty members and conducted follow-up interviews with three of these faculty. As the authors note, this study was conducted in the context of the unusually rapid implementation of a high stakes TPA mandate in the state of New York. They found that faculty responses to the TPA initiative were quite varied, ranging from wholesale rejection of the purpose and legitimacy of the TPA to cautious support. An important finding, however, was that while all faculty participants in this study expressed commitment to their students' learning, differences in faculty stance toward the TPA were associated with substantive differences in both the extent and types of practices with which they engaged the work of preparing candidates to pass the TPA.

The Ratner and Koman (2016) study is quite limited in terms of both its participant sample and methodology (as the authors acknowledge). However, their findings are consistent with those from additional studies of the relationship between faculty stance and TPA implementation (Cohen et al., 2020; De Voto et al., 2020; Ledwell & Oyler, 2016). De Voto et al. (2020) used data from interviews, focus groups, and documents to investigate the ways that 69 faculty, staff, and teacher candidates from 8 teacher

education programs in 2 states interpreted and implemented the edTPA. They found that when faculty, staff, and teacher candidates interpreted the TPA to be in alignment with local program values and practices, their commitment to implementing the assessment and using it for program improvement was relatively high. Conversely, in programs where the TPA was viewed to conflict with local values, both the will and capacity of programs to implement the assessment was reduced. De Voto et al. (2020) noted that the relationships between faculty stance and TPA implementation were not uniform and appeared in some cases to be mediated by additional variables, including the quality of local program leadership, and availability and allocation of resources to support implementation.

Cohen et al. (2020) report additional evidence about the relationship between the interpretive stance of program members and the supports provided to candidates in preparing for the edTPA. Using survey and interview data collected from faculty, field supervisors, and candidates across multiple programs situated in a research-intensive university, Cohen et al. (2020) found that candidates were quite sensitive to differences in faculty stance toward the edTPA. Candidates viewed these differences in faculty stance to be related to differences in the supports they received in preparing for the assessment. Program members' stance about the edTPA was also found to vary by role, with tenure line faculty reporting more negative views about the assessment than field supervisors, being less likely to participate in professional development activities aimed at developing a shared understanding and commitment to implementing the assessment, and less likely to recognize the affordances of the edTPA as a resource for improvement of practice.

Tools: How TPA Data Are Scored, Disaggregated, and Organized

Several recent studies suggest that how TPA data are analyzed and presented can affect the value of the data as a resource for learning and improvement.

Local scoring. One issue that has received considerable attention has to do with local versus external scoring of TPA portfolios. Accounts of TPA implementation consistently describe tensions between the value of local scoring of TPAs by program faculty and staff, and the value of having portfolios scored by individuals with no direct ties to local programs (Sloan, 2013; Warner et al., 2020). For example, Sloan (2013) described the important opportunities for learning and collaboration that attended local scoring, noting that faculty and staff, in learning to score the TPAs consistently, were forced to recognize and engage differences in their understanding and evaluation of teaching that had been previously obscured by “silos” of practice within the program. Sloan also reported that joint examination of candidate portfolios allowed all program members to see what candidates had (and had not) taken up from their coursework and integrated into their classroom practice. Sloan reported that the shared understanding of program outcomes that was afforded through the local scoring process became a resource for negotiating joint focus and commitment to actions related to program improvement. A similar process of faculty learning and engagement with change related to local scoring of TPAs was reported by Cuthrell et al. (2019). In this program, local scoring was gradually displaced by external scoring (via Pearson); however, the value of local

scoring was considered significant enough in this program that 20 percent of TPAs were retained for local co-scoring.

An important question about local scoring is whether results are affected by the positionality of local scorers as faculty and staff of the program in which candidates are enrolled. Pursuing this question, Bastian et al. (2016) compared edTPAs that were scored both locally and externally (via Pearson) in one large university program. Results of these comparisons for 64 teacher candidates showed that although edTPA scores were consistently higher when scored locally, local scores were nevertheless predictive of teacher evaluation ratings collected during candidates' first year of teaching. Bastian et al. (2020) concluded that while the inconsistency of local and external scoring results raises questions about using local scores for high stakes decisions about licensure, local TPA scoring may be valuable for the purposes of improving faculty and staff understanding of program outcomes and in engaging them in collaborative efforts toward program improvement.

Disaggregation of TPA data. Several reports of TPA implementation suggest that faculty learning and program improvement are affected by the ways in which TPA data are disaggregated for analysis and interpretation. Sloan (2013) and Whittaker and Nelson (2013) provide similar accounts of how faculty in two programs in California used specific examples of work from PACT portfolios as a context for collaborative, evidence-based conversation about what candidates had and had not taken up from their coursework and fieldwork experiences. Sloan (2013) commented that program members were particularly engaged in this collaborative work when "candidate work is on the table." Whittaker and Nelson (2013) describe how PACT data evaluating candidates' instruction of English learners were used to design a series of professional development activities for course instructors and field supervisors. These activities were followed by curriculum alignment and integration workshops in which program faculty collaborated to make major program changes aimed at improving candidates' practice related to supporting the academic language development of their P-12 students.

Bunch et al. (2009) also described an analysis of PACT portfolios focused on evaluation and improvement of candidate instructional practices with English learners. Using qualitative data analysis techniques, the authors found evidence that candidates used a variety of research-based curricula and instructional strategies in their teaching, including the use of multiple modes of representation in instruction, connecting to students' experiences in their cultural communities, and supporting students' use of their native language as a resource for learning. However, careful examination of candidate portfolio commentaries also showed the persistence of deficit orientations underlying some candidates' interpretation of student learning, particularly when their students were struggling. Bunch et al. (2009) commented on the value of "going beyond the scores" in using PACT portfolios as formative assessments of candidate learning, and as resources for program self-assessment and improvement.

A very different approach to disaggregation of TPA data was reported by Bastian et al. (2018). Using Latent Class Analysis (LCA), Bastian et al. identified four patterns of candidate instructional practice based on candidate scores on the individual rubrics of the edTPA. They comment on the potential value of LCA as a tool for identifying groups of teacher candidates that may need specific kinds of preservice supports, and

as a source of data for identifying priorities for induction year supports following licensure. The LCA methods described by Bastian et al. may be particularly valuable as a way of empirically identifying subsets of TPA portfolios to analyze in more depth as recommended by Bunch et al. (2009), Sloan (2013), and Whittaker and Nelson (2013).

Organizational Policies and Practices

The organizational conditions in which TPAs are implemented have significant effects on the extent to which opportunities for learning and improvement of practices are taken up by both candidates and program faculty. Several specific issues related to the support of TPA implementation have been identified in the literature.

