

To: Members of the Negotiated Rulemaking Committee on Teacher Preparation Programs

From: Cheryl Achterberg, Dean of the College of Education and Human Ecology,
The Ohio State University
Camilla Benbow, Patricia and Rodes Hart Dean of the Peabody College of Education
and Human Development, Vanderbilt University
Mary M. Brabeck, Gale and Ira Drukier Dean of the Steinhardt School of Culture,
Education, and Human Development, New York University
Jane Conoley, Dean, The Gevirtz Graduate School of Education, University of
California Santa Barbara

CC: Members of the LEARN Coalition

April 2, 2012

We are members of the Learning and Education Academic Research Network (LEARN), a national coalition of leading, research-intensive colleges of education. LEARN members advocate for Federal funding of research on education and public policy to drive improvements in teacher and student performance.

In his recent budget, President Obama requested a number of program consolidations, repurposing of existing program funding, and new programs to improve teacher preparation and recognize and reward high-quality teacher preparation programs. In addition, through the negotiated rulemaking process, the Administration is trying to use its executive authority to effect changes in existing rules and regulations for the teacher preparation (and related) programs it administers, again to meet its improvement goals.

As deans of schools and colleges of education in research intensive universities, we appreciate these efforts to improve teacher preparation, and by extension, the education of US children and youth. However, we understand that some of the changes under consideration in the rulemaking process are significant and will have far-reaching and long-standing consequences. As a result, we believe researchers at our institutions, who have knowledge and experience in rigorous education research can add significant expertise to this debate. Furthermore, we believe that the most effective improvements to teaching and learning in K-12 education are those informed by strong research. We urge you to consider the views outlined in this letter and ask that you take them into account in developing regulations that will meet the President's stated goals.

There are a number of policy issues under discussion in the U.S. Department of Education's (ED) current negotiated rulemaking deliberations on teacher preparation programs. Our focus here is on the research that should inform the components of teacher education program evaluation that are being discussed. However, we note that research is also

relevant to the larger policy discussion and we encourage efforts to use research results to inform the following policy issues:

1. ED proposes to direct *how* the states evaluate and approve professional preparation programs. Historically, this authority has been located in the states. ED should take the necessary time to engage the broader research community in research-informed debate before making significant changes to current laws or regulations.
2. ED proposes to link student eligibility for federal aid to states' evaluation of professional programs. Financial aid should be based on financial need, and possibly merit, but not linked to the program in which students enroll, if the programs are otherwise approved for Title IV aid.
3. ED proposes to mandate that states implement costly assessments of programs, without consideration for how they would be funded. Changes in legislation should have a funding plan or a national program for development of reliable and valid assessments to assess programs.

There is a relevant research base for commenting on the proposed evaluation components of teacher education programs, which we present in some detail below.

In addition to previous reporting requirements (e.g., pass rates on state tests, characteristics of candidates, clinical preparation, and content and pedagogical skills, etc. required in accreditation reports), the proposed changes in Title II reporting systems would require States to report (and programs to demonstrate) the following:

1. Survey results regarding satisfaction of employers of graduates.
2. Survey results regarding satisfaction of graduates.
3. Numbers of graduates serving in high need (as determined by the states) schools.
4. Numbers of graduates in high need academic areas (e.g., STEM, SPED, ELL).
5. Assessment of K-12 learning (value added scores of the K-12 students taught by graduates from the program).
6. Assessment of candidates' performance

The Administration's efforts to ensure that teacher preparation programs are effective is the right goal. We believe the emphasis on outcomes is the right focus for program evaluation. However, educational research must inform the development of the metrics and indicators that will be used to decide on quality and value. In the interest of an informed discussion, we present that research below, and are prepared to add additional supporting materials to our recommendations.

1. and 2. Surveys of Graduates and Employers

Most accreditation and state approval processes require evidence of graduate and employer satisfaction. However, there are a number of practical problems that are entailed in the use of surveys. These measures differ widely across programs, institutions, and

states and, therefore do not provide comparative data. Reliable surveys take significant time and expense to develop and implement to ensure that results are valid.

All data used for consequential decisions by states and the Federal government for determining the allocation of funds to students, should be held to the highest standards of scientific rigor. The rigor of the data should be judged by standards expressed in the *Standards for Educational and Psychological Testing* developed jointly by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education (<http://www.apa.org/science/programs/testing/standards.aspx>).

