Reforming Teacher Preparation and Licensing: What Is the Evidence?

DALE BALLOU
University of Massachusetts, Amherst

MICHAEL PODGURSKY
University of Missouri, Columbia

Using professional self-regulation in medicine as a model, the National Commission for Teaching and America's Future has proposed sweeping changes in the way teachers are trained and licensed. The commission claims that these reforms are well-grounded in a strong base of research. However, a balanced reading of the literature finds far less support for these reforms than the commission has claimed. In many cases the research is misrepresented. Since the commission's proposals would transfer considerable regulatory power out of the public domain into private education organizations, the burden of proof is on the commission to make a convincing case that such changes promote the welfare of the public and not just the interest of the profession. This burden has not been met.

INTRODUCTION

In 1996 the National Commission on Teaching and America's Future (NCTAF) issued a report on the teaching profession that proposed far-reaching changes in the way the nation prepares, licenses, and recruits teachers. Entitled What Matters Most: Teaching for America's Future, the report charged that public schools employ large numbers of unqualified teachers, largely as a result of inadequate and poorly enforced standards for teacher training and licensing. The commission set out a detailed and ambitious policy agenda to "professionalize" teaching, shifting the power to regulate teacher training and licensing from public officials to private professional organizations.

The organization behind the report—the National Commission—is a twenty-six member private body chaired by Governor James Hunt of North Carolina and funded by the Rockefeller and Carnegie foundations. The membership of the commission comprises representatives of various education interest groups, including the presidents of the National Education Association, the American Federation of Teachers, the National Council for
the Accreditation of Teacher Education (NCATE), and the National Board for Professional Teaching Standards (NBPTS). This organization has been very influential in framing policy discussions of teacher training and quality at the federal and state levels. Key components of its agenda include the following:

1. Mandatory accreditation of all teacher training programs by the National Council on Accreditation of Teacher Education (NCATE);

2. Certification of 105,000 “master” teachers by the National Board for Professional Teaching Standards, on the basis of standards developed by the Board for what teachers “should know and be able to do”; and

3. Establishing independent professional boards in all states to control teacher licensing.

Although the commission leaves many of the details of reform to various councils and professional bodies, What Matters Most makes it clear that the NCATE seeks to increase course work and other preservice training before prospective teachers can enter the classroom. The commission approves five-year programs (as opposed to the conventional four-year undergraduate degree) and applauds states that require new teachers to obtain a master’s degree. It disparages reforms that reduce the amount of preservice training in order to streamline entry into the profession (as in many alternative certification programs). The pedagogical training teachers receive ought to reflect “state-of-the-art practices,” “incorporating new knowledge” and an evolving “knowledge base for teaching” that makes clearer than ever before just what teachers should be doing in the classroom.

In 1997 the NCATE issued a second report entitled Doing What Matters Most: Investing in Quality Teaching (DWIM) in which it describes reform efforts underway throughout the states. In both reports, but particularly the latter, the commission makes strong claims about an education research literature that, in its view, offers extensive and well-documented support for its policy recommendations.

In an earlier paper on the commission’s first report, What Matters Most, we pointed out numerous discrepancies between the commission’s claims and what the research literature actually said (Ballou & Podgursky, 1997). Studies cited by the commission frequently contained statements at variance with the conclusions drawn by the NCATE. Even where there were not such discrepancies, the commission overlooked warnings about the poor quality of much of this research expressed by some of the very authorities it cited.

We have similar misgivings about the use the NCATE has made of the research literature in its latest report. It is our conclusion that the evidence
Reforming Teacher Preparation and Licensing

...cited by the commission provides far less support for the reforms advocated by the NCTAF than the commission has claimed. In this paper we review that evidence, showing why we have come to this conclusion.

This discussion raises two larger issues: what is the relationship between teacher characteristics and student outcomes, and what reforms (either those advocated by the NCTAF or other measures) does the evidence support? Unfortunately, an attempt to answer this question would take us far beyond the scope of this paper. Our focus thus remains on the evidence the commission has cited and its bearing on the recommendations made in the commission’s two major reports.