Time and space for collaboration. TPA implementation research strongly suggests that actualizing the affordances of TPAs as resources for learning and program improvement requires strategic allocation of time and space for faculty collaboration (Davis & Peck, 2020; De Voto et al., 2020; Lys et al., 2014; Peck et al., 2010; Sloan et al., 2021). In some institutions, practical routines for supporting faculty collaboration are deeply institutionalized, and new kinds of TPA data are largely integrated into existing policies and practices (Pointer Mace & Luebke, 2021). More often, however, implementation studies suggest that new routines and practices must be developed to support the kinds of collaboration that are required to make TPA data useful for program improvement. For example, Lys et al. describe regular “data summits” designed to create a time and place for faculty analysis and interpretation of TPA scores in the context of decisions about curriculum changes. Similarly, Sloan (2013) described week-long program retreats in which regular coursework and fieldwork activities were suspended to allow program faculty and staff to jointly score TPA portfolios and discuss their implications for program improvement actions. The importance of making time and space for faculty and staff collaboration is underscored by the De Voto et al. (2020) cross-program research, in which these practices were associated with the adoption of an “inquiry” oriented approach to TPA implementation as a resource for program improvement.

Professional development. One of the challenges of TPA implementation is what Lys et al. (2014) refer to as faculty “readiness.” It is important to note that by “faculty” we refer here to university-based course instructors, including those in colleges of education and in arts and science departments, as well as field-based supervisors and mentor teachers—all of whom play essential roles in teacher preparation (Goodlad, 1990). Depth of teacher educators’ knowledge about a TPA is clearly critical to productive implementation, and opportunities to develop a shared and concrete understanding of the required artifacts, evaluation rubrics and procedures for scoring are among the factors contributing to the uses of TPA data for program improvement (Sloan, 2013; Whittaker & Nelson, 2013). Some implementation studies also suggest that opportunities to openly discuss concerns about both the tools and the policy contexts in which they are implemented are related to faculty stance toward the TPA, particularly in the context of external policy mandates (De Voto et al., 2020; Lys et al., 2016; Peck et al., 2010).

Suleiman and Byrd (2016) describe and evaluate a model of edTPA-related professional development specifically related to the role of university field supervisors.

The model focused not only on each supervisor's understanding of the edTPA in terms of purpose, design, and data analysis but also included a feedback process that field supervisors used to support teacher candidates throughout the edTPA process. Follow-up data showed that teacher candidates had a more positive attitude toward their supervisors and felt more prepared to complete the edTPA following this professional development work. In a related study, Steadman and Dobson (2018) reported that a series of professional development meetings focused on implementation of the edTPA provided a new context for collaboration and joint planning for field supervisors that led to more consistency in their support for candidates and a stronger sense of shared identity as a "community of practice." This finding is congruent with reports from faculty in the Peck et al. (2010) study, including the finding that for some faculty and supervisors who had previously felt marginalized in program discussions and decision-making, participating in shared professional development activities related to PACT implementation increased both their sense of belonging and their investment in the program.

Leadership: Orchestrating TPA Implementation

Evidence from both detailed case studies and larger scale cross-program investigations of TPA implementation reflect the crucial role that leadership plays in managing the challenges and accessing the opportunities for learning and program improvement that attend TPA implementation (De Voto et al., 2020; Lys et al., 2014; Peck et al., 2010; Sloan, 2013). For example, De Voto et al. (De Voto & Thomas, 2020; De Voto et al., 2020) found that where academic leaders "set the tone" and "urged faculty and staff to look beyond the TPA mandate and focus on its framework as a tool for inquiry, program improvement, and redesign" (De Voto et al., 2020, p. 8), they were more likely to actively engage the edTPA as an opportunity for learning and program improvement. Conversely, where leaders were less actively engaged in supporting sense-making around the edTPA, there was evidence of cosmetic compliance and active resistance (De Voto & Thomas, 2020; De Voto et al., 2020; Ledwell & Oyler, 2016).

The case study reported by Peck et al. (2010) provides a more concrete example of how program leaders facilitate the kind of sense-making that can support faculty engagement with inquiry-oriented TPA implementation. In this program, academic leaders led a series of discussions in which program faculty reflected on their own values and goals for the program. Faculty then used their list of "valued outcomes" as an analytic framework for decomposing the requirements of the PACT process. These deliberations allowed program members to identify specific areas where the tool was aligned (and misaligned) with local values and identify opportunities to use the tool in ways that supported their own program goals. This general leadership strategy is also evident in other case study reports of TPA implementation (Lachuk & Koellner, 2015; Torgerson et al., 2009).

Sloan (2013) observed that leadership responsibilities in the TPA implementation process were not restricted to program members in formal administrative positions but were distributed across virtually all program members in ways that were related to their specific roles. For example, Sloan describes how field supervisors took on important leadership functions in creating new tools that assisted both candidates and mentor

teachers in developing a clearer understanding of the goals and requirements of PACT and ensuring that planning and communication were well supported as candidates completed the assessment process. The importance of distributing leadership roles and functions widely across program members was also evident in a retrospective analysis of 6 years of edTPA implementation work completed at a large teacher preparation institution (Cuthrell et al., 2019). These authors emphasized the importance of carefully preparing and supporting program members in assuming leadership roles in the TPA implementation process.

SUMMARY REFLECTIONS AND RECOMMENDATIONS

Almost two decades ago, Shulman (2005) commented on the status of teacher education as a field in which “a thousand flowers” were allowed to bloom, a field perhaps charming in its diversity but virtually impossible to cultivate. The enormous body of work that has ensued since that time has moved the field a considerable distance toward establishing a professional consensus on standards articulating what we mean by “teaching quality” (NBPTS, InTASC), along with a variety of performance assessment tools aimed at evaluating and improving teaching practice based on those standards. However, the progress that has been made co-exists with ongoing philosophical and political disputation about the purposes of education, accompanied by related disputes regarding choice of theoretical tools and relevant empirical evidence (Cochran-Smith & Fries, 2001; Wilson & Youngs, 2005).

To these ongoing challenges we add the observation of many scholars, and perhaps almost all practitioners, that teaching is at its core inescapably contextual in nature (Grossman et al., 2009). Good teaching is—perhaps to an uncomfortable extent for some policymakers—contingent on a teacher’s appraisal and response to the demands and affordances of specific situations, including the complex interplay of individual student characteristics and needs, the classroom as a social community with its own history and culture, state and district policy, and (to be sure) the repertoire of knowledge and skill of the teacher. Our review suggests that TPAs can be extremely useful resources for evaluating and improving teacher education programs. However, like other aspects of teaching, they are situated in complex systems of professional activity in which skilled and knowledgeable judgment is fundamental to making good decisions about how they are used. Moss (2013) comments on this in a particularly informative way:

If the goal is to make decisions about how to improve teaching and learning or to make choices among alternative courses of action or policies, evidence of student outcomes alone is insufficient; one must consider information about the conceptual and material resources, the teaching processes and practices, and the organizational routines and cultures that shape or influence those outcomes. (p. 93)

With this contextualist view in mind, we offer some pragmatic suggestions for TPA-related policy, practice, and research based on our review. We believe careful and strategic attention to these concerns is crucial to making TPAs function effectively as resources for meaningful program improvement.

