The Emergence of High-Stakes Accountability Policies in Teacher Preparation: An Examination of the U.S. Department of Education's Proposed Regulations

Maria Teresa Tatto
Corey Savage
Wei Liao
Stefanie L. Marshall
Paul Goldblatt
&
Leonardo Medel Contreras
Michigan State University
United States

Abstract: Using a sociological framework this article explores the emergence and possible consequences of the 2015 U.S. Department of Education's proposed federal regulatory policy on teacher education programs and alternative route providers. After describing the key features of the policy, we examine the research literature looking for evidence of the merits of accountability policies in improving teacher education and preparation quality and outcomes. Although there is some research evidence that increased accountability measures

may indeed contribute to improving the quality and outcomes of teacher education and preparation, the conditions under which this happens are not straightforward. While the stated aim of the regulatory policy, to ultimately advance student learning, finds widespread support in the education community, research evidence points to a number of validity problems with the overall policy. Of particular concern is the policy's attempts at establishing a direct link between teacher preparation graduates' employment and pupil achievement. The policy as conceived could negatively impact program norms and resources and undermine the development of teachers' human, cultural, and social capital. We discuss the accreditation challenges that the policy is likely to confront and implications for the future of teacher education and preparation accountability.
Keywords: Teacher preparation; accountability; effectiveness; United States

El surgimiento de políticas de rendición de cuentas con consecuencias severas en la formación docente: Un estudio de las regulaciones propuestas por el Departamento de Educación EE.UU.

Resumen: Usando un marco conceptual sociológico este artículo explora el surgimiento y las posibles consecuencias de las políticas de reglamentación federal propuestas en 2015 por el Departamento de Educación de EE. UU sobre los programas de formación docente y sobre los modelos de formación alternativos. Después de describir las características claves de esas políticas, analizamos la literatura en busca de evidencia de los méritos de las políticas de rendición de cuentas en la mejora de la formación docente, la calidad de la preparación y sus resultados. Aunque existe alguna evidencia de que el aumento de las medidas de responsabilidad puede contribuir a la mejora de la calidad de la preparación docente, las condiciones en que esto sucede no son sencillas. Mientras que el objetivo declarado de la política, la mejora del aprendizaje de los estudiantes, tiene un amplio apoyo entre la comunidad educativa, las evidencias recogidas en esta investigación identifico una serie de problemas respecto a la validez de esta política. Una preocupación importante es acerca de los intentos de establecer un vínculo directo entre formación docente e indicadores de empleo y rendimiento de los alumnos. Tal como está concebida esta política podría impactar negativamente las normas y recursos del programa y socavar el desarrollo de los docentes en términos de recursos humanos, culturales, y de capital social. Discutimos los desafíos de la acreditación que probablemente esta política enfrentara y las implicaciones para el futuro de las políticas de rendición de cuentas en la mejora de la formación docente.
Palabras clave: formación docente; rendición de cuentas; eficacia; Estados Unidos

O surgimento das políticas de responsabilidade com consequências graves na formação de professores: Um estudo dos regulamentos propostos pelo Departamento de Educação dos EUA

Resumo: Usando uma base conceitual sociológica este artigo explora a ascensão e as possíveis consequências das políticas de regulação federais em 2015 propostas pelo Departamento de Educação dos Estados Unidos para os programas de formação docente e de modelos alternativos de formação. Depois de descrever as principais características dessas políticas, analisamos a literatura procurando evidências sobre o mérito das políticas de responsabilização na melhora da formação de professores, a qualidade da preparação e os resultados. Embora encontramos algumas evidências de que as medidas de prestação de contas poderiam contribuir para melhorar a qualidade, os resultados da educação e da formação de professores, as condições em que isso acontece não são simples. Embora a

meta declarada dessas políticas, a melhoria dos aprendizagem dos alunos, tem amplo apoio entre a comunidade educativa, as provas recolhidas nesta investigação identificaram uma série de problemas relacionados com a validade desta política. Uma grande preocupação é sobre as tentativas de estabelecer uma ligação direta entre a formação de professores e indicadores de emprego e desempenho dos alunos. Como esta formulada esta política poderia impactar negativamente as regras e os recursos dos programa e prejudicar o desenvolvimento dos professores em termos de recursos humanos, culturais e capital social. Discutimos os desafios de credenciamento que esta política deverá enfrentar e as implicações para o futuro das políticas de responsabilidade na melhora da formação de professores.
Palavras-chave: formação docente; prestação de contas; eficácia; Estados Unidos

The Emergence of High-Stakes Accountability Policies in Teacher Preparation: An Examination of the U.S. Department of Education's Proposed Regulations

After soliciting comments from the public, Secretary of Education Arne Duncan announced in early 2015 a federal regulatory plan for teacher education programs and other approaches to preparing teachers. These regulations are expected to be released in the early days of 2016 without any major revisions. The regulations to be implemented at the state level have been justified by arguments that teacher education programs and alternative route providers[1] are of uneven quality. The federal regulations would call for a program of periodic accreditation aligned with the Council for the Accreditation of Education Preparation (CAEP) standards, and an ongoing requirement that providers collect yearly data to allow them to demonstrate the level of success of graduates as indicated by knowledge and satisfaction at graduation, three-year employment outcomes, and pupil outcomes. In return, the states would be expected to produce ratings of teacher education and preparation programs and allocate incentives (e.g., the Teacher Education Assistance for College and Higher Education or TEACH grants) to those providers that demonstrate success according to the aforementioned indicators.

Whether or not these regulations in their current form are implemented, their introduction comes at a time of increasing criticism of higher education generally, and of teacher education programs in particular, and signals an important turn on the social perception of the teacher profession (Levine, 2006; NCTQ, 2014). The field has been responsive to these concerns. For instance, in the past 10 years, after much experimentation, dialogue, and consultation, important changes have occurred in the agencies that have traditionally accredited teacher education programs - National Council for Accreditation of Teacher Education (NCATE) and the Teacher Education Accreditation Council (TEAC) - culminating with the creation of a new agency, the Council for the Accreditation of Education Preparation (CAEP) in July 2013. CAEP was tasked with the development and implementation of new standards for teacher education programs.

---

[1] The terminology used to refer to the diversity of pathways into teaching in the U.S. has been changing. CAEP for instance uses "educator preparation providers." We prefer "teacher education programs" and "alternative route providers" to maintain an important distinction among these pathways, and use the term "teacher education and preparation" respectively to refer to these modalities.

In addition to the creation of new standards, CAEP is seeking to align with other standards, such as the revised Interstate New Teacher Assessment and Support Consortium standards (InTASC), issued first in the 1990s under the Clinton administration and designed to provide curriculum guidelines for teacher education programs. Other standards that CAEP seeks to align with include the National Board for Professional Teaching Standards (NBPTS); content area national standards, such as those issued by the National Council for Teaching Mathematics (NCTM) for mathematics education and the New Generation of Science Standards (NGSS) for science education; and state standards. Teacher education programs have responded quickly to align their programs in time for the new accreditation wave.

In this context of heightened accountability, the proposed regulations are expected to produce indicators for the 'meaningful differentiation' of teacher education and preparation programs as exceptional, effective, low-performing, or at-risk. Programs consistently producing unsuccessful teacher candidates in terms of measured outcomes will be designated as either low-performing or at-risk.

While few seem to disagree with the need to develop an effective system to improve the quality and relevance of teacher education and preparation programs there is concern that the proposed regulations have not been empirically tested and that their implementation may be more internally disruptive and costly than helpful. What is certain is that the regulations will produce more program data, but what is less certain is whether these data collection efforts will support program improvement as proponents argue. Critics argue that the cost and effort from program providers to properly respond to the regulations may have consequences such as discontinuing programs, potentially resulting in a decrease of the supply of teachers to the labor market. In addition, efforts to align with standards at the state and national levels may decrease a program's ability to be more responsive to local needs.

In this article we explore the emergence and possible consequences of the 2015 U.S. Department of Education's proposed federal regulatory policy on teacher education and preparation programs. In the following sections, we outline the framework underlying our analysis and examine the accreditation mandates as stated by the regulatory policy. Using evidence from the research and the policy analysis literature in teacher education and preparation, we then examine each of the indicators of program success as proposed in the regulations, looking for evidence of effectiveness, and discuss how this evidence may help inform regulatory policy in the future. Finally, we consider the many challenges that may be involved in implementing the regulations and discuss potential impacts, positive and negative, that may emerge from the regulations.

Framework

We use a sociological framework to analyze the proposed teacher preparation regulatory policy and its possible consequences (Portes, 2000). Key to the analysis is the tension that exists in different approaches to achieve social goals. On the one hand social goals may be achieved by "community networks capable of governing individual behavior and ensure normative compliance," such as those that have been developed by teacher education programs over the years. On the other hand, social goals may be achieved by "the deliberate application of incentives and coercive power by large organizations in particular the state" such as the strategy upon which the proposed regulations are based (Portes citing Coleman, p.12). Thus the emergence of the regulations marks a clear move toward a more

coercive system in teacher education and preparation accountability. Indeed, the performance accountability movement in education in the United States, which began in earnest with the No Child Left Behind (NCLB) Act of 2001, has expanded its reach from public schools to schools of education, with the state claiming a legitimate role in regulating the manner in which teachers are to be prepared.

From the state perspective, controlling the quality of teacher education and preparation is essential to the nation's survival. Teachers are charged with the development of human capital and their effectiveness depends in part on the extent to which they are themselves able to accumulate not only human, but also cultural and social capital. Teacher education programs placed in institutions of higher education and other alternative programs working closely with school districts and schools, administrators, and teachers, have been instrumental in developing future teachers for the nation's schools. Successful teacher education programs have developed over the years highly effective internal (within the universities and across subject areas) and external (outside of universities with schools and their districts) networks with strong norms that allow for high levels of coherence in the recruitment, preparation, and placement of future teachers. Thus while programs vary in the extent to which they have developed effective accountability systems, the "new accountability" introduced by the proposed regulation is intended to evenly alter programs' norms and social networks under the assumption that increased regulation will improve quality. This is an untested assumption.

Key Aims and Strategies in the Proposed Regulations of Teacher Preparation

The summary of the regulations provided by the U.S. Department of Education (USDOE) highlights the vision it holds for teacher education and preparation programs. The regulations are ambitious concerning the changes they expect to trigger, especially concerning the identification of low performing programs and the creation of a vast system of databases. While this is a federal regulation, the responsibility for teacher education and preparation evaluation will reside with the states and is based on the demonstration of four key indicators of outcomes. The vision, the assignment of the responsibility for implementation to the states, and the four key indicators are summarized in Table 1.

An analysis of the proposed "key indicators" reveals three conceived outcomes of teacher education and preparation. In chronological order these are: first, *graduates' knowledge and ability outcomes* to be measured via graduates' and graduates' employer surveys, and by measures of the knowledge, skills and dispositions attained by graduates at the end of their program (this in the USDOE summary table is included in the fourth bullet); second, *employment outcomes* to be measured by new teacher placement and retention rates; and third, *student learning outcomes*, likely to be measured by teacher evaluation metrics of some kind and evidence of student learning.

Table 1
U.S. Department of Education Vision for the Proposed Regulations for Teacher Education / Preparation Programs and Key Indicators

These Proposed Regulations Will:

Build on innovative state systems and progress in the field to encourage all states to develop their own meaningful systems to identify high- and low-performing teacher preparation programs across all kinds of programs, not just those based in colleges and universities.

Ask states to move away from current input-focused reporting requirements, streamline the current data requirements, incorporate more meaningful outcomes measures and improve the availability of relevant information on teacher preparation.

Reward only those programs determined to be effective or better by states with eligibility for TEACH grants, which are available to students who are planning to become teachers in a high-need field and in a low-income school, to ensure that these limited federal dollars support high-quality teacher education and preparation.

Offer transparency into the performance of teacher preparation programs, creating a feedback loop among programs and prospective teachers, employers, and the public, and empower programs with information to facilitate continuous improvement.

States would have primary responsibility and significant flexibility in designing their systems and evaluating program performance.

Key Indicators

States would report annually on the performance of each teacher preparation program, including alternative certification programs, based on indicators that include at least:

Employment outcomes: New teacher placement and three-year retention rates, including in high-need schools

Teacher and employer feedback: Surveys on the effectiveness of preparation

Student learning outcomes: Effectiveness of new teachers as demonstrated through measures of student growth, performance on state or local teacher evaluation measures that include data on student growth, or both, during their first three teaching years

Assurance of specialized accreditation, or evidence that a program produces candidates with content and pedagogical knowledge and quality clinical preparation, who have met rigorous entry and exit requirements.

Source: Improving Teacher Preparation: Building on Innovation. US Department of Education
http://www.ed.gov/teacherprep

The last or fourth bullet in the "key indicators" in Table 1 above is not an indicator per se but a requirement and refers to the need for teacher education and preparation programs to demonstrate evidence of "specialized accreditation." Specialized accreditation

according to the regulations requires among other things demonstrating evidence of performance according to the indicators described above and is to be achieved via CAEP guidelines to which these regulations are aligned "one to one" as argued and shown in the USDOE summary in Table 2 below:

Table 2
U.S. Department of Education Alignment of Proposed Regulations for Teacher Education /Preparation Programs with CAEP Standards

The key provisions of the proposed regulations align one on one with the standards set by the Council for the Accreditation of Education Preparation (CAEP)

	Proposed Regulations	CAEP
Student outcomes: Academic Gains among K-12 students	●	●
Employment outcomes: Job placement and retention, including in high-need schools	●	●
Customer satisfaction: Surveys of program graduates and their principals	●	●
Program review and accreditation based on content/pedagogical knowledge, high quality clinical proactive, and rigorous entry/exit requirements	●	●
Multiple performance levels resulting from review and accreditation	●	●
Flexibility to states and providers in developing multiple measures of performance	●	●

Source: Improving Teacher Preparation: Building on Innovation. US Department of Education http://www.ed.gov/teacherprep

        In sum, while presented as a proposal for commentary, the regulations seem to be a done deal. The vision is clearly stated; the mechanism for implementation and for the measurement of indicators is laid-out, as is the mechanism through which the regulations will be used in the accreditation of teacher education and preparation programs by CAEP. Given the reliance placed by the USDOE on standards it is important to examine them in more detail before describing the evidence from the literature.

Accreditation Standards

        Under the proposed regulations, to receive accreditation teacher education programs and alternative route providers are asked to demonstrate compliance with the CAEP Standards authorized in 2013. CAEP requires that "educator preparation providers seeking accreditation complete a self-study and host a site visit, during which site visitors determine

whether or not the provider meets CAEP standards based on evidence of candidate performance, use of data in program self-improvement, and teacher education and preparation programs capacity and commitment to quality." Because CAEP is a relatively new development, guidelines from its two precursors, the National Council for the Accreditation of Teacher NCATE and TEAC still mediate accreditation processes[2].

