Evaluating Teacher Preparation Programs:
A Response to the Federal Proposal to Evaluate Teacher Preparation Programs

University Council for Educational Administration
Michelle Young, Executive Director
Ed Fuller and Sheneka Williams, Associate Directors for Policy

The intent of the federal government's proposed legislation is to improve the quality of teacher preparation programs (TPPs) by identifying the effectiveness of preparation programs through the use of various outcome measures. These outcome measures include:

- Effectiveness of graduates in improving student outcomes;
- Employment outcomes, including placement and retention rates of graduates;
- Perceptions of program effectiveness based on survey outcomes from newly hired teachers and their immediate supervisors; and,
- State or CAEP accreditation.

Moreover, states will be required to rate every preparation program—including alternative, non-university based programs—using four categories: exceptional, effective, at-risk, and low-performing.

Major Issues with the USDoE Proposal

There are numerous issues associated with the USDoE's proposal that we review below. There are, however, three over-arching major issues with the proposal: (1) underestimation of fiscal implications and the requisite technical capacity to implement the proposal; (2) lack of evidence of the efficacy of proposal; and (3) lack of a theory of action of how the proposal will improve teacher quality.

1) Fiscal Implications and Technical Capacity

The USDoE dramatically underestimates the fiscal implications of the proposal. States and programs will have to collect data that, in many instances, is currently not collected. At the state level these costs are quite large as evidenced by the approximately $575 million USDOE investment in the Statewide Longitudinal Data Systems (SLDS) grants[i] to improve state data collection systems. Yet, the USDOE proposal for evaluating teacher preparation programs actually notes that only nine states "currently link K-12 teacher data including data on both teacher/administrator evaluations and teacher preparation programs to K-12 student data"[ii] and only 30 states will have such a linkage in the coming years through the SLDS grants. Given that 20 states do not have this linkage and the obvious substantial fiscal investments incurred to establish the linkage in 30 states, there will be substantial costs to establish the linkage in the remaining 20 states.

Further, the USDoE proposal would necessitates the collection of numerous additional variables far beyond the scope of current state data collection systems. This will require all 50 states will need to create new data systems that collect the required information, regardless of whether the state already has an established link between TPPs, graduates, and K-12 student test scores.

Moreover, TPPs will need to create new data collection systems, hire new personnel to create and maintain these systems as well as analyze the data. Further, numerous meetings will need to occur to inform the creation of such data systems. All of this will require substantial fiscal expenditures by TPPs.

Thus, the proposal will be an unfunded mandate that will place a huge fiscal strain on state education agencies (SEAs) and TPPs, many of which are already suffering from a lack of funding and the technical capacity to engage in this work. With respect to SEA capacity, we believe the majority of SEAs do not have the fiscal or technical capacity to implement the proposal in a manner that will provide accurate information about the quality of TPPs. This is not the fault of SEAs, but a result of continued under-funding of SEAs over the last five years and the monumental costs for collecting, analyzing, and reporting the data mandated by USDoE.

2) Lack of Evidence on Efficacy of Proposed Approach

To date, no state has adopted and implemented the set of standards proposed by the USDoE. Thus, we have no evidence about the efficacy of the proposal to improve the quality of TPPs. We also do not know the unintended consequences such a proposal may elicit. Indeed, the well-intentioned adoption of NCLB had numerous unintended consequences[iii] that could not be systematically addressed because the issues were only apparent after the institutionalization of the system. Without piloting the proposal, we simply cannot predict the efficacy or unintended outcomes of the proposal. Given the high-stakes nature of the proposal and the potential damage the proposal could have in terms of the supply of well-prepared teachers to lower-performing schools, the proposal should be piloted without high-stakes until the outcomes can be properly assessed.

For example, in his review of the national evaluation of TPPs by the National Center for Teaching Quality (NCTQ), Fuller[iv] revealed how extremely few research studies cited by NCTQ directly connected the NCTQ metrics with TPP outcomes. NCTQ, in fact, admitted their metrics had very little empirical support. With so little actionable evidence connecting TPPs—especially specific TPP strategies or policies—to outcomes in the USDoE proposal, it is unclear as to why the proposal is being suggested for implementation across the country rather than in some specific states as a pilot study to better understand these mechanisms.

3) Lack of a Theory of Action

Not only is there an extreme paucity of research that would support the wide-scale adoption of the proposal, the USDoE and other proponents of high-stakes accountability systems of TPPs do not even provide a viable theory of action about how the system would spur TPP improvement or teacher quality.

One possibility might be that publication of accountability results might influence the decisions of prospective students. Theoretically, lower performing programs would experience a decline in applicants over time while high performing programs would see an increase in applicants. This, in turn, would spur TPPs to change behaviors in order to improve outcomes and recapture market share. This assumes, however, that prospective students make decisions primarily based on measures of program quality. Research from other fields suggests this may not always be the case.