This dispute over the evidence is not merely academic, however. It is our view that informed public discussion of education reform has been hampered by casual references to “research findings” and sweeping claims about what the research literature shows. The connection between reforms and research needs to be much more carefully drawn if the debate is to move forward, a point that has been forcefully argued by E. D. Hirsch (1996), among other scholars.

In addition, the NCTAF is advocating reforms that would significantly increase the powers of private education organizations represented on the commission. The NCTAF has attempted to persuade the public that critical regulatory functions are best carried out by these organizations. Yet private entities are not directly accountable to the public in the same way as elected officials or their appointees. In addition, regulatory authority empowers these organizations to act in ways that serve private rather than public interests, a significant public policy problem that students of regulation have long recognized. Given these risks, it is incumbent upon the commission to make a strong case that the public will be well served by its agenda. Specifically, the burden of proof is on those who seek to assume regulatory functions to demonstrate they merit the public’s trust. One part of this demonstration is careful, even-handed treatment of the evidence; failure on this count casts doubt on the wisdom of conferring regulatory powers on private education organizations.

Prominent Studies

A section of the commission’s 1997 report, entitled “How Teaching Matters,” opens with the following assertion.

Studies discover again and again that teacher expertise is one of the most important factors in determining student achievement, followed by the smaller but generally positive influences of small schools and small class sizes. That is, teachers who know a lot about teaching and learning and who work in environments that allow them to know
students well are the critical elements of successful learning. (*DWMM*, p. 8)

The commission singles out four studies for particular attention.

a. A study of 900 Texas school districts by Ferguson (1991). According to the commission, Ferguson found that teacher expertise, as measured by scores on a licensing exam, master's degrees, and experience, was the most important set of factors explaining students' reading and mathematics achievement in grades 1 through 11.

b. A 1996 study of Alabama schools by Ferguson and Ladd that found sizable influences of teacher qualifications on student achievement gains in mathematics and reading.

c. A meta-analysis of sixty education production function studies by Greenwald, Hedges, and Laine (1996), which found that teacher education, ability, and experience were associated with significant increases in student achievement.

d. Unpublished research comparing high- and low-achieving schools in New York City in which differences in teacher qualifications accounted for more than 90 percent of the variation in student achievement in reading and mathematics (Armour-Thomas et al., 1990).

The prominence given these studies in *Doing What Matters Most* is only one indication of the importance the commission attaches to them. They have been invoked in numerous articles and interviews in which commission members have articulated their vision of education reform. Charts from *Doing What Matters Most* that purport to summarize the results of these studies have been reproduced in other journals. The findings of Ferguson, Ferguson and Ladd, and Armour-Thomas and colleagues were also offered in commission testimony to the Education Subcommittee of the U.S. House of Representatives in February, 1998, as evidence of the efficacy of the NCTAF's proposed reforms. In a written response to one of our earlier papers, the commission's executive director described these studies as among "the most potent findings undergirding the Commission's efforts" (Darling-Hammond & Berry, 1998).

For all the importance the commission places on this research, however, its claims about the findings of these studies have received little scrutiny. This is unfortunate, for the research cited yields little support for the specific reforms the commission endorses. The commission overstates policy implications, ignoring critical limitations of the research. In many instances, the commission flatly misreports and misrepresents what these studies show.
Modest Policy Implications

Taken at face value, the aforementioned studies show there is a positive association between some measured teacher characteristics and student performance on standardized tests. Which characteristics are most important? In Ferguson’s study of Texas teachers, it was their performance on the Texas Examination of Current Administrators and Teachers—a test of basic literacy. In Ferguson and Ladd’s investigation of Alabama schools, teacher ACT scores were the most important predictor of student test score gains. Similarly, in Greenwald, Hedges, and Laine’s meta-analysis, teacher ability had the largest of the standardized regression coefficients. Less consistently, these studies show that higher levels of teacher education (a master’s degree) and experience are positively related to student performance.