3. and 4. Placements and Retentions

On the face of things, it makes sense to encourage efforts to fill positions where there are shortages, and STEM, SPED, and ELL all have shortages. However, teaching is a national labor force issue and measuring that labor market and attaching those measurements to teacher preparation programs for the purposes of judging quality can result in faulty inferences about program quality. Universities do not control where graduates choose to work. Despite extensive efforts to link our alumni to their alma maters, our employment data are all voluntary and self-reported. In a recent report to Congress on Title II (http://www.americanprogress.org/issues/2012/01/pdf/teacher_preparation_execsumm.pdf) Secretary Duncan acknowledged that 21% of all teachers in the country are working in states other than where they received their initial teacher certification. It may be more effective to seek employment data through the Social Security Agency or Internal Revenue Service, and not the institutions of higher education.

5. Value Added Modeling (VAM)

The ED proposes to use VAM of the achievement of students who were taught by teachers from specific programs to evaluate the programs. VAM is a useful measure for a number of reasons and many of our institutions use it for formative assessment of programs. VAM is a better measure of effective teaching than average scores or a comparison of this year's scores with last year's because classroom composition varies so much from year to year. VAM is a more "objective" measure of teacher effectiveness than principals' or master teachers' evaluations, which are often based on individuals' observations of teaching. VAM can be an important tool in program evaluation, but it must be computed in a valid and reliable way (Corcoran, 2010; Ewing, 2010).

Some limitations to VAM are practical. First, one needs standardized test scores over a number of years to be able to compute it. Since states are revising their standardized tests to conform to the new Common Core Standards, it will be years before reliable scores will be available for VAM analysis on these new measures. Second, most states currently only have math and literacy standardized test scores and can compute value added only in those subject areas. There are plans in many states to add science tests, but it will take states a while to develop the tests, and states will still not assess many other areas, such as social studies/history or the arts. And states do not assess all grades. Finally while Louisiana and Tennessee have built the systems for such analyses, and all eleven Race to the Top states are building them, others have much to do before the systems can compute VAM scores.

Assessing literacy and math gains puts enormous pressure on those teachers and their school leaders to raise the scores, and they may do so at the expense of other important subjects. Jennings and Rentner (2006) reported that as much as 71% of schools were reducing the time spent on subjects not tested, like social studies, geography, art, and foreign languages, never mind the 21st century skills of co-operation, problem finding, and problem solving. Unfortunately, such pressure also leads to more cheating on tests as results from Washington D.C., Atlanta, Houston, and Los Angeles, are showing (e.g., <http://www.theatlantic.com/national/archive/2012/03/investigation-finds-suspicious-achievement-in-schools-across-the-nation/255105/>; <http://www.ajc.com/news/cheating-our-children-suspicious-1397022.html>).

Researchers, psychometricians, and statisticians (e.g., National Academy of Sciences Board on Testing and Assessment, <http://www7.nationalacademies.org/BOTA/>; John Ewing, President of Math for America, <http://www.ams.org/notices/201105/rtx110500667p.pdf>; Sean Corcoran, 2010, <http://annenberginstitute.org/publication/can-teachers-be-evaluated-their-students%E2%80%99-test-scores-should-they-be-use-value-added-meas>) urge caution in using VAM to evaluate the effectiveness of individual teachers. To date, we are not aware of any research that has validated VAM measures for the purpose of program evaluation. VAM scores of individual teachers are highly unreliable in the sense that teachers' ratings are only moderately correlated from one year to the next. Teachers deemed "effective" one year are often found to be "ineffective" the next (Corcoran, 2010). This is particularly true for teachers in their early years of teaching and since research shows that teachers improve significantly in the first four years (Henry, Fortner, & Bastian, 2012), early career teachers should not be included in VAM studies for program evaluation. Scores are also subject to bias: that is, student progress (or lack thereof) due to other schools or out-of-school factors can be inappropriately attributed to the teacher. At the secondary level, where students are taught by a team of teachers, a teacher may be penalized by VAM for having ineffective colleagues.

VAM models were not designed to evaluate teacher preparation programs. While a case can be made for using VAM at the classroom level (particularly in self-contained classrooms), the logic does not extend to teacher training. At the classroom level, VAM attempts to account for differences in incoming student populations. Any exceptionally high or low gains made after accounting for these differences is considered "value added" by the teacher. In contrast, VAM models cannot account for the selection of candidates into teacher training programs, or the non-random patterns of placement by training programs into schools. As a result, VAM is unable to provide the "value added" of a teacher training program.

Finally, the VAM approach does not provide useful information about what practices, characteristics, dispositions, skills, and attitudes of teachers lead to the better test scores. All we know is that some teachers have students who get better scores, and some do more poorly. We do not know, from a VAM analysis, what teachers did that lead to these results. We need a great deal more research across the K-12 years and subjects that are taught to know what teachers do that leads to the best learning outcomes, and we need valid and

reliable observation measures to assess this. Observation measures is covered in the next section.