In order to prepare for accreditation teacher education and preparation programs must continuously collect data to document compliance in each standard. CAEP outlines five standards; here we include a brief description of each as stated by CAEP (2013):

Standard 1: Content and Pedagogical Knowledge: "The provider ensures that candidates develop a deep understanding of the critical concepts and principles of their discipline and, by completion, are able to use discipline-specific practices flexibly to advance the learning of all students toward attainment of college- and career-readiness standards."

Standard 2: Clinical Partnerships and Practice: "The provider ensures that effective partnerships and high-quality clinical practice are central to preparation so that candidates develop the knowledge, skills, and professional dispositions necessary to demonstrate positive impact on all P-12 students' learning and development."

Standard 3: Candidate Quality, Recruitment, and Selectivity: "The provider demonstrates that the quality of candidates is a continuing and purposeful part of its responsibility from recruitment, at admission, through the progression of courses and clinical experiences, and to decisions that completers are prepared to teach effectively and are recommended for certification. The provider demonstrates that development of candidate quality is the goal of educator preparation in all phases of the program. This process is ultimately determined by a program's meeting of Standard 4."

Standard 4: Program Impact: "The provider demonstrates the impact of its completers on P-12 student learning and development, classroom instruction, and schools, and the satisfaction of its completers with the relevance and effectiveness of their preparation."

Standard 5: Provider Quality Assurance and Continuous Improvement: "The provider maintains a quality assurance system comprised of valid data from multiple measures, including evidence of candidates' and completers' positive impact on P-12 student learning and development. The provider supports continuous improvement that is sustained and evidence-based, and that evaluates the effectiveness of its completers. The provider uses the results of inquiry and data collection to establish priorities, enhance program elements and capacity, and test innovations to improve completers' impact on P-12 student learning and development."

While the proposed regulations emphasize these components, the regulations also introduce language that refers to research-based evidence as criteria for quality. For instance, when outlining clinical preparation, the regulations ask that these experiences be grounded in

---

[2] On its website CAEP delineates how the accreditation process could occur for an educator preparation provider: "In completing its standards-focused self-study, a provider selects one of three pathways: Continuous Improvement (CI), Inquiry Brief (IB), or Transformation Initiative (TI). Providers with accreditation visits scheduled for January 2014 through Spring 2016 may choose to write the self-study and host the visit with (1) NCATE Standards or TEAC Quality Principles only (called legacy visits); (2) NCATE Standards or TEAC Quality Principles and CAEP's new standards (called dual accreditation); or (3) CAEP's new standards only (called CAEP pilots). Even these procedures are in a hiatus as recently AACTE declared a "crisis of confidence" toward CAEP.

research-based practices (e.g. observation and analysis of instruction, collaboration, and the use of technology, USDOE, 2014).

A complement to the CAEP standards is the 2013 [InTASC](#) Model Core Teaching Standards and Learning Progressions for Teachers 1.0, an updated version of the 1992 original standards which purport to outline "what teachers should know and be able to do to ensure every PK-12 student reaches the goal of being ready to enter college or the workforce in today's world." Under the proposed regulations, programs will be required to comply with these complex accreditation mandates, yet evidence as to whether they would increase quality is mixed.

In the next sections, we explore the evidence found in the literature regarding the evaluation of teacher education and preparation program *graduates' knowledge and ability outcomes*, *employment outcomes*, and *student learning outcomes*, all key components of the proposed regulations. We end each section with a brief discussion on implications for teacher education and preparation program evaluation.

Accountability in Teacher Education and Preparation: Searching for Evidence in the Literature

Given the importance of the cultural and societal contexts of teacher education and preparation we used U.S.-based studies to inform our analysis. We searched the literature to examine the different dimensions of the proposed regulations, whether and how similar strategies have been attempted, and with what results. The methods we used in our literature search as well as tables summarizing our sources are described in Appendices 1 and 2. We organize the review according to the research evidence we found on the outcomes that teacher education and preparation programs will be required to demonstrate, namely, graduates' knowledge and ability outcomes, employment outcomes, and student (pupils) learning outcomes.

Teacher Education and Preparation Program Graduates' Knowledge and Ability Outcomes

The goal of teacher education accreditation has shifted since the 1990s. Initially, teacher education accreditation in the United States was mostly focused on documenting inputs such as whether institutions had a clear philosophy, sufficient program resources, links with schools, and whether future teachers were exposed to academic and pedagogical content judged by teacher educators as appropriate to prepare them to teach (Bullough et al. 2003, Ingvarson, Beavis & Kleinhenz, 2007). Increasingly, as illustrated by the proposed regulations, programs are required to provide evidence of effectiveness based on graduates' learning outcomes (e.g., evaluation of content and pedagogical knowledge and surveys of satisfaction), and to document processes (e.g., evidence of implementation of quality assurance mechanisms, among others).

The development of measures at the program level to demonstrate the depth and breadth of knowledge attained by graduates is likely to build synergies and provide useful indicators of program outcomes. Yet unless enough effort is invested in developing the norms and the human and social capital that would be required for programs to have equivalent evaluations across all subjects within programs and across individual states it may be difficult to compare programs. With few notable exceptions (Tatto et al.., 2012) the field has lacked the resources to develop rigorous measures of teacher education outcomes and it

is unlikely that a sound and organized undertaking may be developed at the national level by the time the regulations are expected to be in place. However, even if these measures were in place, the benefit to the programs would still depend on the implementation capacity and the norms of learning from self-study within the programs for improvement to occur. As the literature below shows, rigorous and sustained measurement of outcomes has not been a norm in teacher preparation programs; doing so as part of the regulations may be seen as one more requirement to fulfill rather than an opportunity to learn and improve.

The measurement of outcomes of teacher education presents a mixed record. A study documenting program challenges in conducting self-study in response to NCATE requirements to track teacher candidates' development and outcomes illustrates the need to develop adequate systems of data collection, as well as the manpower and capacity to evaluate the system in line with accreditation standards (Bullough et al.., 2003). While programs have a range of ways to document evidence of progress (e.g., portfolio or analysis of similar homework across classes), the challenge is taking these strategies to scale because of the time that it takes to substantially evaluate portfolios. Even if rubrics are developed, it is not clear how to best use the data for accreditation purposes or for program improvement (Bell & Youngs, 2011).

Another method to collect program outcome data is through the use of surveys. This is a method that is suggested by the regulations as a way to collect quantitative and qualitative indicators of program impact. A problematic development is the suggestion to measure "satisfaction" of recent graduates and employers with the preparation received as an indicator of program success. The main motivation to propose this indicator comes from the multiple reports of novice teachers claiming that they do not feel adequately prepared to teach during the first years after graduation. The underlying assumption behind this process is that higher levels of satisfaction from recent graduates and employers imply higher levels of quality in the preparation regarding academic content knowledge and teaching skills from teacher education programs. However, we know that student satisfaction tends to decrease as the level of rigor in courses and requirements for graduation increase as Perlmutter (2015) argues, thus satisfaction and levels of knowledge attained would need to be considered together to develop valid measures of program quality.

One of the methods to evaluate teacher preparation programs proposed by the policy is surveying recent graduates and their employers (principals). Feuer, Floden, Chudowsky, & Ahn (2013) reviewed several sources of literature reporting on teacher education and preparation program evaluations and wrote a comprehensive description of current and proposed evaluation criteria to assess program quality. They reported that usually graduates are asked about their programs (e.g. courses taken, student teaching experiences) and how well prepared they feel to perform different aspects of their jobs, such as teaching their subjects effectively, meeting the diversity of their students' needs, and the like. They found that while surveys of satisfaction may have high face value and may provide some information about the program, the use of satisfaction as indicator of program quality alone is problematic because surveys may be subject to subjectivity and selectivity biases.

Surveys of employers, such as principals, seem to provide another source to assess what is considered quality teaching in school settings, but they fail to provide valid and reliable evidence concerning the relationship between teacher education and preparation and pupil learning.

A review of the literature summarizing results of program evaluation via satisfaction surveys that asked principals about their perception of the extent to which a teacher is

prepared to teach effectively found high correlations between principal assessments and teachers' value added scores (Coggshall, Bivona, & Reschly, 2012). Similarly, Harris & Sass (2009) performed a quantitative analysis with data from interviews with 30 principals from a Florida district and related these with student achievement data from the Stanford Achievement Test. They found low and moderate positive correlations between principal ratings and value-added scores for student tests (varying from .15 to .30). No links were established between the value-added model scores and teacher preparation program quality, or with these scores and teacher effectiveness.

Darling-Hammond (2006) evaluated program outcomes for the Stanford Teacher Education Program (STEP) by analyzing information collected during five years. Among the instruments used were surveys of both graduates and employers. They reported that employers' perception of STEP graduates was highly positive (97% of the principals ranked them with a score of five out of five on overall perception of graduates' preparation). They also found that employers were less critical of graduates' preparedness than the graduates themselves. They concluded that using only survey data is not enough to assess the effects of experience among recent teacher education graduates. This study suggests the need to use multiple measures, as proposed by the regulations.

Jacob and Lefgren (2008) used quantitative methods to analyze data from a western U.S. school district. Namely, they used demographic variables for students, several characteristics for teachers (such as age, experience, license and certification information, higher education institutions attended, among others) and survey responses for all principals in the district. They found that principals were able to effectively identify which teachers produced the largest and the smallest score gains for students in their schools. This ability, however, significantly decreased when identifying teachers with medium gains (i.e. not in the extremes of the distribution).

Crowe (2010) in his Center for American Progress report reviewed a set of 11 sources about teacher education and preparation program evaluation. In general, he proposes increasing efforts to evaluate outcomes and establishing different ways to make programs accountable for their graduates' preparation. He recommends using as indicators of quality value-added models, teacher tests, and surveys of graduates and employers, among others. These recommendations are similar to those proposed by the USDOE regulations.

Loadman, Freeman, Brookhart, Rahman & McCague (1999) report on the administration of a large scale survey to a sample of 3,940 teacher education program graduates from 14 institutions between 1990 and 1995. They considered the survey to be both a reliable and valid instrument, and a valuable tool to provide comparative information across programs. In related subsequent work, Thomas and Loadman (2001) administered the same survey to a cohort of 263 baccalaureate graduates and 171 M.Ed. graduates at a major Carnegie I research university. The authors found that, in general, graduates presented a very positive attitude toward their programs and careers. They also found that responses were more similar than different between both types of graduates across all four measures of the survey. They highlight the positive experience of administering the survey and claim that the use of graduate surveys is necessary to measure program quality. Frequent problems with graduate surveys are the low response rate and selectivity bias, as those who answer the surveys are likely to have favorable views of their experience.

Lessons from the Teacher Education Development Study in Mathematics ([TEDS-M](#)), the largest national and international effort to study the outcomes of teacher education, shows great promise and highlights some challenges. The TEDS-M study used quantitative methods to collect and analyze a diversity of data sources across 17 participating countries

including the U.S. (approximately 22,000 future teachers from 750 programs in about 500 higher education institutions). Among other instruments, the TEDS-M study surveyed teacher education programs, teacher educators, and future teachers. The study also developed assessments of the mathematics knowledge for teaching graduates of pre-service programs. The study followed rigorous procedures to ensure samples representative of the target population and acceptable response rates. The findings of the study show that it is possible to distinguish among programs whose graduates have high levels of knowledge at the time of graduation. Important program characteristics associated with expected outcomes include programs' entry requirements, opportunities to learn before and during the program, graduation requirements, and strong systems of quality assurance (Tatto et al.., 2012). The study also revealed that, in general, programs did not have in place a system that could provide the basic information to evaluate their performance on a regular and long-term basis. The study required that programs collect the needed data, and many did so for the very first time. An important finding from the TEDS-M study is the wide variation in outcomes, which reflect the variation in curriculum and program design within and across the participating countries.

Implications. The existing research provides mixed evidence emerging from the evaluation of teacher education outcomes. Some studies have provided valid and reliable results (e.g. Loadman et al., 2010; Tatto et al., 2012), while others show important limitations (e.g. Jacob & Lefgren, 2008). Among the most salient concern is the considerable amount of resources (such as time, people, and money) that would be required to implement longitudinal surveys of program graduates and whether these surveys would provide valid and reliable indicators of program performance (e.g., Feuer et al., 2013). The use of survey results is a source of concern because of their potential for subjectivity and selectivity biases, and for their misuse (e.g., drawing causal conclusions about program's impact such as a high levels of satisfaction from graduates and employers implying a high level of quality), especially when high stakes are in place for teacher education and preparation programs.

If survey studies are well designed and responsibly used, they offer a plausible way to engage in fair program evaluation. The TEDS-M study's methodology, for instance, provides an excellent model that includes both questionnaires and knowledge assessments of graduates to draw valid conclusions concerning program outcomes. These conclusions would be strengthened if teacher knowledge assessments were to be linked with novice teachers' practices, and with how these practices in turn support learning.

Indeed, absent from the outcomes section of the regulations that concerns graduates' knowledge and ability is evidence that programs are successfully preparing future teachers to teach challenging curriculum in challenging contexts. The regulations do invoke teacher evaluations; however as currently constructed, these have a different purpose which does not align with the purposes of teacher education and preparation evaluation (AERA & NCME, 2014). Unfortunately, the proposed policy document lacks the level of detail or examples that would allow us to have even a minimal idea of how all of these complex components in learning to teach will be considered in the evaluation of programs.

In sum, it is crucial to design efficient yet rigorous measures and methods that generate valid and useable data on teacher education and preparation effectiveness without imposing a burden to programs that distracts them from the fundamental task of preparing future teachers. This is especially relevant for smaller programs where resources are limited.

Employment Outcomes

The rationale for the creation of this indicator is anchored in the belief that the quality of a teacher education and preparation program can be determined by the placement and retention rates of its graduates, which is in turn seen as a reflection of the program's ability to meet the demand for effective and qualified teachers. The federal regulations propose to assess program quality by two indicators of employment outcomes: teacher placement rate and teacher retention rate.

The teacher placement rate is defined as the percentage of new teachers "…who have been hired in a full-time teaching position for the grade level, span, and subject area in which the teachers were prepared" (U.S. Department of Education, 2014, p. 71834). The teacher retention rate is determined by any one of three possible rates: percentage of new teachers working as full-time teachers for three consecutive years in a five-year period; percentage of new teachers granted tenure (or equivalent) within five years of certification to serve as teacher of record; or the percent of teachers whose employment was terminated by their employer within five years of certification to serve as teacher of record.