A second possibility is that the closure of low-performing programs would increase the overall quality of teachers by eliminating poorly prepared teachers from the pool of applicants. Yet, research consistently finds greater variation within TPPs than between TPPs. Thus, closing low-performing programs may only eliminate a very small number of teachers that would be rated as ineffective using student test scores.

A third possibility could be that the public sharing of results and the high-stakes nature of the system would create a strong incentive for TPPs to create an outcome-oriented culture that results in the adoption of strategies

shown to be effective in improving outcomes. A related possibility is that the technical assistance component of the proposal would provide feedback, support, and assistance that would improve the outcomes of TPPs.

One serious issue with the second through fourth possibilities is that the proposal assumes there is a robust body of knowledge about the TPP strategies and policies associated with eliciting positive outcomes. This assumption, when connected to the USDoE's characterization of the majority of TPPs as mediocre, leads to another apparent USDoE assumption: individuals in TPPs necessarily choose to ignore existing evidence about effective practice and, instead, continue using ineffective practice. A careful read of the literature in the area of TPP effectiveness reveals a paucity of research that finds direct connections to specific policies, procedures, or behaviors the consistently lead to positive outcomes for all teachers in all contexts. This directly contradicts the assumptions serving as the foundation of the USDoE proposal.

Technical Issues

In the fo0llowing sections, we further examine the issues associated with the federal government's proposals and outcome measures. The issues are divided into five sections. The first four sections address the first four metrics included in the proposal.

I. Measuring Program Effectiveness Using Graduate Effectiveness

Measures of Effectiveness are Inaccurate, Especially at Program Level

There is still debate about the accuracy of measures of student growth and the validity and reliability of the inferences made from such estimates.[v] Even advocates of the use of metrics acknowledge that there is still a great deal to learn about the utility of such measures, particularly around identifying effective teachers and TPPs.[vi] This is especially true for states using student growth percentiles (SGPs)—a measure specifically designed to be descriptive in nature rather than as an estimate to be used in attributing causality to various factors.[vii] Even when sophisticated VAMs are employed, there is currently no compelling evidence that TPP effectiveness can be estimated independently of other factors that affect student achievement.[viii] Despite several studies from Washington State, Missouri, Florida, and North Carolina, researchers have yet to establish the validity and reliability of such efforts. In fact, some studies have found that such efforts do not yield accurate comparisons of programs.[ix] Further, studies of TPPs generally find greater variation *within* programs than *between* programs.[x] Thus, closing a small number of programs will have, at best, only a minimal impact on teacher quality. Further, given there is no evidence on the ability of TPPs to predict graduate effectiveness—particularly when the outcome measures are inaccurate—TPPs would have little guidance on how to improve graduate effectiveness even if the student growth measures were accurate.

Measures of Effectiveness Would Expand Data Collection and Expand Testing

The proposal uses higher education policy rules to essentially force states to expand their testing of students and/or expand the collection of data on student outcomes. This is a backdoor strategy to implement policies started under NCLB and expanded under NCLB waivers to assess all students in every grade and subject area. This expansion would cause states to incur huge financial costs in developing and implementing such systems as well as for collecting and analyzing the data. Further, there is no clear evidence that the expansion of K-12 testing would

have only positive effects on student achievement[xi] and could have negative effects on student non-cognitive outcomes.

State Student Growth Measures are Biased against Teachers in Certain Schools and of Certain Types of Students

Most states are using growth measures that do not control for student background characteristics other than prior test scores.[xii] Research is *extremely* clear about the outcomes of analyses that exclude student personal background characteristics—the results are biased against teachers in classrooms and schools with certain types of students (typically poor, special education, and ELL students).[xiii] Thus, when aggregated to the program level, the results will be biased against programs placing high percentages of graduates into high-need schools that serve higher percentages of poor, special education, and ELL students. When such biases become apparent to educators, there will be a clear incentive for new teachers to avoid such classrooms and for TPPs to recommend that graduates avoid taking positions in such schools.

II. Measures of Employment

Lack of Solid Research Foundation Linking Program Quality with Employment Outcomes

There is only one peer-reviewed research paper that examines placement rates by individual programs[xiv] and only one peer-reviewed article that examines and reports retention rates by individual programs.[xv] Both studies employed data from the state of Washington--a geographically small, relatively homogeneous state. Thus, it is not representative of the other 49 states in the nation. In addition, both studies encountered difficulty in accurately assessing program quality based on these metrics due to missing data and other factors. The important point is that there is no credible body of research that links individual TPPs to either placement or retention and certainly no body of research that links TPP practices and strategies to employment outcomes. Thus, states will be unable to provide reliable technical assistance to any programs judged to be low-performing in this area.