State Policy

Our review suggests several actions that might be undertaken at the state policy level to support implementation and use of TPAs for program evaluation and improvement.

Increase Program Accountability for the Quality of TPA Implementation

Like others, we recommend that TPAs be considered in conjunction with other measures of teaching competence and performance and used within deliberative processes leading to licensure recommendations at the local program level (Darling-Hammond, 2010; Moss, 2013; Pecheone & Chung, 2006; Reagan et al., 2016). However, we also suggest that this moderated approach to the uses of standardized TPAs be balanced with more strategic and effective accountability policies related to the quality of programmatic support for candidates in completing the TPAs, particularly if they are required for licensure. We are *not* suggesting accountability policies based on program rankings or statewide benchmarks related to TPA pass rates (Reagan et al., 2016). Rather, we are suggesting that programs be expected to measure, evaluate, and, as necessary, improve their supports for candidate preparation related to the performance expectations of any program- or state-level TPA, and that evidence of program-level commitments to continuous monitoring and improvement of TPA preparation practices be incorporated into accreditation reviews and public reports.

This seems particularly important in the context of the considerable evidence suggesting that program supports for candidates related to preparation for TPAs vary widely (Cohen et al., 2020; De Voto et al., 2020) and sometimes reflect intentional faculty disengagement with efforts to prepare candidates to succeed with TPAs (Ledwell & Oyler, 2016; Ratner & Kolman, 2016). While faculty deflection of accountability policies, such as those underlying many contemporary TPA-related initiatives, may be grounded in a thoughtfully considered set of professional and political values (e.g., Kornfeld et al., 2007), it seems reasonable that candidates be apprised of this faculty stance before investing in a program that they expect to prepare them for licensure.

A useful example of program-focused accountability policies related to TPA implementation has been developed in California, in which clear expectations for practicum site selection, candidate support, and program-level analysis of TPA data for the purposes of program improvement are explicit.³ Additionally, Kim and Sato (2019) have developed a set of survey tools that are aligned with questions about the quality of TPA implementation that the research literature suggests are important to monitor and evaluate. Resources such as these could be useful tools in increasing program accountability for the quality of TPA support candidates receive.

Increase State-Level Supports for Cross-Program Learning and Improvement

Several of the studies we reviewed suggested the value of state-level supports for cross-program collaboration, learning, and improvement. These “capacity-building” policies include creation and support of standing state-level collaborative committees

³ See https://www.ctc.ca.gov/docs/default-source/educator-prep/standards/prelimmsstandard-pdf.pdf?sfvrsn=a35b06c_4.

aimed at supporting TPA implementation (De Voto et al., 2020), state-sponsored TPA trainings (Warner et al., 2020), and state conferences sponsored in collaboration with AACTE-affiliated teacher education program networks (Peck et al., 2012).

One of the important affordances of having a common and standardized TPA is the opportunity it represents for state or national level consortia to identify shared needs and undertake collaborative actions for program improvement. For example, SCALE has recently convened a group of more than 50 teacher education programs from around the United States that have been highly successful in preparing their candidates of color for the edTPA over the past 5 years (Pecheone, personal communication, March 11, 2021). These programs are working collectively to identify common themes underlying their successes and develop a set of policy/practice briefs to share what they have learned with other programs. The opportunity to undertake large scale cross-program analyses of this type is predicated on having common measures and a common language that allows practitioners and researchers from diverse institutional and state policy contexts to learn from one another. For a field that has been characterized by its struggles to build a shared and cumulative knowledge base (Sleeter, 2014; Wideen et al., 1998; Zeichner, 2007), this opportunity seems important.

Program Practices

In addition to state-level policies, our review suggests several program-level practices that appear likely to support the implementation of a TPA and enhance its value as a resource for local learning and program improvement.

Develop a Leadership Plan

Strategic leadership is critical to making TPAs useful and used for program improvement. Leadership strategies should include plans for supporting program-level deliberation and sense-making related to examining the TPA itself and considering the purposes and possibilities it may afford for achieving local program improvement. Leadership planning should also clarify strategies for distributing leadership opportunities and related responsibilities broadly across the program (Sloan, 2013) and for balancing the voices of program members that are engaged in coursework and fieldwork (Peck et al., 2010; Steadman & Dobson, 2018).

Invest in Professional Development

A consistent finding from case studies of TPA implementation is the importance of professional development opportunities for faculty, staff, and P-12 colleagues. A variety of resources exist to support informational aspects of training related to the content and process of the assessments, particularly for widely used assessments such as the edTPA (National Education Association, 2021). However, case reports also underscore the critical role that opportunities for collaborative discussion, deliberation, and problem-solving play not just in learning about the tools, but in building a shared vision and commitment to using and learning from the assessments (Cuthrell et al., 2019; Lachuk & Koellner, 2015; Sloan, 2013).

Some of the most powerful professional development experiences reported in these case studies refer to joint work activities undertaken by faculty, field supervisors, and mentor teachers. For example, several studies suggest the value of local program members scoring (at least a portion of) TPAs completed by their own candidates (Bastian et al., 2016; Cuthrell et al., 2019; Peck et al., 2010; Sloan, 2013). One important value of direct participation in scoring lies in the process of calibration, in which program faculty, supervisors, and cooperating teachers can begin to build a common understanding of the affordances (and limitations) of a TPA and a common language for describing candidate teaching practice.

Disaggregate TPA Data

Aggregated TPA scores can be useful as a beginning point for identifying general areas of program strength and weakness. However, further decomposition and analysis of scores (Bastian et al., 2018) and/or systematic qualitative analysis of TPA artifacts is likely to be particularly useful for identifying program-level curriculum and instructional variables related to improvement goals (Bunch et al., 2009; Peck et al., 2010; Sandoval et al., 2020; Sloan, 2013). Several studies suggest that opportunities for collaborative examination of concrete artifacts of teaching practice across a sample of TPA portfolios may ameliorate status differences between university faculty and field-based teacher educators, as field supervisors and mentor teachers are generally well-positioned to interpret assessment findings because of their more intimate knowledge of classroom conditions affecting candidate teaching (Cuthrell et al., 2018; Peck et al., 2010; Sloan, 2013; Whittaker & Nelson, 2013).

Create Specific Organizational Supports for TPA Work

Making a TPA function effectively as a resource for program improvement requires strategic investment of time to support the work. Case studies of TPA implementation suggest that integration of focused TPA implementation planning into regularly scheduled administrative and programmatic meetings as well as the creation of special events such as retreats or “Data Days” are both useful strategies. The success of either requires careful planning and preparation and is contingent on strong and proactive program leadership (Kroeger, 2019; Lys et al., 2014; Sloan et al., 2021).