The teacher placement and teacher retention rates would be annually calculated and reported separately for all schools and for high-need schools. Importantly, the regulations would reward programs that are able to show evidence of graduates' high placement and retention rates in high-need schools. The definition of high-needs schools is complex and should meet at least one of the two criteria. One, schools in the top quartile as "…ranked in descending order by percentage of students from low-income families enrolled in such schools, as determined by the local educational agency based on a single or a composite of two or more of the following measures of poverty: (a) The percentage of students aged 5 through 17 in poverty; (b) the percentage of students eligible for a free or reduced price school lunch under the Richard B. Russell National School Lunch Act; (c) the percentage of students in families receiving assistance under the State program funded under part A of title IV of the Social Security Act; and (d) the percentage of students eligible to receive medical assistance under the Medicaid program" (USDOE, 2014, p. 71834). Two, an elementary school where at least 60% of its students qualify for free or reduced price lunch or a non-elementary school where at least 45% of its students qualify for free or reduced price lunch.

The proposed indicator contains a provision allowing states to exclude certain groups of teachers from their report. These would include: "(a) New teachers who have taken teaching positions in other States, (b) new teachers who have taken teaching positions in private schools, (c) new teachers who are not retained due to market conditions or circumstances particular to the LEA (Local Education Authority) and beyond the control of teachers or schools, or (d) new teachers who have enrolled in graduate school or entered military service" (USDOE, 2014, p. 71834). In addition, states would have the ability to propose different placement and retention rates for traditional teacher education programs and for alternative routes.

Again the idea that teacher education and preparation programs can impact the placement and retention rates of their graduates is not new and has received sufficient attention in the research literature. However, this is the first time that the federal government is proposing to use the link between these two variables (teacher education and preparation and their employment outcomes) as a high-stakes indicator of program success. The evidence to support such actions in the research literature is mixed.

Two features of teacher preparation programs are reported to be correlated to teacher placement: program length and program pathways. For example, Andrew (1990) compared the differences between graduates of a 4-year and a 5-year teacher preparation

program with the data from a 10-year study (1976-1986) of a random sample of students from both programs. He concludes that 5-year teacher education programs tend to have higher rates of graduates entering the teaching profession compared to those from 4-year programs. Darling-Hammond's (2000) review of several studies investigating the same topic also confirmed Andrew's conclusion.

Another feature that has been examined in the literature is the effectiveness of different routes to preparing teachers, with a great deal of attention given to the effectiveness of traditional teacher education versus alternative routes programs. Based on his review of 92 relevant studies on this topic, Allen (1999) suggests that alternative teacher preparation programs are likely to recruit more candidates into hard-to-staff schools. However, the traditional route continues to place the majority of teachers around the country.

Andrew's study, and some others mentioned above, also investigated the relationships between teacher education programs and teacher retention rates. According to Andrew, alternative route preparation programs are more likely to generate greater short-term retention rates, for instance tracks such as [Teach for America](TFA), but concerning long-term retention rates there is no evidence to show whether alternative routes or traditional routes are significantly more successful. More research is needed in this area.

A study seeking to estimate the mathematics and reading effectiveness of out-of-state prepared teachers in North Carolina elementary schools by Bastian and Henry (2015) showed mixed results when considering the diversity in quality among teachers who came from exporting states. The study found that teachers who received their teacher education out-of-state were less effective than teachers prepared in-state or than in-state alternative entry teachers, but they also found a "substantial overlap" in the distribution of effectiveness across groups as judged by value added measures. While the study is problematic because the authors failed to consider the role of schools' social networks and in-school norms and induction support as possible explanations, they conclude that "differences in human capital" (as measured by standardized licensure exam scores) helped explain out-of-state teachers' underperformance. While some out-of-state teachers performed in some cases better than in-state teachers the authors also listed states that seem to export comparable (New York, Michigan, South Carolina, and West Virginia) and less effective (Pennsylvania, Ohio, and Virginia) teachers. The suggestion is that teacher mobility across borders may result in lower pupil performance and that in-state prepared teachers would be more successful. The study recommends higher levels of compensation to recruit and retain high quality teachers. But because of the limitations in this study, these recommendations should be taken with caution.

In short, the length of the program on the one hand, and the preparation route on the other are two identified program factors that may have influence on the placement and retention rates of their graduates, but the degree and direction of the influence is still unclear. Other program factors however show important associations with employment outcomes. A study by Freedman and Appleman (2009) used both qualitative and quantitative data to study how teacher education contributed to teacher retention in high-poverty, urban schools for 26 UC-Berkeley graduates. Their findings suggest that substantive preparation that includes a balance of the practical and the academic may encourage more teachers to stay in hard-to-staff schools. Ingersoll, Merrill and May (2012) reached similar findings in their study of the role that teacher preparation played in retaining teachers. They used two nationally representative datasets: the 2003–2004 Schools and Staffing Survey and the 2004–

2005 Teacher Follow-Up Survey. Based on their analysis, they concluded that teachers who receive less pedagogical training are more likely to leave teaching.

Nurturing the resilience or disposition of working in a challenging context is another aspect that has been identified as a factor that may contribute to teacher retention rates. Yost (2006) reasoned that teachers who stay in the profession after their first year of teaching may have some common characteristics that could be traceable back to teacher preparation programs. Guided by this idea, she designed a qualitative study and interviewed 17 teachers who were teaching in their second year trying to find out the teacher characteristics that hold them in the teaching profession and how they are connected to teacher preparation programs. She triangulated the interview findings with teaching observations and principal interviews. Resiliency and persistence surfaced from the data as the two teacher characteristics that seemed positively associated with teacher efficacy and teacher retention rates. Traits of resiliency and persistence describe people who "are able to recover strength and spirits quickly and persevere in the face of obstacles" (p.1). This study suggests that teacher preparation programs may be able to enhance teacher retention by fostering resiliency and persistence in teacher candidates.

Teachers' overall satisfaction with their teacher preparation seems to be related to their career decisions according to a study by DeAngelis, Wall, and Che (2013). They used survey data from the 2003-2004 academic school year including information from 4,974 teachers and found that lower levels of satisfaction with the teacher preparation program may lead to higher rates of student teachers leaving the teaching profession.

The research reviewed above suggests that some features of teacher preparation programs are associated with teacher retention rates, but in complex ways that are yet to be understood from existing research.

A set of studies have explored the association between teacher quality and employment (as indicted by retention rates) in high-need schools. Taking New York State public schools as an example, less qualified teachers were more likely to teach in schools with higher concentrations of nonwhite, poor, and low achieving students than their more qualified peers (Boyd, Loeb, Lankford & Wyckoff, 2005), although more recently this pattern has changed in New York (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Lankford, Loeb, Mceachin, & Wyckoff, 2014). Similar patterns showing unequal distribution of quality teachers have been identified in other states and countries (Chudgar & Luschei, 2013).

According to several scholars, particular attributes of teacher preparation programs are related to higher rates of success in graduates' placement and retention in hard-to-staff and high-need schools. These include longer preparation programs (Andrew, 1990; Darling-Hammond, 2000), more substantive preparation in pedagogical and methods-related knowledge and skills (Freedman & Appleman, 2009; Ingersoll, Merrill & May, 2012; Ronfeldt, Schwartz & Jacob, 2014), and teachers' overall satisfaction with their preparation (DeAngelis, Wall & Che, 2013). Recent work has also shown the importance of initial student-teaching placement for later employment success (Goldhaber & Cowan, 2014). There are, however, a number of studies that argue that teacher preparation is only one factor in determining where and how long a teacher chooses to teach. Many other factors, such as a teacher's individual characteristics, working conditions, and the job market collectively, shape the distribution of teachers across schools.

For example, Ingersoll (2001) adopted an organizational analysis framework to investigate the relationships between teacher turnover and teacher shortages. By using and analyzing data from the Schools and Staffing Survey and the Teacher Follow-up Survey, he

suggests that teacher turnover is strongly correlated with the individual characteristics of teachers, rather than teacher preparation programs. Similarly, by reflecting on her personal experience in high-need schools, Nelson (2004) argues that teaching at its best is not a codified and prescribed technical "how-to" exercise, but rather a dynamic intellectual activity, and suggests that multiple factors are jointly shaping the teaching force of U.S. public schools.

Thus the evidence from the literature seems to support Kumashiro (2015) who argues that the employment outcome measures proposed in the federal regulations inaccurately presume that placement and retention are the result of program quality, without sufficient acknowledgment of the role of the job economy, personal life circumstances, and preferences that can affect employment and tenure. The research on workplace conditions provides evidence that teacher education and preparation programs alone are not solely responsible for new teachers' employment outcomes. Research on the impact of mentoring support provided to beginning teachers on teacher satisfaction and retention shows that the quality of learning opportunities available to new teachers profoundly affects their career decisions. For example, in their longitudinal study Cameron and Lovett (2015) identified the quality of school leadership and working conditions as key factors influencing teachers' decision to stay, move schools, or leave teaching altogether. A similar claim is made in other studies in the teacher labor market field (Boyd, Lankford, Loeb & Wyckoff, 2005).

Implications. The proposed three-year tracking of employment outcomes for graduates represents a high level of investment by programs which is unlikely to produce the intended results. Programs that are most likely to show success will be those that have substantial resources to engage in the required longitudinal effort, those who have managed to build strong social networks among their graduates, and those for whom a condition for obtaining a teacher credential is employment in schools for a set time period, as in the TFA case. While it may seem legitimate to use placement and retention rates as indicators of teacher education and preparation program success, the unintended consequences of this action may be the abuse or misuse of these measures which may mistakenly attribute the unequal distribution of quality teachers to programs without taking into account individual preferences and other exogenous factors. Tying these results to high-stake decisions (e.g., closing teacher education and preparation programs), may further discourage the whole field from developing unique and sustainable strategies for preparing quality teachers to successfully engage with the variable contexts of schooling across the nation.

Pupil Learning Outcomes

While few would argue against the idea that one important goal of teacher education and preparation is to help develop quality teachers who will in turn influence pupil learning, the research literature demonstrates the difficulty in establishing these causal relationships. According to the proposed regulations, programs would need to link individual graduates to their students' learning in K-12 schools, using measures of student achievement growth and or local evaluation measures for their first three years of teaching. While this proposal raises a series of validity concerns (e.g., AERA & NCME, 2014, p. 14 details the validity issues involved in using tests for a purpose other than for which they were designed) neither the regulations nor the studies we found include evidence that they have attended to this concern, and that they intend or have evaluated the appropriateness of pupil's assessments for the purpose of complying with teacher evaluations mandates or with teacher education and preparation program evaluation studies. With this caution in mind, this section reviews

the research literature exploring the link between the effectiveness of teacher education and preparation programs with K-12 outcomes.

While some recent work is promising, much of the empirical work that attempts to tie individual teacher education and preparation programs to student achievement lacks the data needed to evaluate why a particular program may be effective or more effective than another. Several studies that have attempted to examine teacher education program content and experiences in depth lack a link to K-12 student outcomes, an indicator for individual programs, or a sample size large enough within programs to be able to measure individual program impact. A large number of studies use available data often collected for different purposes to explore questions of program's effectiveness. We describe these below.

Using multiple administrative data sets from the state of New York, Boyd, Grossman, Lankford, Loeb and Wyckoff (2009) analyzed a variety of program graduates' indicators and related them to student outcomes in order to estimate program effectiveness. Using value-added methods, the authors found significant associations between graduates' opportunities to learn during classroom practice (e.g. "listening to a child read aloud for the purpose of assessment, planning a guided reading lesson," or "analyzing student math work") (p.434) and student achievement in NYC schools (Boyd et al., 2009).

Other studies that also use administrative data collected at the state level have found that most programs are indistinguishable from one another as defined by their impact on the achievement of pupils of program graduates in Louisiana (Gansle, Noell, & Burns, 2012), and Missouri (Koedel, Parsons, Podgursky, & Ehlert, 2015). A study conducted in Washington State, however, found some variation across different programs as indicated by graduates' effectiveness at raising student achievement, but this difference was found for only 2 out of the 13 institutions analyzed and only in reading, not in math (Plecki, Elfers, & Nakamura, 2012).

Building on previous methodological concerns (Koedel, Parsons, Podgursky & Ehlert, 2012; Mihaly, McCaffrey, Sass, & Lockwood, 2013), work by Goldhaber, Liddle and Theobald (2013) analyzed administrative databases prepared by Washington State's Office of Superintendent of Public Instruction, concluding that where a teacher receives their credential represents only a small portion of the variation in teacher effectiveness as explained by value-added to pupil achievement. The authors interpreted this result to be meaningful: "the regression-adjusted difference between teachers who receive a credential from the least and most effective programs is estimated to be 3.9% to 13.4% of a standard deviation in math and 9.2% to 22% of a standard deviation in reading" (p. 42). More research is needed to investigate in greater detail the source of the observed variation.

A promising line of research examines the impact of student teaching and field placement in traditional teacher education programs on K-12 student achievement. Using data from New York City and controlling for program fixed effects, Ronfeldt (2012) found that once hired, future teachers who had field placements in easy-to-staff schools during their program had a positive effect on student achievement, even if the teachers ended up teaching in hard-to-staff schools. Digging deeper into student teaching, Ronfeldt (2015) found that highly collaborative field placements had an influence on novice teachers' retention and on their ability to raise student achievement once they enter the field. This is an area that merits careful study to better understand how field placement influences graduates' impact on their future students' achievement gains.

Several studies compare the performance of teachers that have graduated from traditional and alternative pathways. Constantine et al.(2009) conducted a large scale study to evaluate differences between alternatively certified and traditionally certified teachers.

While the main aim of this study was to examine the outcomes of teachers trained through different pathways, program structure was also analyzed. The evaluation included 2,600 students and their teachers in 63 schools and 20 districts in several states across the country and used a purposive sample of 87 matched pairs of alternatively certified and traditionally certified teachers who were randomly assigned to students. The authors found that neither the pathway (alternative or traditional teacher education) nor the amount or content of coursework were associated with teacher effectiveness as measured by their pupils' test scores.

Research on Teach for America (TFA), a highly selective alternative certification program with a two year teaching commitment in hard to staff areas reports mixed impact on student achievement. Several non-experimental studies find that students of TFA teachers perform about the same or worse as traditionally trained teachers. Raymond, Fletcher, and Luque (2001) found no statistically significant difference between the students of TFA and non-TFA teachers in Houston. Similarly, Darling-Hammond et al. (2005) found that grade 4 and 5 Houston TFA teachers are about as effective in raising math and reading achievement as their traditionally prepared counterparts. Boyd et al. (2006) found that TFA teachers in New York City were slightly less effective than their traditionally trained colleagues. These differences, however, did disappear after the first few years. Kane et al. (2008) confirmed these findings two years later in New York City.