Employment Measures will be Inaccurate

Currently, states that already report employment measures rely on simple percentages of graduates obtaining employment and remaining in teaching. This is problematic given that a multitude of other factors outside the control of TPPs influence placement and retention rates.[xvi] Note that the two studies by Goldhaber and his colleagues that have examined placement and retention issues at the individual TPP level employed highly sophisticated statistical procedures in an attempt to isolate the effect of programs apart from other factors outside the control of programs. The authors did so because they clearly understood that simple percentages would not provide accurate assessments about TPP effectiveness relative to placement and retention.

Given that the USDoE will not be providing the fiscal or human capital necessary to use the sophisticated techniques necessary to accurately identify TPP effectiveness, states will likely default to reliance on simple percentages. This reliance will create incentives for TPPs to attempt to game the system in ways that have unintended consequences. For example, TPPs could focus on recruiting White individuals, younger individuals, and females to increase placement and retention percentages. TPPs could also recommend to graduates that they seek employment in high-performing schools, particularly high-need schools with high levels of performance. The incentives created through the use of simple percentages could seriously erode the precious little progress states have made in diversifying their teaching workforces.

III. High-Need Schools

While the emphasis on serving high-needs schools is laudable, the metrics used to identify high-needs schools is problematic. High-needs schools are not always low-performing or hard-to-staff schools.[xvii] If the purpose of including the high-needs metric in the proposal is to create a more equitable distribution of teachers and, hence, reduce the achievement gap between economically disadvantaged and not economically disadvantaged students, the reliance on a simplistic metric identifying high-needs schools will be inadequate to accomplish that purpose. Other factors such as working conditions, in fact, have a much stronger association with hard-to-staff schools than student demographics.[xviii]

If the identification of high-needs schools is based only on the percentage of economically disadvantaged students, then there will be an incentive for programs to place graduates in high-performing schools identified as high-need schools—thus subverting the goal to close the achievement gap. Therefore, the proposal should use other or, at least additional, measures to identify high-needs if the USDoE wants to address equity issues in placing and retaining effective teachers in all schools.

IV. Surveys of Newly Hired Teachers and Supervisors

There is no research on the reliability, validity, or accuracy of the inferences made from such measures. Further, simple percentages will ignore the fact that graduates are not randomly distributed across schools and that innumerable other factors influence the degree to which graduates perceive their training as effective. The use of simple percentages of graduates' perceptions of TPP efficacy without adjusting for other factors and establishing the necessary psychometric properties violates the basic program evaluation standards set forth by the Joint Committee on Standards for Educational Evaluation.

V. Other Issues

Judgments about Low-Performing Programs Should Include Site Visits

Data that can be collected at the state level will never be sufficient to make accurate judgments about programs. Because each program is unique and labor markets may differ significantly within states, site visits are a necessary component of any evaluation system that will make high-stakes judgments about programs. The visits associated with CAEP will not be sufficient because of the long time-span between accreditation visits. Thus, state accreditation should include site visits when a program is designated as "low-performing."

Weighting Decisions Should be Determined after Data are Collected and Analyzed

The weighting of the components is going to be critical and the USDoE provides no guidance on how each component should be weighted except to state that employment and retention in high-needs schools should be considered "in significant part." Weighting of the components is a critical decision, particularly with respect to the weighting of high-needs schools. If placement in high-needs schools has a relatively low-weight and student growth is negatively associated with the percentage of economically disadvantaged students enrolled in the school, then programs may game the system by choosing to counsel students to seek employment in non-high needs schools.

Adopt High Stakes Consequences after Phase-in

Because errors will be made in the calculation of data and in the determining the weights associated with various factors, states should be required to calculate, analyze, and publish the data for at least two years before

high-stakes consequences are attached.[xix] This will ensure initial unintended consequences are identified and addressed before the consequences have high-stakes for programs.

Adjusting Measures

There should be a mechanism to adjust results when schools close or school boundaries change. Programs with smaller numbers of graduates concentrated in particular schools could be significantly impacted by these changes that are clearly outside the control of programs.

Alternative versus Traditional Programs

Given that employment in a teacher-of-record position is generally a prerequisite for enrollment in an alternative preparation program, placement and retention rates for alternative programs should be different than for traditional programs. Retention rate calculation should start in year two for alternative programs. In other words, employment in year one during the probationary period should be the base year and retention should be calculated as the percentage of graduates initially placed in year 1 that return for year 2, year 3, etc.