7. Observations of Teaching Skills. We know more today about teacher behaviors that are associated with student achievement as measured by standardized tests than 2 years ago. A recent study, funded by the Gates Foundation, *Measures of Effective Teaching* (2010; 2012) (MET), critically examined the 5 observational tools that are most extensively being used to assess what effective teachers do. The researchers used 7,491 videos of 1,332 teachers who each taught between 4-8 lessons. The videotaped classes were rated by 900 trained raters using the 5 teacher observation measures. Preliminary findings of the MET study are instructive. They also indicate a great deal more research and development of additional observation tools that are needed.

First, all 5 measures were associated with student achievement gains, using value added modeling of standardized test scores in mathematics and literacy, and two assessments to supplement the state test results: the *Balanced Assessment in Mathematics* and the open-ended version of the *Stanford 9* reading test. Other subjects like science, history, and the arts were not studied and only teachers and their students in grades 4-8 were included. Thus, we do not have valid and reliable observation tools for other grades or disciplines.

Second, teachers did better on tasks like classroom management and time management than on higher order skills like questioning, analysis, and problem solving. This information is useful for those who prepare teachers, and suggest where we should put our efforts. However, it is not clear what data will be required for candidates to “pass” the evaluation of teaching skills and strategies. The science of determining cut scores (the score that separates pass/fail) suggests setting the criteria is essential and complex, and the history tells us cut scores vary remarkably in the states (National Mathematics Advisory Panel, 2008). If there are a variety of cut scores used, data across states will not be comparable.

Third, researchers (MET, 2012) reported that the combination of scores on the observation measures, student feedback, and student achievement gains (value added) are better in predicting student performance than graduate degrees or years of teaching experience. These findings contradict other research and beg for programmatic study of the kinds of graduate experience (and undergraduate experience) that promote student achievement and the types of schools that promote continual teacher development. However, the authors caution that reliable ratings of a teacher’s practice require multiple observations and suggest that any high-stakes decisions should combine observations of teachers with student achievement data and student feedback.

Finally the authors of the MET study emphasize training in the proper use of the observation tool, impartial second raters for at least some of the data, and checks on the accuracy of those who are using the measure:

"Even with systematic training and certification of observers, the MET project needed to combine scores from multiple raters and multiple lessons to achieve high levels of reliability. A teacher's score varied considerably from lesson to lesson, as well as from observer to observer. For four out of five observation instruments, we could achieve reliabilities in the neighborhood of 0.65 only by scoring four different lessons, each by a different observer." (MET, 2012b p. 8)

This is an expensive assessment system and funding sources need to be indentified before policy changes are implemented.

In conclusion, there is a clear national interest in developing the best teachers for our students. Deans at research-intensive teacher education programs welcome the challenge to be part of this important nation building activity, and are eager to bring our knowledge and experience to the discussion about how to achieve valid and reliable evaluations of teacher preparation programs.

For all of the reasons outlined above, we do not think it is advisable at this time for the Committee to make such sweeping changes.

Sincerely,



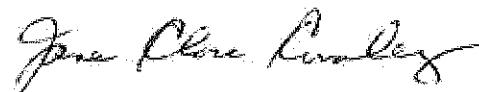
Cheryl Achterberg
Dean, College of Education
and Human Ecology
The Ohio State University



Camilla P. Benbow
Dean, Peabody College of
Education and Human Development
Vanderbilt University



Mary Brabeck
Dean, Steinhardt School of Culture,
Education, and Human Development
New York University



Jane Close Conoley
Dean, Gevirtz Graduate
School of Education
University of California Santa Barbara

References

Corcoran, S. (2010). *Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value Added Measures of Teacher Effectiveness in Policy and Practice*. Brown University: Annenberg Institute of School Reform.

Henry, G. T., Fortner, C. K., & Bastian, K. C. (2012) The effects of experience and attrition for novice high-school science and mathematics teachers. *Science*, 335.

MET (2010a). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. Policy Brief*. http://www.metproject.org/downloads/Preliminary_Finding-Policy_Brief.pdf Posted March 25, 2012.

MET (2010b). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. Research Paper*.
http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf Posted March 25, 2012.

MET (2012a). *Gathering Feedback for Teaching: Gathering High Quality Observations with Student Surveys and Achievement Gains. Policy and Practice Brief*.
http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

MET (2012b). *Gathering Feedback for Teaching: Gathering High Quality Observations with Student Surveys and Achievement Gains. Research Paper*.
http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

National Mathematics Advisory Panel (2008). *Foundations for Success*, Vol. 2, Reports of the Task Groups and Sub-Committee. pp. 8-6 - 8-17. U.S. Department of Education.