Other studies have found positive impacts of TFA teachers. Using longitudinal data in North Carolina, Xu, Hannaway, and Taylor (2011) found that high school TFA teachers were significantly more effective at raising student achievement than their peers, especially in science. Henry et al. (2014) reported similar findings in North Carolina high schools. Several experimental studies from Mathematica have found positive impacts as well. Using random assignment of students to TFA and non-TFA teachers in 6 regions, Decker, Mayer, and Glazerman (2004) found that the average math scores of students of TFA teachers were about .15 standard deviations higher than students of non-TFA teachers. The TFA teachers did not appear to have an impact on reading however. With the sample restricted to only novices, TFA teachers were even more effective in math, about a .26 effect size. Another experimental study by Mathematica found that TFA teachers were more effective at raising student achievement in secondary math compared to both traditionally trained teachers and other less selective alternative pathways (Clark et al., 2013).

Attention has also been given by researchers to other alternative routes to teaching. Findings for the Teaching Fellows, another selective alternative certification program, tend to be null with regard to value-added to student achievement relative to traditionally trained teachers and other less selective alternative pathways (Boyd et al., 2006; Clark et al., 2013; Kane et al., 2008).

The American Board for Certification of Teacher Excellence (ABCTE) is another alternative route to teaching gaining a foothold in recent years. A Mathematica study (Tuttle et al.., 2009) used propensity score matching to analyze the effectiveness of ABCTE teachers relative to those traditionally trained, finding no difference in pupils' reading scores; in mathematics the ABCTE prepared teachers were less effective in math with an effect size of .25. In contrast, Sass (2015) found that ABCTE teachers in Florida were more effective (about 6-8% of a standard deviation higher) in math than those traditionally trained.

Another recent pathway to entry into the teacher workforce is the urban teacher residencies, which typically include a full year of classroom apprenticeship with master's level education coursework. These programs tend to also include attractive stipends, scholarships,

and a commitment to three or more years in the district beyond the residency year. Papay et al. (2012) analyzed the impact of one such program, the Boston Teacher Residency. Analyzing student achievement in grades 4-8, the authors found that Boston Teacher Residency graduates with available value-added data are no more effective at raising student achievement than other novices in English Language Arts (ELA) and are less effective in math. Their effectiveness improves rapidly by years 4 and 5, however, outperforming veteran teachers. While some alternatively trained teachers do appear to be more effective than their peers, the above findings taken together suggest the need for more research to analyze the factors associated with performance.

Given the mixed evidence found in the existing studies it is no wonder that this section of the regulations has resulted in a strong response from the field highlighting the limitations of the proposed methods to evaluate program impact using pupil achievement. Those in the field are no stranger to the debates in K-12 around the use of value-added measures in high-stakes accountability climates (e.g. Amrein-Beardsley & Collins, 2013; Corcoran, 2010; Hill, 2009; Rothstein, 2008). These debates include recent evidence which finds confusion and distrust of the measures in K-12 settings and a preference for principal ratings (Goldring et al.., 2015; Jiang, Sporte, & Luppescu, 2015). Others, however, argue favorably for their potential uses in impacting the quality of the teacher workforce (e.g. Goldhaber, 2015). In spite of these concerns some scholars agree that there is utility for value-added measures (VAMs) in empirical research, yet issues appear to arise when utilizing these measures in high-stakes accountability contexts.

Similar to the hesitations and cautions in K-12, scholars have questioned the validity of using VAMs to determine teacher education and preparation programs quality for accountability purposes. Feuer et al. (2013) recognizes the utility of VAM's to analyze program impact while controlling for factors external to teachers. In addition, they argue that these measures can avoid the issues that exist when evaluating individual teachers, as they would aggregate to the program level. However, they also recognize that program VAM scores could potentially reflect programs' selection criteria, rather than program impact, and recommend the need for multiple measures. Floden (2012) highlights the limitations of VAMs in teacher education and preparation program evaluation including the need to analyze more than student achievement and the need to use multiple metrics, whether average VAM scores are an appropriate measure of teacher education and preparation program quality, and that the employment outcomes of teachers can bias estimates.

While theoretically the limitations of VAMs could be mitigated through the use of other measures as the regulations intend, the authors' cautions are still important when deciding which measures to use and how to weight them. We would add that just as important is the appropriateness of using existing measures and already existing data for purposes other than originally intended (AERA & NCME, 2014, p. 15). Significantly, the current draft of the proposed regulations does not necessarily require the use of VAM's, but rather the metric that is used at the local district level which increasingly include teacher observation scores. The research on such metrics has been for the most part in the K-12 arena alone (e.g. MET Project), thus more research is needed to investigate whether program impact can be measured through these observation scores.

Flexibility to use local metrics does not fully resolve validity concerns and may actually exacerbate them making comparison across teacher preparation programs, within and across states, difficult or impossible. In their literature review Henry, Kershaw, Zulli, and Smith (2012) identify different options states must consider as they explore ways to evaluate program effectiveness for accountability purposes including which tests teachers are

to be held accountable for, how teachers in tested and non-tested subjects are included, and whether VAM's or a more transparent method is more appropriate for evaluation. They conclude by arguing that states need to address issues of accuracy (using multiple measures such as observations), fairness (addressing out-of-school factors statistically), transparency (ensuring stakeholders understand the methods used), and inclusiveness (measuring outcomes other than test scores, such as graduation rates and student engagement).

A recent report from the American Psychological Association (APA) was optimistic about the potential use of VAM's in an evaluation context, providing careful recommendations for its uses alongside student learning objectives, surveys, and observation instruments during the progression of a program, at its completion, and at post-graduation (Worrell et al.., 2014). While several of the measures that the APA report recommends are similar to those in the regulations from the USDOE, the construction of dedicated measures for the purposes of program evaluation and a focus on evaluation at several points *during* the program and after--suggesting the notion of developmental progressions or growth--is the major difference. Such a structure would provide more data points for the programs and stakeholders to utilize for improvement of practice and would render more valid results.

Implications. The empirical research evidence to support the efficacy of the proposed regulations with regard to linking teacher preparation outcomes to pupil outcomes is mixed. Lessons from K-12 may help inform the discussion around accountability metrics for teacher education and preparation programs. The inclusion of multiple metrics appears to be a step in the right direction. However different issues are likely to arise with program accountability. For example, measuring the effectiveness of a teacher preparation program based on existing measures of K-12 student achievement has validity problems as pupil assessments were developed for a purpose other than evaluating a teacher education and preparation program. A more fundamental problem, however, is the question of how teacher education and preparation programs should be evaluated. Previous work supports the notion that teacher education and preparation success should be measured according to a program's goals and content and its theory of action (Tatto, 2001). This approach would likely be the most useful element for programs seeking to improve their own practice and outcomes.

While the ultimate goal of teacher education preparation programs is for their graduates to be effective as demonstrated by their practices and pupil's success, the methodology to carry out valid evaluations to document teacher effectiveness as a result of teacher education and preparation still needs to be developed. The most obvious gap in the teacher preparation evaluation equation as formulated by the regulations concerns the links between program experiences and effective teaching practices of novice teachers as a moderating factor in pupils' learning. The degree to which novice teachers manage to teach the complex curriculum as required by emerging ambitious standards is a key question. To properly evaluate novice teachers' practices, it is important to understand the knowledge that novice teachers bring with them from their teacher preparation experiences and the ways in which schools' norms and social capital (e.g., knowledge, practices, and expectations of peer teachers; compliance with mandated assessments) support their planning and enacting of such ambitious standards. In other words, the evaluation of teacher preparation effectiveness needs to factor the degree to which it prepares future teachers to teach in challenging environments successfully. These kinds of studies are at the core of the basic research that is more likely to provide rigorous and useful information for program and school improvement. The current and growing movement to evaluate teachers would mostly

answer the question of pupil achievement for individual teachers. But it will not explain the kinds of practices that are most conducive for more effective teaching and deeper learning for all children, and what conditions, internal and external, for teacher education and preparation are more supportive of these practices.

Implementation, as often is the case, appears to be a major issue for evaluating teacher performance using student data. Some states are likely to have both the capacity and infrastructure for reporting and actually utilizing the data for productive purposes while others will struggle to do so. State level data collection of student progress seems to be central to the current regulation effort. After passage of the Every Student Succeeds Act (ESSA) in 2015, however, there is uncertainty as to what teacher evaluation will look like both within and between states, as well as what data will be collected in the process. Fidelity is also likely to be an issue. For example, the American Association of Colleges for Teacher Education (AACTE), many legislators, and other organizations have pushed back on the regulations, creating a negative narrative. Such a negative outlook could result in programs and states only going through the motions as a means to comply with the regulations rather than using the data to inform practice, as has been reiterated throughout this article. Lastly, it remains to be seen how sensitive programs will be to the potential loss of TEACH grants with many of the top programs in the U.S. (as ranked by *U.S. News and World Report*) administering very few grants to program graduates.

Additionally, thought needs to be given regarding how teachers and children can be supported by collaborative social networks in schools and communities (Bryk & Schneider, 2002). Focusing predominantly on student achievement when measuring schools' or programs' outcomes would overlook these other important areas.

Though there are many concerns, attempts to push teacher education and preparation programs to demonstrate effectiveness is held to be the next step in developing policies that seek to increase the quality of the educational system. Information on what program graduates learn as a result of their teacher education and preparation, and how they manage to enact a challenging curriculum to improve their pupils' learning could potentially help researchers to further disentangle the teacher preparation – teacher effectiveness question, an issue the field continues to struggle with.

Accountability Implementation Challenges

At the time of writing, the Department of Education had declared its intention to implement these regulations across the country without much pilot research. It is thus essential that the USDOE and teacher education and preparation program community, with support from AACTE and CAEP, ensure the production of rigorous empirical evaluations of the implementation, outcomes, and impact of the policy.

The literature review has highlighted a number of challenges that may provide direction for evaluating the implementation of the regulations. Among those the most salient challenges to implementation are definition and measurement, capacity, and use concerns. Accreditation of teacher education is dependent on the degree to which programs are able to provide evidence of effectiveness, but defining effectiveness is in itself a challenge. According to the proposed regulations, effectiveness is indicated by compliance with CAEP and InTASC standards and more broadly by constituents' levels of satisfaction with their preparation, their employment outcomes, and effective teaching as measured by their pupil's outcomes. While there may be some agreement around these definitions, the most important challenge is how to measure these different aspects of effectiveness. Problems of

measurement are closely associated with the capacity that institutions have to conceptualize and develop such measures and the resources to actually engage in a long-term agenda of self-evaluation. Finally, the effort invested in the development and implementation of self-study measures would ideally result in program improvement, yet this may only happen if the results of the self-study are used by those who have the power to engage in change. These challenges are discussed in more depth below.

Definition and Measurement Challenges

Measuring implies the identification of indicators and their definition, followed by the development of instruments or methods to measure them with rigor. In this area, however, the field seems highly incoherent and contested. A few examples will suffice. For instance, Norris (2013) looked at assessment challenges for teacher licensing as a response to accreditation. The study identified problems having to do with the determination of what the expected standards should be, how they should be measured, and whether they should be the same for all teachers and programs. The author points out that even if future teachers show good results on assessments, this may only mean that the participants are good learners and not necessarily that a quality teacher is produced.

Failing to obtain materials and access to teacher education program data, the National Council on Teacher Quality (NCTQ) undertook a study of programs in an effort to find alternative ways to evaluate teacher education programs. NCTQ looked specifically to analyze syllabi and used the results of these analyses to draw conclusions about program quality. Paulson and Marchant (2012) studied whether NCTQ's effort was a valid strategy and more generally asked what should be the basis for evaluating the quality and success of teacher preparation programs. Their exploration concluded that a focus on syllabi only reflects what is intended by individual instructors and fails to measure what future teachers actually learn. NCTQ continues its attempt at independently and comprehensively evaluating teacher education programs with mixed success. An authoritative report by Feuer et al. (2013) informed by a rich national and international research literature is critical of the NCTQ and other similar approaches and provides valid evidence-based alternatives to teacher education and preparation program evaluation.

Freeman, Simonsen, Briere and MacSuga-Gage (2014) selected a sample of all approved traditional and alternative track teacher preparation programs (this constituted 18% of teacher preparation programs in the nation; n = 1,940) to study their pedagogy offerings (e.g., classroom management) in depth. The course catalogues of these programs were analyzed and emails were sent in order to gain access to additional course materials. Of these, 10% responded due to concerns with sharing professors' intellectual property. The study found that while self-study policies may be in place, the methods used to collect information for accreditation processes may not be rigorous enough. Programs complying with NCATE accreditation had more stringent and comprehensive requirements.

The examples above illustrate that there is a high level of variability across the field as to what counts as indicators of program effectiveness and the tools and methods that are within the reach of programs to measure such indicators.

Capacity Challenges

An important challenge for teacher education program's accreditation is related to the issue of capacity. To date teacher education programs have been unable to develop valid, reliable, and sustainable self-study efforts, or if these are undertaken they are short-lived and

have limited impact beyond meeting accreditation demands (Cibulka, 2009). There are few yet notable exceptions, particularly among teacher education programs that in the late 1990s and early 2000 began implementing TEAC self-study guidelines (see Papanastasiou & Tatto, 2011; Tatto, 2003; Tatto & Papanastasiou, 2008).

Another example is the TEDS-M Study which designed a model and developed the methods to study the outcomes of mathematics teacher education in representative samples of teacher education programs in 17 countries including the U.S. (Tatto, et al.., 2012). The study not only collected data which contributed significant knowledge to the field but it also helped participating countries and institutions build capacity to carry-out rigorous research-based-self-study. According to the study's researchers it took a comprehensive country-wide effort and vigorous within and cross-country dialogue about teacher education approaches and existing and desirable indicators of program success to engage institutions in productive self-study (Tatto, 2013). In other words, building capacity to produce usable knowledge takes specialized expertise and sustained effort.

In another study, Bell and Youngs (2011) conducted case studies of 5 of the 17 institutions of higher education in Connecticut. They found that the self-study required by the accreditation process implicitly assumed that institutions would have in place an adequate system of collecting the needed information, as well as the manpower and capacity to evaluate the system in line with accreditation standards, which was not always the case. A particular issue is how to demonstrate the progression of students in a program such as the building of a portfolio. Their study found that teacher preparation programs implemented these data collection tasks in different ways. One way described was to assign homework in a similar format across classes. Student work was then evaluated using a rubric system. This system provided a way to progressively collect data throughout a student's time in the program so that student growth could be potentially measured. If this deceptively simple task is to be used to evaluate a program, it should be implemented in a standard manner within programs across time and across programs as well.

Coupland (2011) did a case study of Hillsdale College in Michigan, a small liberal arts school. Hillsdale College lost accreditation once NCATE and TEAC accreditation became a requirement. Coupland's account presents a potential scenario where small colleges end up discontinuing their teacher preparation programs due to costly accreditation requirements, with little support to develop the kind of information infrastructure necessary for compliance. In a perverse way, requests for accreditation could become a strategy to "squeeze out" smaller programs and gain more control of local institutions, rather than a program improvement strategy as the policy arguably intends.