Research Priorities

The research base related to TPAs is extremely fragmented. Psychometric studies evaluating the reliability and validity of these measures have been useful for identifying important questions about these measures and how they are used (Bastian et al., 2016; Gitomer et al., 2019; Goldhaber et al., 2017; Stanford Center for Assessment, Learning, and Equity, 2015). However, with notable exceptions (e.g., Cohen et al., 2020; De Voto et al., 2020; Kim & Sato, 2019), implementation studies have been limited in both scope and method. Based on this general appraisal of the status of TPA-related research, we suggest several priorities for inquiry.

TPAs and P-12 Student Achievement

Investment in TPAs is predicated on the assumption that variability in teaching performance as measured by these tools is related to learning outcomes for P-12 students. The evidence on this point is inconclusive but promising (Bastian et al., 2016; Chen et al., 2021; Goldhaber et al., 2017; National Research Council, 2008; Stanford Center for Assessment, Learning, and Equity, 2015). Evidence related to the effects of teacher education program improvement interventions on the learning of P-12 students is even more limited. Gansle et al. (2015) report one well-documented example of the kind of research linking program improvement actions to P-12 student outcomes that would be helpful. While the notion of “using data for program improvement” represents an intuitively appealing and highly rational theory of action, the fact remains that the field is in serious need of research that clarifies the conditions under which faith in this theory is justified.

Equity Issues

Existing research has raised important questions about differences in TPA performance between historically marginalized racial and cultural groups and their White peers (e.g., Goldhaber et al., 2017). However, results of these studies have been inconsistent. There is a clear and pressing need to replicate and extend research in this area. The troubling findings from Williams et al. (2019) suggest the importance of undertaking more process-oriented studies of program-level policies and practices that may contribute to TPA performance of minoritized teacher candidates.

Several efforts to engage equity concerns about TPAs have also been undertaken through collaborations among teacher education practitioners. For example, as described above, SCALE has recently assembled a cadre of programs that have either consistently achieved high edTPA scores with candidates of color or substantial improvements on these scores over time (Pecheone, personal communication, March 11, 2021). The members of this group are currently working to identify innovations, as well as shared themes in practice, that may contribute to improvements in preparation of candidates of color to succeed with these assessments at the same levels as White peers. Other efforts to engage equity issues related to TPAs have focused on critical analysis of the instruments themselves, as well as improving procedures for analyzing the data they produce (Escalante et al., 2021; Stillman et al., 2013). Taken together, these reports suggest that equity-related concerns about TPAs, like other standardized assessments in education, have become a focal point for both critique and improvement of the both the instruments themselves and the ways they are used in evaluation of preservice teacher quality.

Learning Processes

Despite the rationale and promise of performance assessments as resources for candidate and program-level learning and improvement (Darling-Hammond & Snyder, 2000; Haertel, 1991; Pecheone & Chung, 2006), we need to know much more about the conditions under which these learning processes take place. The contexts and condi-

tions under which teacher candidates learn and improve their practice is an important locus of this research (Chung, 2008; Lin, 2015). We also need to know more about faculty learning, including how faculty learning at the individual level is related to collective learning and program-level decision-making (Peck et al., 2009), and how these processes themselves may be mediated by professional networks of affiliation and influence (Cohen et al., 2020).

Intervention Studies

The research on TPA implementation is almost entirely descriptive and retrospective in nature. Some of these studies have produced rich descriptions of promising clinical practice (e.g., Alloway & Lesh, 2019) that merit further evaluation using more prospective, intervention-oriented methodologies. The case-based intra-subject experimental designs used in clinical research fields such as psychology and special education may be particularly useful in evaluation of specific program practices (Perdices & Tate, 2009; Sidman, 1960). Intra-subject research methods support focused evaluation of specific changes in local practice, using pre-intervention measures in each individual case as a basis for comparison with post-intervention outcomes aimed at program improvement goals. While systematic analysis and improvement of local practice is the first goal of intra-subject research methodology, systematic replication of these kinds of studies, even with small numbers of participants in each, can provide a basis for developing and evaluating innovations in practice that are useful across programs.

Formative intervention methodologies (Penuel, 2014) such as design-based implementation research (Fishman et al., 2013), developmental work research (Sannino et al., 2016), and improvement science (Bryk et al., 2015) are particularly useful for studying and improving processes of learning, negotiation, and organizational change at the program and institutional level. A distinguishing feature of these methods is their focus on *collaborative processes* of problem identification, data collection, and action related to change—collaborations that are essential to making TPA data matter for program improvement. Both intra-subject and formative intervention research methodologies share an additional feature, which is that their focus on evaluation of program improvement actions with small sample sizes allows these methods to be “locally owned and operated” in a way that is consonant with faculty values around local program authority and autonomy.

CONCLUSION

In an early review of the present paper, a colleague asked us “so ... do you think that TPAs are a good thing or a bad thing?” Our response to this provocative question is that TPAs are *tools*; like any tool, their value very much depends on how they are used and what they are used for. Our review suggests that the kinds of rich and contextualized data that TPAs produce can create valuable opportunities for learning, innovation, and program improvement (Denner et al., 2009; Lys et al., 2014; Sloan, 2013; Peck et al., 2009). However, taking up these opportunities requires strategic organizational supports for learning that are often weak or missing in both state policy and local program

routines and practice. Making TPAs more useful and more regularly *used* as resources for program evaluation, learning, and improvement will require not only the continuous improvement of the tools themselves, but more focused and strategic attention to improving the organizational conditions in which they are implemented.