There may be some important findings here, but the policy implications are much more modest than the commission acknowledges. Ferguson’s Texas study, for example, demonstrates that it is important to have literate teachers. But is there anyone who thinks otherwise? Other research indicates that schools do well to attract teachers who had high scores on their college entrance exams or other tests of cognitive ability, and that teacher education and experience matter. But these are rather general findings. The relevant question is what, if anything, this evidence says about the specific reforms advocated by the NCTAF.

Consider, for example, the commission’s recommendation that all programs of teacher education be accredited by the national accrediting body, NCATE. One implication of the research cited by the commission would be to insist that academic standards for admission to teacher education programs be raised. However, a comparison of NCATE and non-NCATE programs finds little evidence to suggest that admissions selectivity plays a significant role in accreditation: NCATE accredits many programs whose students have very low scores on tests of academic skills or achievement.1

The NCTAF also recommends that teachers receive better training before they enter the classroom. There appears to be some support in the data for requiring that new teachers first earn master’s degrees (however, much more on this below). But there is no evidence in these studies that teachers should be “better trained” in any specific sense—for example, that their training follow the guidelines of the National Board for Professional Teaching Standards, as the commission would prefer. The studies cited above offer no support for any particular pedagogical approach over another. These investigations did not identify the kind of training instructors received and were silent on whether they were exposed to the latest, state-of-the-art methods espoused by the commission. It is quite possible that the teachers in these studies earned their master’s degrees in the kinds of programs that the commission deplores.1
Moreover, the commission overlooks one of the important implications of this research. Expanding preservice training (e.g., stipulating that new teachers earn master's degrees) will deter some prospective teachers from entering the profession. Among those most apt to be deterred are college students with strong academic records who have attractive career options outside teaching. Yet deterring prospective teachers of this sort would be precisely the wrong thing to do, given the relationship between teacher ability and student achievement. As we read it, the research the commission has cited supports more flexible approaches to teacher training and licensing, approaches that would relax requirements for individuals of high ability if that will help to bring more of them into the classroom.

To conclude, simply showing that teachers matter does not advance the case for the commission's specific reforms. No one denies that teachers are important. The argument becomes much more problematic when the commission cites such research as if it offers support for its particular vision of education reform.

Limitations of the Research

Before drawing conclusions about policy from the research literature, it is also important to understand some of the limitations of studies of this kind. The studies cited above are examples of education production functions, a category of research in which investigators attempt to relate student outcomes (typically scores on standardized tests or changes in such scores over time) to a variety of educational inputs reflecting the quality of schooling the student has received. Other variables are generally introduced as controls for family background, community characteristics, and other factors that investigators do not want to confound with the effect of schooling.

These background variables are critically important. Students and teachers are not randomly assigned to schools. Students and teachers wind up together because of decisions that have been made on both sides: parents deciding to live in a particular kind of community, teachers drawn to work in certain types of schools. The nonrandomness of this process makes it difficult to draw inferences about the relationship of teacher characteristics to student achievement. For example, if students learn more when their teachers hold master's degrees, is that because the training involved in acquiring a master's degree makes an instructor more effective? Or is it because teachers who work with high-achieving students are also more likely to hold master's degrees for other reasons? A priori it is quite difficult to say. The great majority of new teachers enter the profession with only a bachelor's degree and earn master's degrees over time by taking courses at night and in the summer. It follows that schools with high staff turnover will not have as many teachers with master's degrees, while those that are
attractive workplaces that retain many of their faculty from year to year will have a higher proportion of teachers with advanced degrees. In addition, teachers employed elsewhere in the school system may take advantage of their seniority to transfer into the most attractive schools; many of them will have earned master's degrees by the time they effect their transfers.