Tindle, Freund, Maxine, Belknap, Green, and Shotel (2011) conducted a case study of an urban teacher residency program at George Washington University. They report that in order to meet specific NCATE diversity requirements there is a need to engage in continuous analysis of teacher's knowledge, beliefs, and practices. This approach, while labor-intensive, points to the need for aligning field instruction with coursework and the need to develop these experiences over time; in other words, the development of new norms to increase program integrity and coherence.

The examples above illustrate that developing capacity for self-study is a costly and time consuming endeavor and that it requires careful planning and expertise, and likely a dedicated team within the teacher education and preparation programs.

Use Challenges

The optimistic view of engaging in rigorous self-study in teacher education would be its pragmatic implications. The effort could result in improved programs as indicated by improved outcomes, both for teachers and their students. However, this could only happen if the faculty is fully engaged in the undertaking or if somehow the faculty found ways to use the results of ongoing studies (Cibulka, 2009). This is not the typical pattern however. The study by Bell and Youngs (2011) of 5 of the 17 institutions of higher education in Connecticut found that several were complying with standards by collecting data; however, they lacked the institutional capacity to analyze and make sense of this data for program improvement. They also observed that while some institutions used the accreditation demands as an opportunity for learning and self-renewal, other institutions saw it as one more bureaucratic requirement to fulfill, and these efforts remained marginal when considering program implementation and improvement.

Bullough, Clark, and Patterson (2003) conducted a critical analysis of accreditation using a case study of Brigham Young University. The study sheds light on how accreditation has led to a "troubling reduction" of the curriculum due to market assumptions about teacher education. The authors conclude that in many cases, compliance becomes the goal rather than programmatic improvement: "[w]hen means become aims, journeys become long and pointless and life weary" (p. 41).

Heafner, McIntyre and Spooner (2014) studied the links between clinical partnerships and program impact. The study found that comprehensive evaluation models are necessary in order to effectively document complex outcomes and concluded that a successful clinical partnership is greater than simply having positive relationships between universities and schools, mentorship, and a commitment to supporting diverse spaces.

In a recent literature review (Tatto, 2015) searched for comprehensive evidence of the impact of quality assurance (QA) mechanisms on teacher education programs' outcomes. Four country case studies were analyzed (i.e. United States, Singapore, Finland, and Chile) selected according to different levels of educational success as reported by Mourshed, Chijioke, and Barber (2010). While the available research in the area is sparse, the author found that programs that use research as part of a process of learning to teach and use this research for program improvement (e.g., self-study and use research-based practices) are more effective than those that do not. The author calls for a clear definition of what is meant by quality assurance and rigorous evaluations of the differential effectiveness of such mechanisms across socio-cultural settings.

A study of the impact of quality assurance mechanisms on teacher education programs' quality is documented in the TEDS-M study. Programs in countries where compliance with QA policies was emphasized had teacher education programs whose graduates demonstrated higher knowledge levels at the moment of graduation than those who did not have those regulatory policies or where compliance was not as rigorous (Tatto et al., 2013).

In sum, while the idea of accountability translated into formal accreditation mandates is not new and such policies have the potential to produce organizational learning, they may not immediately result in improved programs or better teacher learning outcomes. The response from teacher education and preparation programs to high-stakes accountability pressures has simply not been researched thoroughly; however, research in K-12 can provide some insight into consequences of high stakes accountability. For example, while there does appear to be evidence of improved academic performance due to increased accountability

pressures under No Child Left Behind (e.g. Chiang, 2009; Winters & Cowen, 2012), a number of other studies find negative consequences including teaching or staffing to the test (Cohen-Vogel, 2011; Jennings & Bearak, 2014), and an increased focus on particular students and academic subjects at the expense of others (Booher-Jennings, 2005; Dee, Jacob, & Schwartz, 2012; Krieg, 2011; Reback, 2008). Another strand of research finds confusion around data reporting among consumers in K-12 (Jacobsen, Saultz, & Snyder, 2013). This finding is particularly important considering the push for more data use and data reporting for programs. While this does not imply that accountability provisions and data reporting in higher education are necessarily a bad idea, caution should be taken given the lessons learned in K-12. A key question here is how to improve teacher education and preparation programs quality by introducing accountability strategies that are relevant and useful rather than just another hurdle to jump on the way to attaining accreditation.

The uneven quality and performance of teacher education and preparation programs across the states is an important element to be considered in the potential impact of regulatory policy, as norms and human and social capital within the programs themselves are likely to mediate the level of success that can be derived from such mandates. Programs that have already strong norms and resources in support of self-evaluation may derive important benefits from the regulations, but those that are weak would have to work hard to obtain the needed resources and to develop the required human and social capital, or they may become weaker if these conditions are not met. In sum there is no evidence that the newly proposed legislation with its periodic visits by accreditation agencies and the continuous collection of program information would change what is already a marginal task in teacher education and preparation programs (see GAO [Government Accountability Office], 2015).

Discussion and Conclusion

The federal initiative to regulate teacher education and preparation programs is seen by many as a necessary step to improve the quality of the profession, and to address the uneven performance of teacher education programs across the nation. Yet the lack of consistent research evidence to guarantee a desirable level of confidence that the regulatory measures would attain the intended effect in a country as decentralized and deeply divided on educational matters as is the U.S. should be a cause for concern.

The response from the field illustrates the challenges ahead. We use the concerns with the regulations on employment outcomes and the attempt to link teacher education with pupil outcomes to highlight problematic areas, but similar methodological and logical issues affect other sections of the regulations.

Concerning the use of employment outcomes to evaluate the effectiveness of teacher education and preparation programs, AACTE, in a written response to the Department of Education, has argued that "Using teacher retention rates in this manner—with a high-stakes result for their preparation programs—is inappropriate. Teachers leave schools for multiple reasons that are not directly a result of the teacher preparation program, including life changes, lack of resources needed for effective instruction, weak school leadership, and conditions at the school that are often related directly to the level of poverty found in the community" (AACTE, p. 14). The association also expressed strong concern about the differing and less-stringent standards for alternative teaching routes which would result in unfair comparisons; ones in which teacher education programs would come up short. Finally, AACTE argues that the high-need schools requirement could result in more new teachers being placed in these schools, which "…contradicts the ESEA requirement that

states work against congregating new teachers in high-need schools" (AACTE, p. 14). This view is shared by the Association of Mathematics Teacher Educators (AMTE), which argued that "…these proposed regulations could promote that very practice by incentivizing preparation programs to place first-year teachers in high-need schools" (AMTE, 2015, p. 2).

The AACTE and the AMTE arguments point to issues of reliability and validity of the measures that would be used to link program's graduates to pupil outcomes. These concerns echo the work by the National Academy of Education report above. Both further criticize the regulations in their current form for pushing "test and punish" forms of accountability in teacher education and preparation, similar to recent reforms in K-12 education, rather than a focus on improvement, capacity building, and innovation. Indeed, conceptions of teaching and learning are severely limited within these regulations.

In sum, the process of improving teacher education and preparation should be based on valid empirical research and feedback from the field. As discussed above, such research is relatively limited and should be promoted and funded by the U.S. Department of Education (for instance supporting pilot evaluations across targeted teacher education programs) if the goal is to develop and implement valid, useful, and relevant policy. Without these actions the regulations may promote a closed system where existing measurements are confused with norms as the main reference point and where the effect is policing the status quo and squelching innovation.

We conclude by pointing out a number of potential unintended consequences that need to be addressed during implementation. We restrict ourselves to five.

One, the focus on accountability measures may lead to the diverting of resources from day-to-day program activities such as maintaining program norms and social networks that are crucial for the successful functioning of teacher preparation programs (Heafner, McIntyre & Spooner, 2014). As indicated earlier, these internal and external networks are vital to the development of effective teachers. Students in programs with limited resources may have fewer opportunities to benefit from coherent norms and strong school networks (Bell, & Youngs, 2011; Bullough, Clark, & Patterson, 2003; Cibulka, 2009; GAO, 2015).

Two, the pressure to implement these regulations under very tight deadlines inherently favors stronger programs that already have embedded research as part of their modus operandi, established internal regulatory mechanisms, and the resources or networks to create accountability mechanism within a short time frame. But even well-resourced teacher education and preparation programs may not have the wherewithal to collect and/or make meaning of the data required by these regulations as has occurred in K-12 (Jacobsen, Saultz, & Snyder, 2013). While much organizational learning and self-regulation occurs in the day-to-day functioning of teacher education and preparation programs, demonstrating accountability in terms of knowledge and skills, employability, and effective practice, requires a distinct effort. Given this challenge, programs and state education agencies may end up subcontracting much of their evaluation work, adding yet another stakeholder to the process. Accountability structures could be integrated to support program learning (e.g., assessing knowledge throughout their program and adjusting courses to maximize desirable outcomes), yet they could also be conceived as parallel structures functioning independently of programs' day to day work. This potential separation between those who do the job of teacher education and those who evaluate it could hinder programs' normative capacity required to self-regulate their work (Tatto, 2011; Tindle, Freund, Maxine, Belknap, Green, & Shotel, 2011).

Three, increased regulations may lead to positive and fruitful alliances (e.g., Deans for Impact) but they also may lead to the 'balkanization' of teacher preparation programs, with highly effective programs creating exclusive networks among themselves which may open spaces to maintain some independence while at the same time finding a way to moderate federal scrutiny. This may happen, however, at the expense of smaller or less effective programs, to the detriment of the profession as a whole (Coupland, 2011). If the result of the regulatory process is seen as a zero-sum game with winners and losers the healthy collaboration that may have existed across and within states (and across programs, schools, and teacher educators), as once occurred with the Holmes Group, could become a thing of the past.

Four, in addition to capacity and use concerns, there are a number of unresolved methodological issues. Indicators of effectiveness need to be carefully conceptualized, defined, and measured. These indicators would only be as good as the methods used to collect the data; thus issues having to do with what and who gets measured and how (e.g., validity and fidelity of the measures) would need to be addressed before regulations are fully implemented. For instance, valid outcomes of teacher education and preparation would need to include indicators of knowledge for teaching, as well as carefully designed studies to evaluate the effectiveness of novice teachers' practices in challenging contexts (AERA & NCME, 2014; Rossi & Freeman, 2004; Weiss, 1998). Suggested indicators of programs' employment outcomes are particularly problematic and may lead to the overestimation of such outcomes for school-based routes, and the disproportionate placement of novice teachers in underserved schools (Ingersoll, Merrill, & May, 2012; Kumashiro, 2015).

Five, much data will be produced by these regulations which may be potentially publicly available before its fidelity can be properly confirmed. This presents a high risk for programs in a policy environment characterized by lack of trust and a high level of vulnerability for the teaching profession. Consequently, issues having to do with appropriate data use at the national, state and institutional level would in itself need to be regulated to avoid abuse by unscrupulous critics of teacher education (Darling-Hammond, 2013; Paulson and Marchant, 2012).

Policy makers, researchers, and practitioners should be on the same page when it comes to improving the preparation and supply of knowledgeable, effective and creative teachers who are well equipped to begin their careers, not as sole practitioners, but as members of a supportive collective dedicated to their ongoing professional growth and that of their students.

References

Allen, M., Palaich, R., & Anthes, C. (1999). *Teacher recruitment, preparation and retention for hard-to-staff schools.* A Report of the August 29-20, 1999 Meeting in Chicago, IL. Retrieved from http://eric.ed.gov/?id=ED440948

American Association of Colleges for Teacher Education (2015). *AACTE response to ESED.* Retrieved from http://aacte.org/resources/regulations

American Educational Research Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing (US) (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Amrein-Beardsley, A., & Collins, C. (2013). Value-added model (VAM) research for educational policy: Framing the issue. *Education Policy Analysis Archives*, *21*(4). Retrieved from http://epaa.asu.edu/ojs/article/view/1311.

Andrew, M. D. (1990). Differences between graduates of 4-year and 5-year teacher preparation programs. *Journal of Teacher Education*, *41*(2), 45–51. http://dx.doi.org/10.1177/002248719004100206

Association of Mathematics Teacher Educators (2015). *Official AMTE statement on proposed regulations.* Retrieved from http://amte.net/issue/federal-regulations

Bastian, K. C. & Henry, G.T. (2015). Teachers without borders: Consequences of teacher labor force mobility. *Educational Evaluation and Policy Analysis*, *37*(2), 163-183. http://dx.doi.org/10.3102/0162373714535662

Bell, C. A., & Youngs, P. (2011). Substance and show: Understanding responses to teacher education programme accreditation processes. *Teaching and Teacher Education*, *27*(2), 298-307. http://dx.doi.org/10.1016/j.tate.2010.08.012

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268. http://doi.org/10.3102/00028312042002231

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy, 1*(2), 176–216. http://doi.org/10.1162/edfp.2006.1.2.176

Boyd, D., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440. http://dx.doi.org/10.3102/0162373709353129

Boyd, D., Grossman, P. L., Hammerness, K., Lankford, R. H., Loeb, S., McDonald, M., … Wyckoff, J. (2008). Surveying the landscape of teacher education in New York City: Constrained variation and the challenge of innovation. *Educational Evaluation and Policy Analysis*, *30*(4), 319–343. http://dx.doi.org/10.3102/0162373708322737

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). Explaining the short careers of high achieving teachers in schools with low performing students. *American Economic Review, 95*(2), 166-171. http://dx.doi.org/10.1257/000282805774669628

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement.* New York, NY: Russell Sage Foundation.

Bullough Jr., R. V., Clark, D. C., & Patterson, R. S. (2003). Getting in step: Accountability, accreditation and the standardization of teacher education in the United States. *Journal of Education for Teaching*, *29*(1), 35-51.http://dx.doi.org/10.1080/0260747022000057945

Cameron, M., & Lovett, S. (2015) Sustaining the commitment and realising the potential of highly promising teachers. *Teachers and Teaching, Theory and Practice*, *21*(2), 150-163. http://dx.doi.org/10.1080/13540602.2014.928132

Chiang, H. S. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, *93*(9-10), 1045–1057. http://doi.org/10.1016/j.jpubeco.2009.06.002

Chudgar, A., & Luschei, T. (2013). *Final report: Study of teachers for children marginalized by social origin, economic status or location.* Retrieved from https://www.unicef.org/education/files/Final_Report-Teachers_for_Marginalized_Children.pdf

Cibulka, J. G. (2009). Improving relevance, evidence, and performance in teacher preparation. *Education Digest: Essential Readings Condensed for Quick Review*, *75*(2), 44-49.