REFERENCES

- Adie, L., & Wyatt-Smith, C. (2020). Fidelity of summative performance assessment in initial teacher education: The intersection of standardisation and authenticity. *Asia-Pacific Journal of Teacher Education*, 48(3), 267-286.
- Alloway, M., & Lesh, J. (2019). Compliance, affirmation, alignment, or learning? Influencing the edTPA process through one's stance. In L. Barron (Ed.), *A practical guide for edTPA implementation* (pp. 25-45). Information Age Publishing.
- Anderson, R. S., & DeMeulle, L. (1998). Portfolio use in twenty-four teacher education programs. *Teacher Education Quarterly*, 25(1), 23-31.
- Athanases, S. Z. (1994). Teachers' reports of the effects of preparing portfolios of literacy instruction. *The Elementary School Journal*, 94(4), 421-439.
- Bacon, J., & Blachman, S. (2017). A disability studies in education analysis of the edTPA through teacher candidate perspectives. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, 40(4), 278-286.
- Bales, B. (2006). Teacher education policies in the United States: The accountability shift since 1980. *Teaching and Teacher Education*, 22, 395-407.
- Bastian, K. C., Henry, G. T., Pan, Y., & Lys, D. (2016). Teacher candidate performance assessments: Local scoring and implications for teacher preparation program improvement. *Teaching and Teacher Education*, 59, 1-12.
- Bastian, K. C., Lys, D., & Pan, Y. (2018). A framework for improvement: Analyzing performance-assessment scores for evidence-based teacher preparation program reforms. *Journal of Teacher Education*, 69(5), 448-462.
- Bastian, K. C., Lys, D., & Whisenant, W. R. L. (2020). *Student teaching environments, performance assessment scores, and the tradeoffs between authenticity and standardization in teacher candidate assessments* [Unpublished manuscript].
- Behizadeh, N., & Neely, A. (2019). Testing injustice: Examining the consequential validity of edTPA. *Equity & Excellence in Education*, 51(3-4), 242-264.
- Bloxham, S., Huges, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment and Evaluation in Education*, 41(4), 638-653.
- Boreham, N. (2004). A theory of collective competence: Challenging the neo-liberal individualization of performance at work. *British Journal of Educational Studies*, 52(1), 5-17.
- Borko, H., Michalec, P., Timmons, M., & Siddle, J. (1997). Student teaching portfolios: A tool for promoting reflective practice. *Journal of Teacher Education*, 48(5), 345-357.
- Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.
- Buchmann, M., & Floden, R. E. (1992). Coherence, the rebel angel. *Educational Researcher*, 21(9), 6-9.
- Bullough, R. V. (2010). Proceed with caution: Interactive rules and teacher work sample scoring strategies, an ethnomethodological study. *Teachers College Record*, 112(3), 775-810.
- Bunch, G. C., Aguirre, J. M., & Telléz, K. (2009). Using candidate responses on high stakes performance assessment to inform teacher preparation for English learners. *Issues in Teacher Education*, 18(1), 26.
- Campbell, C., Ayala, C. C., Railsback, G., Freking, F. W., McKenna, C., & Lausch, D. (2016). Beginning teachers' perceptions of the California Teaching Performance Assessment (TPA). *Teacher Education Quarterly*, 43(2), 51-71.
- Campbell, K. E., Oliver, D. E., & Shortliffe, E. H. (1998). The unified medical language system: Toward a collaborative approach for solving terminologic problems. *Journal of the American Medical Informatics Association*, 5(1), 12-16.

- Chen, B., Cowan, J. Goldhaber, D., & Theobald, R. (2021). *From the clinical experience to the classroom: Assessing the predictive validity of the Massachusetts Candidate Assessment of Performance*. National Center for Analysis of Longitudinal Data in Educational Research.
- Choppin, J., & Meuwissen, K. (2017). Threats to validity in the edTPA video component. *Action in Teacher Education, 39*(1), 39-53.
- Chung, R. R. (2008). Beyond assessment: Performance assessments in teacher education. *Teacher Education Quarterly, 35*(1), 7-28.
- Chye, S., Zhou, M., Koh, C., & Liu, W. C. (2019). Using e-portfolios to facilitate reflection: Insights from an activity theoretical analysis. *Teaching and Teacher Education, 85*, 24-35.
- Clark-Gareca, B. (2015). Where practicum meets test preparation: Supporting teacher candidates through EdTPA. *CATESOL Journal, 27*(2), 211-220.
- Clayton, C. D. (2018a). Policy meets practice in New York state: Understanding early edTPA implementation through preservice candidates' eyes. *Teacher Education Quarterly, 45*(3), 97-125.
- Clayton, C. D. (2018b). Voices from student teachers in New York: The persistence of a subtractive experience of the edTPA as a licensure exam for initial certification. *Education Policy Analysis Archives, 26*(27).
- Coburn, C. E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision-making in the district central office. *Teachers College Record, 111*(4), 1115-1161.
- Cochran-Smith, M., Carney, M. C., Keefe, E. S., Burton, S., Chang, W.-C., Fernández, M. B., Miller, A. F., Sánchez, J. G., & Baker, M. (2018). *Reclaiming accountability in teacher education*. Teachers College Press.
- Cochran-Smith, M., & Fries, M. K. (2001). Sticks, stones, and ideology: The discourse of reform in teacher education. *Educational Researcher, 30*(8), 3-15.
- Cochran-Smith, M., Piazza, P., & Power, C. (2013). The politics of accountability: Assessing teacher education in the United States. *The Educational Forum, 77*(1), 6-27.
- Cohen, J., Hutt, E., Berlin, R. L., Mathews, H. M., McGraw, J. P., & Gottlieb, J. (2020). Sense making and professional identity in the implementation of edTPA. *Journal of Teacher Education, 71*(1), 9-23.
- Conant, J. B. (1963). *The education of American teachers*. McGraw-Hill.
- Cooner, D., Stevenson, C., & Frederiksen, H. (2011). Teacher work sample methodology: Displaying accountability of U.S. teacher education program effectiveness. *Journal of College Teaching & Learning (TLC), 8*(10), 17.
- Cronenberg, S., Harrison, D., Korson, S., Jones, A., Murray-Everett, N. C., Parrish, M., & Johnston-Parsons, M. (2016). Trouble with the edTPA: Lessons learned from a narrative self-study. *Journal of Inquiry and Action in Education, 8*(1), 109-134.
- Crowe, E. (2011). *Race to the Top and teacher preparation: Analyzing state policies for ensuring real accountability and fostering program innovation*. Center for American Progress.
- Cuthrell, K., Fogarty, E., Lys, D., Stapleton, J., & Tschida, C. (2019). 1,800 teacher candidates later: Lessons learned with large program implementation. In L. Barron (Ed.), *A practical guide for edTPA implementation: Lessons from the field* (pp. 115-130). Information Age Publishing.
- Darling-Hammond, L. (1986). A proposal for evaluation in the teaching profession. *The Elementary School Journal, 86*(4), 531-551.
- Darling-Hammond, L. (2006, March/April). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*(2), 120-138.
- Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching* (pp. 1-27). Center for American Progress.
- Darling-Hammond, L., Newton, S. P., & Wei, R. C. (2013, April). Developing and assessing beginning teacher effectiveness: The potential of performance assessment. *Educational Assessment, Evaluation and Accountability, 25*(1), 179-204.
- Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5-6), 523-545.
- Davey, T., Ferrara, S., Holland, P., Shavelson, R., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessments*. Educational Testing Service.
- Davis, S. C., & Peck, C. A. (2020). Using data for program improvement in teacher education: A study of promising practices. *Teachers College Record, 122*(3), 1-48.
- De Voto, C., Olson, J. D., & Gottlieb, J. J. (2020). Examining diverse perspectives of edTPA policy implementation across states: The good, the bad, and the ugly. *Journal of Teacher Education, 72*(1), 1-14.