Conclusion

UCEA strongly believes in holding educator preparation programs accountable, and supports a variety of ways to do so. However the USDoE needs to move away from using high-stakes judgments with attached punishments. High-stakes accountability systems that rely on questionable measures of performance have not spurred program improvement or led to authentic reforms. Instead we encourage the USDoE to consider more thoughtful and evidence-based approaches to accountability, and to invest its scarce resources in capacity building, supportive technical assistance, and guidance.

---

[i] See http://nces.ed.gov/programs/slds/stateinfo.asp

[ii] Office of the Federal Register (2014). *Teacher preparation issues.* Retrieved Dec. 28, 2014, from https://www.federalregister.gov/articles/2014/12/03/2014-28218/teacher-preparation-issues.

[iii] Numerous studies have identified multiple unintended consequences of NCLB—particularly with respect to the mandated state testing associated with NCLB. A few of these studies include the following:

Hollingworth, L. (2008). Unintended educational and social consequences of the No Child Left Behind Act. *J. Gender Race & Just.*, *12*, 311.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of 'No Child Left Behind'. *Race Ethnicity and Education*, *10*(3), 245-260.

Perna, L. W., & Thomas, S. L. (2009). Barriers to college opportunity the unintended consequences of state-mandated testing. *Educational Policy*, *23*(3), 451-479

[iv] Fuller, E.J. (2013). Shaky methods, shaky motives: A Critique of the National Council of Teacher Quality's review of teacher preparation programs. *Journal of Teacher Education 65*: 63-77.

[v] See, for example, the following:

American Statistical Association (2014, April 8). *ASA statement on using value-added models for educational assessment.* Retrieved Dec. 28, 2014, from http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf;

Baker, E.L., Barton, P.E., Darling-Hammong, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., Shepard, L.A. (2010) Problems with the Use of Student Test Scores to Evaluate Teachers. Washington, DC: Economic Policy Institute. http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf

Floden, R. E. (2012). Teacher Value Added as a Measure of Program Quality Interpret With Caution. *Journal of Teacher Education*, *63*(5), 356-360.

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.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35-62.

National Research Council and National Academy of Education (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press. Retrieved Dec. 28, 2014, from http://www.nap.edu/openbook.php?record_id=12820.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *education policy analysis archives*, *18*(23), n23.

Papay, J. P. (2011). Different Tests, Different Answers The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, *48*(1), 163-193.

[vi] Kennedy, M. (Ed.). (2010). *Teacher assessment and the quest for teacher quality: A handbook*. John Wiley & Sons.

[vii] Baker, B. D., Oluwole, J., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Evaluation and Policy Analysis Archives*, *21*, 1-71.

[viii] Floden, R. E. (2012). Teacher Value Added as a Measure of Program Quality Interpret With Caution. *Journal of Teacher Education*, *63*(5), 356-360.

Feuer, M.J., Floden, R., Chudowsky, N., Ahn, J. (2013*). Evaluation of teacher preparation programs: Purposes, methods, and policy options*. Washington, DC: National Academy of Education.

[ix] 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*, *8*(4), 459-493.

[x] Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012). Teacher preparation programs and teacher quality: Are there real differences across programs. *Education Finance and Policy*.

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.

[xi] Hamilton, L. (2003). Assessment as a policy tool. *Review of research in education*, 25-68.

[xii] Collins, C., & Amrein-Beardsley, A. (2013). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1)

Fuller, E.J., Hollingworth, L., & Liu, J. (in press). A Fifty-State Analysis of Efforts to Evaluate Principals. *Journal of Research in Leadership Education*.

[xiii] Baker, B. D., Oluwole, J., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Evaluation and Policy Analysis Archives*, *21*, 1-71.

[xiv] Goldhaber, D., Krieg, J., Theobald, R. (in press). Knocking on the Door to the Teaching Profession? Modeling the Entry of Prospective Teachers into the Workforce. *Economics of Education Review*

[xv] Goldhaber, D., & Cowan, J. (2014). Excavating the Teacher Pipeline Teacher Preparation Programs and Teacher Attrition. *Journal of Teacher Education*, *65*(5), 449-462.

[xvi] Feuer, M.J., Floden, R., Chudowsky, N., Ahn, J. (2013*). Evaluation of teacher preparation programs: Purposes, methods, and policy options*. Washington, DC: National Academy of Education.

[xvii] Opfer, D. (2011). Defining and Identifying Hard-to-Staff Schools The Role of School Demographics and Conditions. *Educational Administration Quarterly*, *47*(4), 582-619.

[xviii] Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*(10), 1-39.

[xix] Lincove, J. A., Osborne, C., Dillon, A., & Mills, N. (2013). The politics and statistics of value-added modeling for accountability of teacher preparation programs. *Journal of Teacher Education*, 0022487113504108.