Clark, M. A., Chiang, H.S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A. & Puma, M. (2013). *The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows Programs* (NCEE 2013-4015). Retrieved from https://ies.ed.gov/ncee/pubs/20134015/pdf/20134015.pdf

Coggshall, J. G., Bivona, L., & Reschly, D. J. (2012). *Evaluating the effectiveness of teacher preparation programs for support and accountability: Research and policy brief.* Retrieved from http://www.gtlcenter.org/sites/default/files/docs/TQ_RandP_BriefEvaluatingEffect iveness.pdf

Cohen-Vogel, L. (2011). "Staffing to the test": Are today's school personnel practices evidence based? *Educational Evaluation and Policy Analysis, 33*(4), 483–505. http://doi.org/10.3102/0162373711419845

Constantine, J., Player D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification, final report.* (NCEE 2009- 4043). Retrieved from https://ies.ed.gov/ncee/pubs/20094043/pdf/20094043.pdf

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? The use of value-added measures of teacher effectiveness in policy and practice.* Retrieved from https://eric.ed.gov/?id=ED522164

Coupland, D. B. (2011). The cost of accreditation: Hillsdale ends its teacher certification program. *Academic Questions*, *24*(2), 209-221. http://dx.doi.org/10.1007/s12129-011-9225-z

Crowe, E. (2010). *Measuring what matters: A stronger accountability model for teachers.* Retrieved from https://www.americanprogress.org/issues/education/reports/2010/07/29/8066/measuring-what-matters/

Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education*, *51*(3), 166–173. http://dx.doi.org/10.1177/0022487100051003002

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education*, *57*(2), 120–138. http://dx.doi.org/10.1177/0022487105283796

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives, 13*(42), 1–48. Retrieved from http://epaa.asu.edu/ojs/article/view/147

Darling-Hammond, L. (2013). Why the NCTQ teacher prep ratings are nonsense. *Washington Post.* Retrieved on June 18, 2013 from http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/06/18/why-the-nctq-teacher-prep-ratings-are-nonsense/

DeAngelis, K. J., Wall, A. F., & Che, J. (2013). The impact of preservice preparation and early career support on novice teachers' career intentions and decisions. *Journal of Teacher Education*, *64*(4), 338-355. http://dx.doi.org/10.1177/0022487113488945

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The effects of Teach For America on students: Findings from a national evaluation.* Retrieved from: http://www.mathematica-mpr.com/~/media/publications/PDFs/teach.pdf

Dee, T., Jacob, B. A., & Schwartz, N. L. (2012). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252–279. http://dx.doi.org/10.3102/0162373712467080

Feuer, M. J., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options*. Retrieved from https://eric.ed.gov/?id=ED565694

Floden, R. E. (2012). Teacher value added as a measure of program quality: Interpret with caution. *Journal of Teacher Education*, *63*(5), 356–360. http://dx.doi.org/10.1177/0022487112454175

Freedman, S. W., & Appleman, D. (2009). "In it for the long haul"—How teacher education can contribute to teacher retention in high-poverty, urban schools. *Journal of Teacher Education*, *60*(3), 323-337. http://dx.doi.org/10.1177/0022487109336181

Freeman, J., Simonsen, B., Briere, D. E., & MacSuga-Gage, A. S. (2014). Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *37*(2), 106-120. https://doi.org/10.1177/0888406413507002

Gansle, K. A., Noell, G. H., & Burns, J. M. B. (2012). Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education*, *63*(5), 304–317. http://doi.org/10.1177/0022487112439894

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher, 44*(2), 87–95. http://doi.org/10.3102/0013189X15574905

Goldhaber, D., & Cowan, J. (2014). Excavating the teacher pipeline: Teacher preparation programs and teacher attrition. *Journal of Teacher Education*, *65*(5), 449–462. http://dx.doi.org/10.1177/0022487114542516

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29–44. http://dx.doi.org/10.1016/j.econedurev.2013.01.011

Goldring, E., Grissom, J. A, Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96–104. http://doi.org/10.3102/0013189X15575031

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* Retrieved from http://www.urban.org/url.cfm?ID=1001431

Heafner, T., McIntyre, E., & Spooner, M. (2014). The CAEP standards and research on educator preparation programs: Linking clinical partnerships with program impact. *Peabody Journal of Education*, *89*(4), 516-532. http://dx.doi.org/10.1080/0161956X.2014.938998

Henry, G. T., Kershaw, D. C., Zulli, R. A., & Smith, A. A. (2012). Incorporating teacher effectiveness into teacher preparation program evaluation. *Journal of Teacher Education*, *63*(5), 335–355. http://doi.org/10.1177/0022487112454437

Henry, G. T., Purtell, K. M., Bastian, K. C., Fortner, Ck., Thompson, C. L., Campbell, S. L. (2014). The effects of teacher entry portals on student achievement. *Journal of Teacher Education, 65*(1), 7–23. http://doi.org/http://dx.doi.org/10.1177/0022487113503871

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, *28*(4), 700–709. http://dx.doi.org/10.1002/pam.20463

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499–534. http://dx.doi.org/10.3102/00028312038003499

Ingersoll, R., Merrill, L., & May, H. (2012). Retaining teachers: How preparation matters. *Educational Leadership*, *69*(8), 30-34. Retrieved from https://eric.ed.gov/?id=EJ988729

Ingvarson, L., Beavis, A., & Kleinhenz, E. (2007). Factors affecting the impact of teacher education programmes on teacher preparedness: Implications for accreditation policy *European Journal of Teacher Education*, *30*(4), 351-381. http://dx.doi.org/10.1080/02619760701664151

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–136. https://doi.org/10.1086/522974

Jacobsen, R., Saultz, A., & Snyder, J. (2013). When accountability strategies collide: Do policy changes that raise accountability standards also erode public satisfaction? *Educational Policy*, *27*(2), 360–389. http://doi.org/10.1177/0895904813475712

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher, 43*(8), 381-389. http://doi.org/10.3102/0013189X14554449

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher*, *44*(2), 105–116. http://doi.org/10.3102/0013189X15575517

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615–631. http://doi.org/10.1016/j.econedurev.2007.05.005

Kennedy, M. M. (2007). Defining a literature. *Educational Researcher*, *36*(3) 139-147. http://dx.doi.org/10.3102/0013189X07299197

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012). *Teacher preparation programs and teacher quality: Are there real differences across programs?* Retrieved from http://economics.missouri.edu/working-papers/2012/WP1204_koedel_et_al.pdf

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs? *Education Finance and Policy, 10*(4), 508–534. http://doi.org/10.1162/EDFP_a_00172

Krieg, J. M. (2011). Which students are left behind? The racial impacts of the No Child Left Behind Act. *Economics of Education Review*, *30*(4), 654–664. http://doi.org/10.1016/j.econedurev.2011.02.004

Kumashiro, Kevin, K. (2015) *Review of proposed 2015 federal teacher preparation regulations.* Retrieved from http://nepc.colorado.edu/thinktank/review-proposed-teacher-preparation

Lankford, H., Loeb, S., McEachin, A., Miller, L. C., & Wyckoff, J. (2014). Who enters teaching? Encouraging evidence that the status of teaching is improving. *Educational Researcher*, *43*(9), 444-453. http://dx.doi.org/10.3102/0013189X14563600

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

Loadman, W. E., Freeman, D. J., Brookhart, S. M., Rahman, M. a., & McCague, G. J. (1999). Development of a national survey of teacher education program graduates. *The Journal of Educational Research*, *93*(2), 76–89. http://dx.doi.org/10.1080/00220679909597632

Mihaly, K., McCaffrey, D., Sass, T. R., & Lockwood, J. R. (2013). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy*, *8*(4), 459-493. http://dx.doi.org/10.1162/edfp_a_00110

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better.* Retrieved from http://www.mckinsey.com/client_service/social_sector/latest_thinking/worlds_most_improved_schools 25

NCTQ (2014). *Teacher Prep Review 2014*. Retrieved from http://www.nctq.org/dmsView/Teacher_Prep_Review_2014_Report

Nelson, C. (2004). Reclaiming teacher preparation for success in high-needs schools. *Education*, *124*(3), 475-481.

Norris, J. M. (2013). Some challenges in assessment for teacher licensure, program accreditation, and educational reform. *Modern Language Journal*, *97*(2), 554-560. http://dx.doi.org/10.1111/j.1540-4781.2013.12022.x

Papanastasiou, E.C., & Tatto, M.T. (2011). Programme theory, programme documents, and state standards in evaluating teacher education. *Assessment and Evaluation in Higher Education*, *37*(3), 305-320. http://dx.doi.org/10.1080/02602938.2010.534760

Papay, J. P., West, M. R., Fullerton, J. B., & Kane, T. J. (2012). Does an urban teacher residency increase student achievement? Early evidence from Boston. *Educational Evaluation and Policy Analysis, 34*(4), 413–434. http://doi.org/10.3102/0162373712454328

Paulson, S., & Marchant, G. (2012). Voices in education: Accountability in teacher education and the National Council on Teacher Quality. *The Teacher Educator*, *47*(1), 1-8. http://dx.doi.org/10.1080/08878730.2012.633483

Perlmutter, D.D. (2015). *How to use student evaluations wisely.* Retrieved from https://chroniclevitae.com/news/1035

Portes, A. (2000). The hidden abode: Sociology as analysis of the unexpected. *American Sociological Review*, *65*(1), 1-18. http://dx.doi.org/10.2307/2657287

Raymond, M. E., Fletcher, S., & Luque, J. A. (2001). *Teach for America: An evaluation of teacher differences and student outcomes in Houston, Texas.* Retrieved from http://credo.stanford.edu/downloads/tfa.pdf

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5-6), 1394–1415. http://doi.org/10.1016/j.jpubeco.2007.05.003

Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education, 66*(4), 304–320. http://doi.org/10.1177/0022487115592463

Ronfeldt, M., & Reininger, M. (2012). More or better student teaching? *Teaching and Teacher Education, 28*(8), 1091–1106. http://doi.org/10.1016/j.tate.2012.06.003

Ronfeldt, M., Schwartz, N., & Jacob, B. (2014). Does pre-service preparation matter? Examining an old question in new ways. *Teachers College Record, 116*(10), 1–46.

Rossi, P.H., Lipsey, M.W., & Freeman, H.E. (2003) *Evaluation: A systematic approach*, Newbury Park, CA: Sage Publications.

Rothstein, J. (2008). *Teacher quality in educational production: Tracking, decay, and student achievement.* Retrieved from http://dx.doi.org/10.3386/w14442

Sass, T. R. (2015). Licensure and worker quality: A comparison of alternative routes to teaching. *The Journal of Law and Economics*, *58*(1), 1–35. http://doi.org/10.1086/682904

Tatto, M. T. (2015). The role of research in the policy and practice of quality teacher education: an international review. *Oxford Review of Education*, *41*(2), 171-201. http://dx.doi.org/10.1080/03054985.2015.1017405

Tatto, M. T., Peck, R., Schwille, J., Bankov, K., Senk, S. L., Rodriguez, M., ... Rowley, G. (2012). *Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M).* Amsterdam: International Association for the Evaluation of Educational Achievement.

Tatto, M. T. (Ed.) (2013). *Teacher education and development study in mathematics (TEDS-M): Technical Report.* Amsterdam: International Association for the Evaluation of Educational Achievement.

Tatto, M.T. (2011). Reimagining the education of teachers: The role of comparative and international research. *Comparative Education Review*, *55*(4), 495-516. http://dx.doi.org/10.1086/661769

Tatto, M.T. (2003). *Accountability in teacher education: Conceptual, methodological, and implementation challenges.* East Lansing: Michigan State University.

Tatto, M.T. (2001). The value and feasibility of evaluation research on teacher preparation: Contrasting the experiences in Sri Lanka and Mexico. *International Journal of Education and Development*, *22*(6), 637-657. http://dx.doi.org/10.1016/S0738-0593(01)00032-3

Tatto, M.T., & Papanastasiou, E. (2008). *The influence of teacher preparation on recent graduates' knowledge and views about teaching mathematics and literacy: A case study.* East Lansing: Michigan State University.

Thomas, A. M., & Loadman, W. E. (2001). Evaluating teacher education programs using a national survey. *The Journal of Educational Research*, *94*(4), 195–206. http://dx.doi.org/10.1080/00220670109598753

Tindle, K., Freund, M., Belknap, B., Green, C., & Shotel, J. (2011). The urban teacher residency program: A recursive process to develop professional dispositions, knowledge, and skills of candidates to teach diverse students. *Educational Considerations, 38*(2), 28-35.

Tuttle, C. C., Anderson, T., & Glazerman, S. (2009). *ABCTE teachers in Florida and their effect on student performance.* Retrieved from http://www.nctq.org/docs/ABCTE_FL_teachers.pdf

United States Department of Education (2014). *Teacher Education Issues. Proposed Rule.* Federal Register, *79*, 232 [http://www.gpo.gov/fdsys/pkg/FR-2014-12-03/pdf/2014-28218.pdf, consulted December 3, 2014].

United States Government Accountability Office (2015). *Teacher Preparation Programs: Report to the Ranking Member, Subcommittee on Health, Employment, Labor, and Pensions, Committee on Education and the Workforce, House of Representatives (GAO-15-598).* Retrieved from http://www.gao.gov/assets/680/671603.pdf

Weiss, C. (1998). *Evaluation.* Upper Saddle River, NJ: Prentice-Hall.

Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis*, *34*(3), 313–327. http://doi.org/10.3102/0162373712440039

Worrell, F. C., Brabeck, M. M., Dwyer, C. A., Geisinger, K. F., Marx, R. W., Noell, G. H., & Pianta, R. C. (2014). *Assessing and evaluating teacher preparation programs*. Retrieved from http://dx.doi.org/10.1037/e520852014-001

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach for America in high school. *Journal of Policy Analysis and Management, 30*(3), 447–469. http://doi.org/10.1002/pam.20585

Yost, D. S. (2006). Reflection and self-efficacy: Enhancing the retention of qualified teachers from a teacher education perspective. *Teacher Education Quarterly, 33*(4), 59-76. Retrieved from https://eric.ed.gov/?id=EJ795226

Appendix 1

Literature Search Procedures and Sources Used

We developed careful definitions and followed rigorous procedures in searching the literature and in selecting the studies we discuss. We found however that if we excluded studies because they failed to describe important features characterizing rigorous research reports (e.g., the methods for selecting a sample including the full description of the actual sample studied; whether existing data or measures used had been examined to confirm appropriate use for other purposes; how validity and reliability had been ensured) we would have excluded studies that have been or have the potential to be influential in the current policy debate. The same reason persuaded us to include significant literature reviews. Thus our goal was to arrive to a literature review that would represent the conceptual issues in the field (Kennedy, 2007) according to their relevance to the regulations' mandate. We point out the studies' limitations in the "sources of data" column.

We acknowledge that these studies' limitations also limit the power of our conclusions and call for more methodologically rigorous studies and reports. A summary of the references and findings is available in Appendix 2.