- De Voto, C., & Thomas, M. K. (2020). Cultural sensemaking and the implementation of edTPA technological tools: Lessons for the field. *Educational Technology Research and Development*, 68, 2729-2751.
- Delandshere, G., & Arens, S. A. (2003). Examining the quality of the evidence in preservice teacher portfolios. *Journal of Teacher Education*, 54(1), 57-73.
- Delandshere, G., & Petrosky, A. R. (2010). The use of portfolios in preservice teacher education: A critical appraisal. In M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 9-42). Wiley.
- Denner, P., Norman, A., & Lin, S. (2009). Fairness and consequential validity of Teacher Work Samples. *Educational Assessment, Evaluation, and Accountability*, 21(3), 235-254.
- Denner, P., Pankratz, R., Norman, T., & Newsome, J. (2004). *Building credibility into performance assessment and accountability systems for teacher preparation programs: A "how to" manual for teacher educators who want to collect, use and report valid and reliable performance data on teacher candidates with a link to P-12 student learning*. Renaissance Partnership for Improving Teacher Quality.
- Dover, A. G. (2018). Your compliance will not protect you: Agency and accountability in urban teacher preparation. *Urban Education*. <https://doi.org/10.1177/0042085918795020>.
- Educational Testing Service. (2020). *Glossary of standardized testing terms*. https://www.ets.org/understanding_testing/glossary.
- Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research* (1st ed.). Cambridge University Press.
- Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, 14(1), 133-156.
- Escalante, K., Ervin-Kassab, L., & Soodjinda, D. (2021, April). Re-envisioning teaching performance assessments: How the CalTPA can confront inequities through an anti-racist lens. Paper presented at the Annual Meeting of the American Educational Research Association.
- Fishman, B., Penuel, W., Allen, A. R., Cheng, B., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *Teachers College Record*, 115(14), 136-156.
- Foot, K. A., & Groleau, C. (2011). Contradictions, transitions, and materiality in organizing processes: An activity theory perspective. *First Monday*, 16(6).
- Gansle, K. A., Noell, G. H., Grandstaff-Beckers, G., Stringer, A., Roberts, N., & Burns, J. M. (2015). Value-added assessment of teacher preparation: Implications for special education. *Intervention in School and Clinic*, 51(2), 106-111.
- Gay, G. (2001). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.
- Gellman, E. S. (1993). The use of portfolios in assessing teacher competence: Measurement issues. *Action in Teacher Education*, 14(4), 39-44.
- Gitomer, D., Brown, T., & Bonett, J. (2011). Useful signal or unnecessary obstacle? The role of basic skills tests in teacher preparation. *Journal of Teacher Education*, 62(5), 431-445.
- Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2019). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, 58(1), 3-31.
- Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, 68(4), 377-393.
- Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal*, 47(1), 218-251.
- Goodlad, J. I. (1990). *Teachers for our nation's schools*. Jossey-Bass.
- Govaerts, M. J., Van der Vleuten, C. P., Schuwirth, L. W., & Muijtjens, A. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education*, 12(2), 239-260.
- Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9), 2055-2100.
- Grossman, P., Hammerness, K. M., McDonald, M., & Ronfeldt, M. (2008). Constructing coherence: Structural predictors of perceptions of coherence in NYC teacher education programs. *Journal of Teacher Education*, 59(4), 273-287.
- Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education*, 17, 3-29.

- Hall, R., & Horn, I. S. (2012). Talk and conceptual change at work: Adequate representation and epistemic stance in a comparative analysis of statistical consulting and teacher workgroups. *Mind, Culture, and Activity, 19*(3), 240-258.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.
- Khatti, N., Reeve, A. L., & Kane, M. B. (1998). *Principles and practices of performance assessment*. Routledge.
- Kim, S. J., & Sato, M. (2019). A consequential validity framework for performance assessment implementation based on a statewide adoption of edTPA. In L. Barron (Ed.), *A practical guide for edTPA implementation: Lessons from the field* (pp. 3-23). Information Age Publishing.
- Kohler, F., Henning, J. E., & Usma-Wilches, J. (2008). Preparing preservice teachers to make instructional decisions: An examination of data from the teacher work sample. *Teaching and Teacher Education, 24*(8), 2108-2117.
- Kornfeld, J., Grady, K., Marker, P., & Ruddell, M. (2007). Caught in the current: A self-study of state-mandated compliance in a teacher education program. *Teachers College Record, 109*(8), 1902-1930.
- Kroeger, L. (2019). Going beyond rubric scores: Meaningful use of edTPA data. In L. Barron (Ed.), *A practical guide for edTPA implementation: Lessons from the field* (pp. 365-376). Information Age Publishing.
- Lachuk, A. J., & Koellner, K. (2015). Performance-based assessment for certification: Insights from edTPA implementation. *Language Arts, 93*(2), 84-95.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
- Ledwell, K., & Oyler, C. (2016). Unstandardized responses to a “standardized” test: The edTPA as gatekeeper and curriculum change agent. *Journal of Teacher Education, 67*(2), 120-134.
- Levine, A. (2006). *Educating school teachers*. The Education Schools Project.
- Lin, S. (2015). *Learning through action: Teacher candidates and performance assessments* [Unpublished doctoral dissertation]. University of Washington
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.
- Lit, I. W., & Lotan, R. (2013). A balancing act: Dilemmas of implementing a high-stakes performance assessment. *The New Educator, 9*(1), 54-76.
- Lys, D. B., L'Esperance, M., Bullock, A., Dobson, E., Patriarca, L. A., & Maynard, E. E. (2016). Data and dialogue: Cultivating transformative change in teacher preparation programs. In D. Polly (Ed.), *Evaluating teacher education programs through performance-based assessments* (pp. 34-51). IGI Global.
- Lys, D. B., L'Esperance, M., Dobson, E., & Bullock, A. A. (2014). Large-scale implementation of the edTPA: Reflections upon institutional change in action. *Current Issues in Education, 17*(3).
- McConney, A. A., Schalock, M. D., & Del Schalock, H. (1998). Focusing improvement and quality assurance: Work samples as authentic performance measures of prospective teachers' effectiveness. *Journal of Personnel Evaluation in Education, 11*(4), 343-363.
- McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis, 9*(2), 171-178.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.
- Meyer, S. J., Espel, E. V., & Nelson, N. J. (2018). State education agency use of teacher candidate performance assessments. In J. E. Many & R. Bhatnagar (Eds.), *Implementing and analyzing performance assessments in teacher education* (pp. 191-216). Information Age Publishing.
- Milner, H. R., & Tenore, F. B. (2010). Classroom management in diverse classrooms. *Urban Education, 45*(5), 560-603.
- Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.
- Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement, 50*(1), 91-98.
- Moss, P. A., Schutz, A. M., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education, 12*(2), 139-161.
- National Association for Multicultural Education. (2014). *NAME position statement on the edTPA*. <https://www.nameorg.org/docs/Statement-rr-edTPA-1-21-14.pdf>.