Accreditation in the Teacher Education Literature. We used ProQuest as our primary search engine. We first used "teacher preparation" and "accreditation" restricted to peer-reviewed articles. This resulted in many articles that were not necessarily relevant. Next we searched for "teacher preparation" and "accreditation process," which resulted in 420 articles. By restricting the search to articles since 2010 we identified 60 relevant articles. We then reviewed all of the articles within this final search for relevance. Michigan State University SearchPlus was used to download the full text of the selected articles. "Relevant", "high-quality", and "recent" were the three principles that we used to select articles from the initial searching results. We determined the relevance of an article based on whether the article addressed the accreditation process of a teacher preparation program, rather than for example a specific class of an accredited program, or the implementation of a policy in a K-12 setting. The focus was on the effects of accreditation on teacher education and preparation programs. Based on the aforementioned searching and selecting process, we ended up with 10 pieces of literature. They served as the data for this article (see Appendix 2, Table 1).

Teacher Education and Preparation Program Graduates' Knowledge and Ability Outcomes Literature. We began the search by using the Google Scholar database, and Michigan State University SearchPlus. We searched for peer reviewed articles and other influential publications. The following combinations of keywords were searched: "teacher education /preparation /effectiveness use of surveys", "teacher education /preparation /effectiveness survey graduates /employers". Those articles that did not meet our criterial for relevant, high-quality, or recent were discarded. Next, the design, methods, results and discussion sections of the selected articles was studied leaving 16 articles. We looked for peer reviewed research articles that evaluated teacher preparation programs using teachers or principal surveys and that included careful design and methods.

Two experts were presented with the body of the selected literature and asked whether anything relevant was missing. New works were suggested in this step. After analyzing their suggestions, it was considered that all were pertinent to this study and, therefore, were incorporated into the final set of literature that informed this review. This set considered both peer-reviewed and non-peer-reviewed literature. Cases of the latter were

included when either one or some of the authors were prominent in the field or when the associated agency is recognized and/or actively involved in the scene (see Appendix 2, Table 2).

Employment Outcomes Literature. We used Google Scholar Database as our primary search engine and supplemented it with Michigan State University SearchPlus. We first searched for all combinations of "teacher education/preparation" and "placement /employment/ recruitment/ retention/ hard-to-staff schools/high-need schools/assess/evaluate." We believed such a blanket search reached the majority of the relevant literature in recent years. The searching results were ranked by relevance and recent date of publication. The reviewers read the abstracts of the top 50 entries returned by each combination. For example, searching "teacher preparation and retention" pulled up 115,000 entries, but we only reviewed the abstracts up to the 50th entry and determined which ones should be included or not. The entries became significantly irrelevant to our topic from the 20th or 30th ones. Michigan State University SearchPlus was used to download the full text of the selected articles. In order to control the quality of the reviewed literature, we used "peer-reviewed" as a filter for selecting journal articles; as for research reports, we used both the selectivity of the publishing institutions and the times of citation as two proxies of quality. In terms of "recent," we narrowed our horizon to the studies published after 1995, with only one exception—an article published in 1990 but was very closely connected to the topic of this literature review (see Appendix 2, Table 3).

Pupil Outcomes Literature. The literature review in this section used Google Scholar and Michigan State University SearchPlus to gather peer reviewed scholarly articles with the following keywords: *teacher education and preparation effectiveness, teacher education, preparation quality, pupil achievement*. Reference lists of selected articles were used to discover more articles. The search identified 24 works relevant to the topic, including a book published by the National Academy of Education (see Appendix 2, Table 4).

Appendix 2

Literature Review Sources

Table 1
Teacher Preparation Accreditation Literature Review

Author(s) (year)	Title	Sources of data	Major findings
Bullough, Clark & Patterson (2003)	Getting in step: Accountability, accreditation and the standardization of teacher education in the United States.	A case study of Brigham Young University. A critical document analysis of NCATE accreditation's history to shed light on how accreditation has led to the reduction of education to what is described as "troubling reduction" due to market assumptions about teacher education.	Portfolios would be a significant tool in teacher preparation programs, however how to best use the data was found to be challenging. In many cases, compliance becomes the goal rather than programmatic improvement: "When means become aims, journeys become long and pointless and life weary (Bullough et al.., p. 41, 2003)."
Cibulka (2009)	Improving Relevance, Evidence, and Performance in Teacher Preparation	Written by President of NCATE and describes the goals of accreditation, providing examples	This article describes the overarching goals of accreditation. The goal of accreditation is to implement changes that promote excellence in addition to collegiality, inclusivity, partnerships, and cost-effectiveness. However, there may be little motivation to make or continue to make, changes once accreditation is achieved.
Coupland (2011)	The cost of accreditation: Hillsdale ends its teacher certification program	A case study of Hillsdale College using administrative data related to the NCATE /TEAC accreditation process.	Small colleges such as Hillsdale are giving up teacher preparation because accreditation requirements such as those from NCATE and TEAC. This example from Michigan can be seen as an attempt by the state to "squeeze out" independent programs and gain more control of local institutions.

Freeman et al. (2014)	Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs	Conducted document analysis of each state's policies concerning classroom management and analyzed reviews from student teachers and course offerings. Authors did not report sampling information.	While accreditation policies require data collection, the methods used may not be research based. Those requiring NCATE had more stringent and comprehensive data collection requirements.
Heafner, McIntyre & Spooner (2014)	The CAEP standards and research on educator preparation programs: Linking clinical partnerships with program impact	A case of a tutoring program at the University of North Carolina, Charlotte, and the study of the tutoring program's impact for candidates, high school students, clinical educators and faculty. Takes a critical look at the intersection of two standards: Clinical Partnerships and Practice (Standard 2) and Program Impact (Standard 4)	(1) comprehensive evaluation models are needed to document complex outcomes, which will make program impact more visible and measureable; (2) the authors recommend using the Feuer et al., decision-making framework to initiate and guide programmatic design and evaluation model development; (3) the authors affirm CAEP's expectations: clinical partnerships need to go beyond positive relationships, to the careful selection of mentor teachers and candidates, and commitment to diverse settings to ensuring P-12 student achievement. These are seen as attainable goals for educator preparation accreditation.
Norris (2013)	Some challenges in assessment for teacher licensure, program accreditation, and educational reform.	Summarizes the literature on accreditation and specifically emerging concerns. Author did not provide information on how the literature was searched.	The author finds potential conflict of interest among entities that accredit programs (if one company is paid for the process it may result in a monopoly); in determining expected standards for accreditation (how to insure the same standards are held for all? e.g., good learners do not necessarily mean a good teacher or a good teacher education program).

Paulson & Marchant (2012)	Voices in education: Accountability in teacher education and the National Council on Teacher Quality	12 editorial Advisory Board members of *The Teacher Educator* were asked to answer the following: 1. What is your impression of this (NCTQ) accountability effort for teacher preparation programs? 2. What should be the basis for evaluating the quality and success of teacher preparation programs?	Focusing on syllabi rather than focus on what is happening in the classroom is problematic and provides only a partial impression of what the program is doing. There is no measure of what student-teachers actually learn.
Tindle et al. (2011)	The urban teacher residency program: A recursive process to develop professional dispositions, knowledge, and skills of candidates to teach diverse students	A case study of the Urban Teacher Residency Program at George Washington University (GWU). Using multiple forms of data collected by the program (e.g., interview data with program participants) to examine how the program meets the 4a and 4d accreditation requirements of the NCATE through recursive analysis of teacher's knowledge, beliefs, and practices.	When considering factors that determine accreditation, field instruction must align with coursework, and build experiences over time, which indicates that specific factors should not be considered independently. The study concludes that a labor-intensive approach contributes in the success of GWU's program.

Table 2

Teacher Education and Preparation Program Graduates' Knowledge and Ability Outcomes Literature Review

Author(s) (year)	Title	Sources of data	Major findings
Coggshall, Bivona & Reschly (2012)	Evaluating the effectiveness of Teacher Preparation Programs for support and accountability (A research and policy brief)	The study examined the literature about TPPs evaluation from several sources. Authors did not provide information on the number of studies included and the total found.	Some studies found high correlations between principal assessment and teachers' value added scores. Few studies found correlations between principal's survey responses about preparation programs and teacher effectiveness. Surveys of graduates, when designed and administered carefully, can provide useful information to states and TPPs.
Crowe (2010)	Measuring what matters: A stronger accountability model for teacher education	The study examined the literature about TPPs evaluation from several sources. Authors did not provide information on the number of studies included and the total found.	The review suggests a need for increasingly evaluating teacher education outcome measures, including VAM, teacher tests, surveys of graduates and employers, among others.
Darling-Hammond (2006)	Assessing teacher education: The usefulness of multiple measures for assessing program outcomes	Based on research and assessments strategies used to evaluate program outcomes in the Stanford Teacher Education Program (STEP) during 5 years, including surveys of graduates and employers. Information on sampling was missing.	Employers' perception of STEP graduates was very positive (97% of them assigned 5/5 on overall perception of graduates' preparation); in fact, employers were less critical of graduates' preparedness than graduates themselves. Using survey data is not enough to determine the experience and effects from a TPP.

Feuer, et al. (2013)	Evaluation of teacher preparation programs: Purposes, methods and policy options.	Examined the literature about TPPs evaluation from several sources. Authors do not mention the number of studies searched and which were included or methods for inclusion.	Comprehensive description of TPPs evaluation instances. Authors propose a framework for evaluation. Surveys have a higher face value than other instruments and can provide with useful information, but there is concern about the amount of resources necessary to reach all graduates and the subjectivity and selectivity biases that might appear.
Harris & Sass (2009)	What makes for a good teacher and who can tell?	The study was carried out in a mid-size school district in Florida. The authors interviewed 30 principals; and assessed 31,645 grade 3-10 students for math test scores FCAT-Sunshine State Standards, and 30,974 students' reading scores on the same test.	The authors found (weak) positive correlations between principal ratings and value added scores (.15 to .30)
Jacob & Lefgren (2008)	Can principals identify effective teachers? Evidence on subjective performance evaluation in education	Data from a midsize school district in western US. Surveys were used to collect demographic indicators for students, and effectiveness indicators and other characteristics for teachers and principals in the district.	Principals are effective in identifying teachers that produce the most and the least student gains. They are not as effective in identifying teachers in the middle of the distribution.
Loadman et al. (1999)	Development of a national survey of teacher education program graduates	Data from the National Survey of Teacher Education Program Graduates collected from 1990 to 1996 for a sample of 3,940 baccalaureate graduates from 14 institutions.	According to the authors, the instrument proved to be valuable in providing comparative information from programs and their graduates. Reportedly, this survey is considered a reliable and valid instrument.

Tatto et al. (Eds.) (2012)	Policy, practice and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)	Data from case study reports from 17 countries. Data from four surveys administered to teacher education institutions and programs, teacher educators and future primary and lower-secondary school teachers (app. 22,000 future teachers from 750 programs in about 500 institutions).	Abundant information about teachers' mathematical knowledge, characteristics of TPPs, features of the education system, employment and working conditions, beliefs and teachers' opportunities to learn for each of the participating countries.
Thomas & Loadman (2001)	Evaluating teacher education programs using a national survey	Results from the National Survey of Teacher Education Program Graduates in a major Carnegie I research university. Data from a cohort of 263 baccalaureates and 171 M.Ed. graduates (the class of 1996).	Graduates were very positive towards their programs and careers. Responses in all 4 measures of the survey were more alike than different between baccalaureate and M.Ed. graduates. Use of graduate surveys is necessary to measure TPPs' quality.

Table 3
Employment Outcomes Literature Review

Author(s) (Year)	Title	Source of Data	Major Findings
Andrew (1990)	Differences between graduates of 4-year and 5-year teacher preparation programs	The study collected survey data from random samples of graduates of 4- and 5-year programs at the University of New Hampshire from 1976-1986; and yearly program evaluation questionnaires sent to students at the end of either the 4-year or 5-year program from 1981-82 through 1988-89. Number of teachers surveyed is not mentioned.	5-year TE program tends to enhance placement and retention rates.
Cameron & Lovett (2015)	Sustaining the commitment and realising the potential of highly promising teachers	The data of the study was collected for a survey of 57 teachers and interviews with 21 teachers.	The study finds that school-level practices are major contributors to the job satisfaction and organizational commitment of teachers who, early in their careers, have been predicted to make a significant contribution to teaching.
Darling-Hammond (2000)	How teacher education matters	Review of prior studies (7) comparing 4-year and 5-year programs' graduates retention rates (cited on p.170).	5-year TE programs are more likely to place and retain their graduates.
DeAngelis, Wall & Che (2013)	The impact of preservice preparation and early career support on novice teachers' career intentions and decisions	The study uses 2003-2004 survey data of 4,974 four-year undergraduate teaching degree completers from 12 public higher education institutions; and archived state administrative data from 2005-2006 and 2006-2007 academic years.	A direct association was found between perceived preparation quality and leaving teaching.

Freedman & Appleman (2009)	"In it for the long haul"—How teacher education can contribute to teacher retention in high-poverty, urban schools	The study used program-level administrative materials (e.g., background information, coursework records, and interviews) and survey data of the first five years' teaching of 26 UC-Berkley graduated novice teachers.	Substantive and coordinated preparation in theory and practice may help retain teachers in hard-to-staff (high-poverty, urban) schools
Goldhaber & Cowan (2014)	Excavating the teacher pipeline: Teacher preparation programs and teacher attrition	The study used Washington State administrative databases including teacher assessment data, teacher certification data, and school-level data on teacher assignments of all teachers who entered Washington public schools as a beginning teacher between 1989-1990 school year and 2011-2012 school year.	Teacher preparation programs may contribute to teacher placement and retention rates.
Ingersoll (2001)	Teacher turnover and teacher shortages: An organizational analysis	The study used data from the 1990–1991 Schools and Staffing Survey (SASS) (teacher and administrator questionnaires, which had a randomly selected sample); and from the Teacher Follow-up Survey (TFS) the sample of which comprises 6,733 elementary and secondary teachers.	Teacher turnover is strongly correlated with the individual characteristics of teachers.
Ingersoll, Merrill & May (2012)	Retaining teachers: How preparation matters	The study used data from the 2003–04 Schools and Staffing Survey, and the 2004–05 Teacher Follow-up Survey. Authors do not mention whether they used the whole dataset or sub-sets.	Teachers who receive less pedagogical training are more likely to leave teaching.

Kumashiro (2015)	The review of the Proposed Federal Regulations for Teacher Preparation Programs	The author searched the literature on the relationships between teacher preparation and multiple measures of outcomes. Author did not explicitly say how he searched for the literature or how many pieces his analysis resides on.	The indicator of employment outcomes inaccurately presumes that placement and retention are the result of program quality, without sufficient acknowledgment of the role of the job economy, work conditions, personal life circumstances, and preferences that can affect employment and tenure.
Nelson (2004)	Reclaiming teacher preparation for success in high-needs schools	The author uses personal experience working in inner-city schools. No sample information was included in the report.	Teaching at its best is not a codified and prescribed technical "how-to" exercise, but rather a dynamic intellectual activity; knowledge-based preparation is not enough.
Ronfeldt, Schwartz & Jacob (2014)	Does pre-service preparation matter: Examining an old question in new ways	The study uses the Schools and Staffing Survey (SASS), and Teacher Follow-up Survey (TFS). Specifically, a sample of full-time 3,237 public school teachers from the 2003–2004 and 2007–2008 SASS who were also included in TFS.	Teachers who completed more methods-related coursework and practice teaching felt better prepared and were more likely to stay in teaching.
Yost (2006)	Reflection and self-efficacy: Enhancing the retention of qualified teachers from a teacher education perspective	The study used interviews with teachers and their principals, and observations of teachers' teaching performance. The data comes from 17 teachers who graduated from a four-year, undergraduate teacher preparation program in a small, liberal arts university of approximately 3,000 students.	Teacher education programs could enhance teacher retention by fostering resiliency and persistence in teacher candidates.

Table 4
Student Learning Outcomes Literature Review

Author(s) (Year)	Title	Sources of Data	Major Findings
Boyd et al. (2006)	How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement	The study used NY and NYC administrative data between the years of 1998 and 2004 with information on students, schools, and teachers, including their pathway into teaching. The sample includes all teachers in tested grades and subjects.	Both TFA and the Teaching Fellows in NYC have a smaller impact on student achievement than those traditionally trained, although these differences disappear after the early years. The differences early on are also fairly small in magnitude. Most of the variation in effectiveness was found to be within groups.
Boyd et al. (2009)	Teacher preparation and student achievement	The study used multiple NY state administrative data sets, including demographic data for students, teachers, and schools for each year from 2000-2001 to 2005-2006; and program-level data of 31 teacher education programs across 18 institutions in NY.	The authors found variation across programs in the average effectiveness, defined by value-added to student achievement, of graduates in NY. They also found that preparation linked to practice benefits first year teachers.
Clark et al. (2013)	The Effectiveness of Secondary Math Teachers from Teach For America and the Teaching Fellows Programs	The study used an experimental design with random assignment of students to teachers with different training. They used a purposive sample spanning across 8 states. The TFA study sample included 4,573 students, 111 classroom matches, 136 math teachers, 45 schools, and 11 districts in 8 states. The Teaching Fellows study sample consisted of 4,116 students, 118 classroom matches, 153 math teachers, 44 schools, and 9 districts in 8 states.	The major findings from the study were that TFA teachers were more effective with teaching secondary math, defined by value added to student achievement, and Teaching Fellows were indistinguishable relative to the comparison group, which was either traditionally trained teachers or less selective alternative programs.

Constantine et al. (2009)	An Evaluation of Teachers Trained Through Different Routes to Certification	A purposive sample of 87 matched pairs of alternatively certified and traditionally certified teachers was used. The evaluation included 2600 students and their teachers in 63 schools and 20 districts. These included 5 districts in California, 7 districts across Illinois, Wisconsin, Louisiana, and Georgia, 3 districts in New Jersey, and 5 districts in Texas. The study used random assignment of teachers to students.	The authors found no evidence that the amount or content of coursework in teacher training was associated with teacher effectiveness. They also found no average difference between alternatively certified teachers and traditionally trained teachers as measured by their students' test scores.
Darling-Hammond et al. (2005)	Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness	The study used administrative data from Houston, Texas linking student and teacher data between the years of 1995 and 2002. The sample includes all students and their teachers in grade 4 or 5 during this timeframe.	The authors found that TFA teachers in Houston are about as effective in raising student achievement in math and reading as other traditionally trained teachers with similar experience.
Decker et al. (2004)	The effects of Teach for America on students: Findings from a national evaluation	The study used data from 2000 students and their teachers in 100 classrooms in17 schools in Chicago, Los Angeles, Houston, New Orleans and the Mississippi Delta during the 2002-2003 school year. Students in grades 1-5 were randomly assigned to TFA and non-TFA teachers.	Average math scores were significantly higher for students of TFA teachers relative to the comparison group with an effect size of about .15. The TFA teachers did not appear to have an impact on reading however. With the sample restricted to only novices, TFA teachers were even more effective in math, about a .26 effect size.
Feuer et al. (2013)	Evaluation of Teacher Preparation Programs: Purposes, Methods, and Policy Options	Literature Review	Authors recognize utility of VAM's, including the ability to control for factors external to the teacher and program's influence. In addition, such measures, they argue, can avoid the issues that exist when evaluating individual

			teachers, as these measures would aggregate to the program level. The authors also recognize the limitations in the information that VAMs can provide and highlight the need for more research in other metrics.
Floden (2012)	Teacher Value Added as a Measure of Program Quality: Interpret with Caution	Literature Review	Author highlights the limitations of VAMs in teacher education and preparation programs evaluation including the need for multiple metrics to analyze more than student achievement, whether average VAM scores are an appropriate measure of teacher education and preparation programs quality, and that where teachers are hired, and who they teach can bias estimates of program effectiveness.
Gansle et al. (2012)	Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in LA.	The study uses Louisiana state administrative data including data sets describing students, teachers, classes, and schools in Louisiana which were merged using a series of data cleaning steps.	Using value-added measures, the authors find that most programs in Louisiana are not distinguishable from one another, though few are significantly different as defined by impact on student achievement.
Goldhaber et al. (2013)	The gateway to the profession: Assessing teacher preparation programs based on student achievement	The study used five administrative databases prepared by Washington State's Office of Superintendent of Public Instruction (OSPI) including 8718 teachers (17,715 teacher-years) whose initial teacher training programs were either from one of 20 state accredited teacher preparation programs, or from outside of the state.	The study found variation in the effectiveness of graduates in Washington teacher preparation programs. While graduates of few programs are identified as differentially impacting student achievement, the effect sizes are argued to be educationally meaningful, with slightly higher effect sizes in reading than math.

Henry et al. (2012)	Incorporating Teacher Effectiveness into Teacher Preparation Program Evaluation	Literature Review	Authors identify different options for states to consider, including which tests teachers are to be held accountable for, how teachers in tested and non-tested subjects are included, and whether VAM's or more transparent methods are more appropriate for evaluation. They conclude by arguing that states need to address issues of accuracy (using multiple measures such as observation), fairness (addressing out-of-school factors statistically), transparency (ensuring stakeholders understand the methods used), and inclusiveness (measuring outcomes like graduation rates and student engagement).
Henry et al. (2014)	The Effects of Teacher Entry Portals on Student Achievement	The study used administrative data from North Carolina and included all students and their teachers in tested grades and subjects with less than five years of experience.	Using undergraduate-prepared teachers from in-state public institutions as the reference group, the authors found that TFA teachers were more effective in STEM subjects and secondary grades. They also found that other alternative entry teachers were less effective than their traditionally trained counterparts in high school, as defined by their adjusted average test score gains.
Kane et al. (2008)	What does certification tell us about teacher effectiveness? Evidence from New York City	The study used administrative data from NYC and NY and includes all students and their teachers in grades 4 through 8 who teach math or reading between the years of 1998 and 2005.	With the traditionally trained teachers as the reference group, the authors find negligible differences between both TFA teachers and Teaching Fellows. They found large differences existed within groups, similar to earlier work.

Koedel et al. (2015)	Teacher preparation programs and teacher quality: Are there real differences across programs?	The study used Missouri administrative data which includes 1,309 unique teachers who were certified from one of the 24 major preparation programs in the state.	Using value-added measures, the author finds that differences in effectiveness between teachers from different programs in the state of Missouri were very small. There was more variation within programs than between programs.
Mihaly et al. (2012)	Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates	The study used Florida administrative data from teachers who taught in an elementary school in grades 4 and 5 at some point during 2000-2004 in a dataset provided by The Florida Education Data Warehouse (FL-EDW)	Authors attempt to separate school effectiveness from teacher education and preparation program effectiveness measures. Authors argue that both school hiring processes and the type of statistical model used can create very different results in evaluating programs. Their recommendations provide caution for future analyses.
Papay et al. (2012)	Does an Urban Teacher Residency Increase Student Achievement? Early Evidence from Boston	The study used administrative data from Boston including all student and teachers in grades 4-8 in ELA and math beginning in 2004.	Boston Teacher Residency graduates with available value-added data are not more effective at raising student achievement than other novices in ELA and are less effective in math. Their effectiveness improves rapidly by years 4 and 5, however, outperforming veteran teachers.
Plecki et al. (2012)	Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures	The study used a sample of 2864 in math and 2874 teachers in readings teachers who were trained in 22 institutions in Washington state. The administrative data contains information on students, teachers, and schools.	Using value-added measures of student achievement, the authors found that few institutions were distinguishable from one another with regard to the student achievement of those taught by their graduates in reading. There were no institutions with a significant association regarding student achievement in math.

Raymond et al. (2001)	Teach for America: An evaluation of teacher differences and student outcomes in Houston, Texas	The study used administrative data from Houston, Texas linking student and teacher data between the years of 1996 and 2000. The included information for students, teachers, and schools, including the training pathway for the teacher.	The authors found positive but not statistically significant differences between TFA teachers and the average non-TFA teacher. While there was larger variation within groups of teachers with the same pathway, the variation among TFA teachers was smaller than the other groups analyzed.
Ronfeldt (2012)	Where Should Student Teachers Learn to Teach?: Effects of Field Placement School Characteristics on Teacher Retention and Effectiveness	The study used administrative and survey data on nearly 3000 first-year New York City teachers who responded to a survey in 2005 (response rate of 70%).	Authors find that future teachers trained in traditional programs with field placements in easy-to-staff schools had a positive effect on student achievement, even if the teachers ended up teaching in hard-to-staff schools once hired.
Ronfeldt (2015)	Field Placement Schools and Instructional Effectiveness	The study used a sample of 752 teachers who did their field placement in 259 different schools and were later employed by 308 different schools in an anonymous district. The sample was drawn from 3 years of survey data and 10 years of administrative data.	Authors find that teachers trained in traditional programs with field placements where there is stronger teacher collaboration, achievement gains and retention were more effective at raising student achievement once hired.
Sass (2015)	Licensure and Worker Quality: A Comparison of Alternative Routes to Teaching	The study used administrative data from Florida and included rich longitudinal data on all students and their teachers who taught math and/or reading in tested grades between 2000 and 2010.	The author analyzes three types of alternative certification in FL: district alternative certification, education preparation institute (EPI) option, and the American Board for Certification of Teacher Excellence (ABCTE) Passport. The author finds that the alternatively certified teachers typically have stronger preservice academic skills; however their effectiveness is mixed. District certified teachers'

			value added is generally only 1-2% of a standard deviation higher than traditionally trained teachers. EPI graduates are about 2-4% higher than traditionally trained. The ABCTE teachers however are about 6-8% of a standard deviation higher in math than traditionally trained teachers.
Tuttle et al. (2009)	ABCTE Teachers in Florida and Their Effect on Student Performance	The study used data on 30 ABCTE teachers in Florida over two years. The sample was derived from all ABCTE teachers in Florida and limited to those in tested grades and subjects. Comparison teachers were assigned using propensity score matching.	The authors found no difference in reading between students of ABCTE teachers and the comparison group. ABCTE teachers were less effective in math, with a large effect size of .25.
Worrell et al. (2014)	Assessing and Evaluating Teacher Preparation Programs	Literature Review	Authors were optimistic about the potential use of VAM's in an evaluation context, providing careful recommendations for its uses alongside student learning objectives, surveys and observation instruments during the progression of a program, at its completion, and at post-graduation.
Xu et al. (2011)	Making a difference? The effects of Teach For America in high school	The study used administrative data from North Carolina including rich student and teacher data between the years of 2000 and 2007. The sample was limited to 23 local education agencies who hired at least one TFA teacher to a high school during the time period.	The authors find that TFA teachers are more effective at raising student test scores than traditionally trained teachers in North Carolina high schools, particularly in science.

About the Authors

Maria Teresa Tatto
Michigan State University
mttatto@msu.edu
http://orcid.org/0000-0003-4955-1420
Maria Teresa Tatto, Ed.D., is an Associate Professor of Educational Policy, Curriculum, Instruction, and Teacher Education at Michigan State University. Her research is characterized by the use of an international-comparative framework to study the impact of educational reform on educational systems, particularly teacher education, teaching, and learning. Her other research interests include the influence of early childhood education on improved knowledge levels for children living in disadvantaged communities, the development of effective policies to support the education of children of migrant workers in the U.S., and the role of values education on citizenship formation. Her work combines the use of quantitative and qualitative approaches and methods. She has done research in Mexico, Sri Lanka, and several countries in Latin America and has served as a consultant to the World Bank and USAID for the governments of the Dominican Republic, Columbia, Guatemala, Mexico and Peru. She is currently the director and principal investigator of two large scale international studies funded by the U.S. National Science Foundation, the Teacher Education and Development Study in Mathematics, or TEDS-M, and FIRSTMATH a study of the characteristics of novice teachers who are effective teachers of mathematics in primary and secondary school across several contexts and countries.

Corey Savage
Michigan State University
csavage@msu.edu
Corey Savage is a PhD student in educational policy at Michigan State University. His research interests focus on issues of teacher quality generally. Currently, his work focuses on disentangling opportunities to learn in teacher training and analyzing their impacts on a variety of outcomes.

Wei Liao
Michigan State University
liaowei@msu.edu
Wei Liao is a PhD student in the department of teacher education at Michigan State University. His research revolves around comparative and international education with an emphasis on the relationship between educational policy and teacher quality in China and the United States.

Stefanie L. Marshall
Michigan State University
marsh413@msu.edu
Stefanie Marshall is a PhD student in educational policy at Michigan State University. Her research focuses on the role of school principals in implementing the Next Generations

Science Standards. Stefanie's research interests also include issues around race and equity in science education.

Paul Goldblatt
Michigan State University
goldbla7@rhs.msu.edu
Paul Goldblatt is a PhD candidate in educational policy at Michigan State University. His area of interest is in student success and retention, focusing on students of color.

Leonardo M. Contreras
Michigan State University
medelleo@msu.edu
Leonardo Medel Contreras is a PhD student in the Program in Mathematics Education (PRIME) at Michigan State University. His research focuses on teachers' mathematical content and pedagogical knowledge. He is also interested in educational policy and how it relates to social processes and equity.

education policy analysis archives

Please contribute commentaries at http://epaa.info/wordpress/ and send errata notes to Gustavo E. Fischman fischman@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.