- National Board for Professional Teaching Standards. (2021a). *Teacher career continuum*. <https://www.nbpts.org/teacher-career-continuum>.
- National Board for Professional Teaching Standards. (2021b). *Vision*. <https://www.nbpts.org/vision-and-impact/vision>.
- National Education Association. (2021). *edTPA*. <https://www.nea.org/professional-excellence/professional-learning/teacher-licensure/edtpa>.
- National Research Council. (2008). *Assessing accomplished teaching: Advanced-level certification programs*. The National Academies Press. <https://doi.org/10.17226/12224>.
- Nicolini, D., Mendis, J., & Swan, J. (2012). Understanding the role of objects in cross-disciplinary collaboration. *Organization Science*, 23(3), 612-629.
- Nolen, S. B., Horn, I. S., Ward, C. J., & Childers, S. A. (2011). Novice teacher learning and motivation across contexts: Assessment tools as boundary objects. *Cognition and Instruction*, 29(1), 88-122.
- Okhremtchouk, I., Newell, P., & Rosa, R. (2013). Assessing pre-service teachers prior to certification: Perspectives on the Performance Assessment for California Teachers (PACT). *Education Policy Analysis Archives*, 21(56).
- Olson, J. D., & Rao, A. B. (2017). Adopting the edTPA as a high-stakes assessment: Resistance, advocacy, and reflection in Illinois. *Mid-Western Educational Researcher*, 29(4), 377-402.
- Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education*, 57(1), 22-36.
- Peck, C. A., & Davis, S. (2019). Building capacity and commitment for data use in teacher education programs. In E. B. Mandinach & E. S. Gummer (Eds.), *Data for continuous programmatic improvement: Steps colleges of education must take to become a data culture* (pp. 68-92). Routledge.
- Peck, C. A., Gallucci, C., & Sloan, T. (2010). Negotiating implementation of high-stakes performance assessment policies in teacher education: From compliance to inquiry. *Journal of Teacher Education*, 61(5), 451-463.
- Peck, C. A., Gallucci, C., Sloan, T., & Lippincott, A. (2009). Organizational learning and program renewal in teacher education: A socio-cultural theory of learning, innovation and change. *Educational Research Review*, 4(1), 16-25.
- Peck, C., Muzzo, M., & Sexton, P. (2012). *Program implementation of the edTPA in Washington State* [Unpublished manuscript].
- Penuel, W. R. (2014). Emerging forms of formative intervention research in education. *Mind, Culture, and Activity*, 21(2), 97-117.
- Perdices, M., & Tate, R. L. (2009). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognized and undervalued? *Neuropsychological Rehabilitation*, 19(6), 904-927.
- Petchauer, E., Bowe, A. G., & Wilson, J. (2018). Winter is coming: Forecasting the impact of edTPA on Black teachers and teachers of color. *The Urban Review*, 50(2), 323-343.
- Peterson, K., & Lyness, S. (2015). Factors affecting the reliability of TPA-PACT. *CC News*, 26(2), 38-40.
- Pointer Mace, D., & Luebke, P. (2021). Living in a culture of evidence and outcomes: Engaging with data at Alverno College. In C. Peck, K. Cuthrell, D. Pointer Mace, T. Sloan, & D. Lys (Eds.), *Using data to improve teacher education: moving evidence into action* (pp. 142-152). Teachers College Press.
- Porter, J. M., & Jelinek, D. (2011). Evaluating inter-rater reliability of a national assessment model for teacher performance. *International Journal of Educational Policies*, 5(2), 74-87.
- Ratner, A. R., & Kolman, J. S. (2016). Breakers, benders, and obeyers: Inquiring into teacher educators' mediation of edTPA. *Education Policy Analysis Archives*, 24(35).
- Reagan, E. M., Schram, T., McCurdy, K., Chang, T.-H., & Evans, C. M. (2016). Politics of policy: Assessing the implementation, impact, and evolution of the Performance Assessment for California Teachers (PACT) and edTPA. *Education Policy Analysis Archives*, 24(9).
- Rennert-Ariev, P. (2008). The hidden curriculum of performance-based teacher education. *Teachers College Record*, 110(1), 105-138.
- Reusser, J., Butler, L., Symonds, M., Vetter, R., & Wall, T. J. (2007). An assessment system for teacher education program quality improvement. *International Journal of Educational Management*, 21(2), 105-113.
- Richmond, G., Bartell, T., Carter Andrews, D. J., & Neville, M. L. (2019). Reexamining coherence in teacher education. *Journal of Teacher Education*, 70(3), 188-191.

- Riggs, M., Verdi, M., & Arlin, P. (2009). A local evaluation of the reliability, validity, and procedural adequacy of the teacher performance assessment exam for teaching credential candidates. *Issues in Teacher Education*, 18(1), 13-38.
- Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, 66(4), 304-320.
- Sandoval, Jr., C., van Es, E. A., Campbell, S. L., & Santagata, R. (2020). Creating coherence in teacher preparation: Examining teacher candidates' conceptualizations and practices for equity. *Teacher Education Quarterly*, 47(4), 8-32.
- Sannino, A., Engeström, Y., & Lemos, M. (2016). Formative interventions for expansive learning and transformative agency. *Journal of the Learning Sciences*, 25(4), 599-633.
- Sarason, S. B. (1993). *The case for change: Rethinking the preparation of educators*. Jossey-Bass.
- Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education*, 65(5), 421-434.
- Schalock, H., Schalock, M., & Girod, G. (1997). Teacher work sample methodology as used at Western Oregon State College. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure* (pp. 15-45). Corwin Press.
- Schalock, M. D. (1998). Accountability, student learning, and the preparation and licensure of teachers: Oregon's teacher work sample methodology. *Journal of Personnel Evaluation in Education*, 12(3), 269-285.
- Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, 20(7), 2-16.
- Shulman, L. (1987a). Assessment for teaching: An initiative for the profession. *Phi Delta Kappan*, 69(1), 38-44.
- Shulman, L. (1987b). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-23.
- Shulman, L. (2005, Fall). Teacher education does not exist. *Stanford Educator*, 7.
- Shulman, L. S., & Sykes, G. (1986). *A national board for teaching?: In search of a bold standard*. Carnegie Forum on Education and the Economy, Task Force on Teaching as a Profession.
- Sidman, M. (1960). *Tactics of scientific research*. Basic Books.
- Sleeter, C. (2014). Toward teacher education research that informs policy. *Educational Researcher*, 43(3), 146-153.
- Sloan, T. (2013). Distributed leadership and organizational change: Implementation of a teaching performance measure. *The New Educator*, 9(1), 29-53.
- Sloan, T., Cuthrell, K., & Pointer Mace, D. (2021). Making time and space for data use. In C. Peck, K. Cuthrell, D. Pointer Mace, T. Sloan, & D. Lys (Eds.), *Using data to improve teacher education: Moving evidence into action* (pp. 142-152). Teachers College Press.
- Snyder, J., Lippincott, A., & Bower, D. (1998). The inherent tensions in the multiple uses of portfolios in teacher education. *Teacher Education Quarterly*, 25(1), 45-60.
- Spillane, J. P. (2000). Cognition and policy implementation: District policymakers and the reform of mathematics education. *Cognition and Instruction*, 18, 141-179.
- Spillane, J. P., & Miele, D. B. (2007). Evidence in practice: A framing of the terrain. In P. A. Moss (Ed.), *Evidence and decision-making* (pp. 46-73). Blackwell.
- Stanford Center for Assessment, Learning, and Equity. (2015). *Educative assessment and meaningful support: 2014 edTPA administrative report*. <http://edtpa.aacte.org/news-area/2014-edtpa-administrative-report-provides-new-data-on-validity-and-performance.html>.
- Steadman, S. C., & Dobson, E. E. (2018). From isolation to a community of practice: Redefining the relationship of faculty and adjunct university supervisors during the implementation of edTPA. In J. E. Many & R. Bhatnagar (Eds.), *Implementing and analyzing performance assessments in teacher education* (pp. 39-62). Information Age Publishing.
- Stewart, S., & Haynes, C. (2015). An alternative approach to standardized testing: A model that promotes racial equity and college access. *Journal of Critical Scholarship on Higher Education and Student Affairs*, 2(1), 9.
- Stewart, A. R., Scalzo, J. N., Merino, N., & Nilsen, K. (2015). Beyond the criteria: Evidence of teacher learning in a performance assessment. *Teacher Education Quarterly*, 42(3), 33-58.
- Stillman, J., Anderson, L., Arellano, A., Wong, P. L., Berta-Avila, M., Alfaro, C., & Struthers, K. (2013). Putting PACT in context and context in PACT: Teacher educators collaborating around program-specific and shared learning goals. *Teacher Education Quarterly*, 40(4), 135-157.

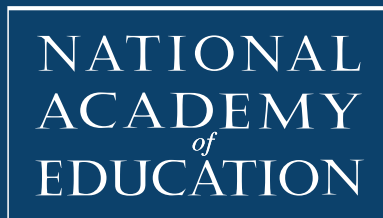
- Suleiman, R., & Byrd, C. (2016). edTPA preparation: Building support structures for teacher candidates. In D. Polly (Ed.), *Evaluating teacher education programs through performance-based assessments* (pp. 138-165). IGI Global.
- Torgerson, C. W., Macy, S. R., Beare, P., & Tanner, D. E. (2009). Fresno Assessment of Student Teachers: A teacher performance assessment that informs practice. *Issues in Teacher Education, 18*(1), 63-82.
- Tröhler, D., Meyer, H. D., Labaree, D. F., & Hutt, E. L. (2014). Accountability: Antecedents, power, and processes. *Teachers College Record, 116*(9), 1-12.
- Truijens, F. L., Cornelis, S., Desmet, M., De Smet, M. M., & Meganck, R. (2019). Validity beyond measurement: Why psychometric validity is insufficient for valid psychotherapy research. *Frontiers in psychology, 10*, 532.
- Tuck, E., & Gorlewski, J. (2016). Racist ordering, settler colonialism, and edTPA: A participatory policy analysis. *Educational Policy, 30*(1), 197-217.
- Warner, C. K., Bell, C. V., Alvarez McHatton, P., & Atilas, J. (2020). Navigating mandates and working toward coherence: Our journey with a high-stakes teacher performance assessment. *Educational Forum, 84*(3), 226-239.
- Whittaker, A., & Nelson, C. (2013). Assessment with an "end in view." *The New Educator, 9*(1), 77-93.
- Whittaker, A., Pecheone, R. L., & Stansbury, K. (2018). Fulfilling our educative mission: A response to edTPA critique. *Education Policy Analysis Archives, 26*, 30.
- Wideen, M., Mayer-Smith, J., & Moon, B. (1998). A critical analysis of the research on learning to teach: Making the case for an ecological perspective on inquiry. *Review of Educational Research, 68*(2), 130-178.
- Williams, J. A., Hart, L. C., & Algozzine, B. (2019). Perception vs. reality: EdTPA perceptions and performance for teacher candidates of color and White candidates. *Teaching and Teacher Education, 83*, 120-133.
- Wilson, S., & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education*. Routledge.
- Wolf, K. (1991). The schoolteacher's portfolio: Issues in design, implementation, and evaluation. *Phi Delta Kappan, 73*(2), 129-136.
- Zeichner, K. (2007). Accumulating knowledge across self-studies in teacher education. *Journal of Teacher Education, 58*(1), 36-46.
- Zeichner, K., Payne, K. A., & Brayko, K. (2015). Democratizing teacher education. *Journal of Teacher Education, 66*(2), 122-135.
- Zeichner, K., & Wray, S. (2001). The teaching portfolio in US teacher education programs: What we know and what we need to know. *Teaching and Teacher Education, 17*(5), 613-621.

AUTHOR BIOGRAPHIES

Charles (Cap) Peck is a professor of teacher education and special education at the University of Washington (UW). Peck began his career as an instructional aide serving children with intellectual and developmental disabilities. After completing his doctorate at the University of California (UC), Santa Barbara, Peck continued his work in special education for the next decade, focusing on research and program development related to the inclusion of students with disabilities in school and community settings. Convinced that the goals of inclusion for children with disabilities could not be achieved without significant change in general education, Peck shifted his work to focus on teacher preparation, serving as the director of teacher education at both UW and UC Santa Barbara. Peck's research in teacher education has focused on policy implementation and program improvement. His early work in this area (Peck, Gallucci, & Sloan, 2010) received commendation as the Outstanding Article in the *Journal of Teacher Education* for 2011. During Peck's tenure at UW, he served as the co-director for the *Teachers for a New Era* project sponsored by the Carnegie Foundation, where his perspectives on organizational change in teacher education were deeply influenced by his opportunity to work with John Goodlad, Bill McDiarmid, and Roger Soder. For the past decade Peck has pursued a line of research on teacher education program renewal using the theoretical and methodological tools of Cultural Historical Activity Theory. His most recent work in this area is *Using Data to Improve Teacher Education: Moving Evidence into Action* (2021), co-edited with Cuthrell, Pointer Mace, Sloan, and Lys and published by Teachers College Press.

Maia Goodman Young is a National Board Certified Teacher and a Ph.D. student at the University of Washington College of Education in the Teaching, Learning & Curriculum Program. Prior to joining the Ph.D. program, she taught middle and high school English Language Arts in California and Washington. She has worked as a mentor teacher for the Stanford Teacher Education Program and an on-site mentor for Rainier Prep Charter School. Her current research is the use of data-based intervention methodologies for program improvement in teacher education.

Wenqi Zhang is a second-year doctoral student of the Educational Policy, Organizations, and Leadership in P-12 Systems program at the University of Washington College of Education. She seeks to draw lessons from teacher preparation programs to support the understanding and implementation of principal preparation programs. Using a lens from organizational theories and policy implementation, her current research interests are about how universities, schools, and districts work together to create a coherent principal preparation program with a combination of coursework and internship. Her work employs comparative qualitative case studies.



The **National Academy of Education (NAEd)** advances high-quality research to improve education policy and practice. Founded in 1965, the NAEd consists of U.S. members and international associates who are elected on the basis of scholarship related to education. The Academy undertakes research studies to address pressing educational issues and administers professional development fellowship programs to enhance the preparation of the next generation of education scholars.

naeducation.org