Administrators’ preferences also play a role. Principals and superintendents who hire faculty may believe a master's degree makes a teacher more effective, even if there is no basis in fact for that belief. Such a belief then becomes a self-fulfilling prophecy, given that better schools with a wider choice of applicants can hire the kind of instructors they prefer. This creates a positive statistical association between the level of student achievement and the proportion of teachers with advanced degrees, even when the latter has done nothing to contribute to the former. In short, the education level of the faculty may be more a consequence of school quality than a cause.

Thus the background controls play a critical role. By including additional explanatory variables such as the median income in the community, the education level of the adults in the community, per-pupil expenditure, and perhaps prior measures of student achievement, investigators attempt to control for factors that make a school or district an attractive workplace. Thus any remaining influence of teacher education above and beyond the effect of these variables would presumably represent the true impact that an advanced degree has on student learning.

Or so it would in principle. Yet we know that there can be great variation in the effectiveness of one school compared to another, even when they serve similar kinds of student populations. The leadership provided by school and district administrators and the degree of parental involvement have much to do with the success of the school. Much of this influence consists of intangibles or factors (e.g., school discipline and morale) rarely measured in production function studies. As a result, the fact that higher-performing students are taught by teachers with more advanced degrees and experience can be a reflection of other factors not controlled for in the research, rather than a causal relationship. Teacher characteristics cease to be independent explanatory variables and instead become endogenous or dependent variables: the attributes of the teachers, along with the achievement of the students, are joint outcomes of deeper causal processes that make some schools more effective (and attractive to work in) than others. We are not the only researchers concerned about this phenomenon and its implications for research. As John Bishop of Cornell University has written:

Much of the economic research on elementary and secondary education has employed a production function paradigm. Conventionally, test scores measuring academic achievement are the outputs, teachers
are the labor input and students are goods in process. Even though I have written papers in this tradition myself, I am concerned that many of the inputs that conventionally appear on the right in these models are really endogenous and that severely biased findings may result. (Center for Advanced Human Resource Studies, 1994)

A good illustration of the dangers inherent in this technique is provided by the unpublished study of New York City schools cited in Doing What Matters Most (Armour-Thomas et al., 1990). The commission makes much of this study, which found that “differences in teacher qualifications accounted for more than 90% of the variation in student achievement in reading and mathematics at all grade levels tested” (DWMM, p. 9). The study in question compared teachers in six elementary and two middle schools. The schools selected had similar student populations with respect to socioeconomic status, limited English proficiency, student turnover, etc. They differed (and were specially chosen for this reason) in that four of the schools were high-performing (as indicated by student test scores) and the other four were low-performing. That is, the sample was specifically chosen to contain schools that were exceptionally good and exceptionally bad, yet serving similar student populations. In attempting to explain the differences in outcomes, the researchers then found a very high correlation between teacher experience and education and the test scores of the students. But this is neither surprising nor especially revealing. The poor schools, as the authors themselves report, had high teacher turnover and large numbers of inexperienced faculty. The high-performing schools were much more successful at retaining teachers. Is it any wonder that statistical tests showed that students were doing better in the four schools that had more experienced teachers with more advanced degrees? By the same token, is there any reason to think that the differences in measured teacher characteristics caused the differences in performance? This is not to deny that the students in the high-performing schools may have had better teachers. Clearly, they were getting more of something. But it is a stretch to assume that the teacher attributes measured in the study—attributes so clearly dependent on how long one has chosen to stay in teaching—are the critical causal factors. It is noteworthy that the authors of this study are far more circumspect than the commission in the claims they make for their findings:

Although the findings from the present study offer useful information for designing school improvement initiatives, the study is limited in a number of ways... [A] correlational design identified characteristics associated with effective schools but cannot be instructive with regard to which of those functional and status characteristics actually caused schools to be unusually effective. (p. 43; emphasis in original)
Misrepresentations and Errors

Like its 1996 predecessor, the NCTAF's latest report contains numerous errors and misrepresentations of the evidence. Two practices are worth noting in advance. First, the commission frequently combines several different variables into a single category it calls "teacher expertise." This practice obscures how much each of these factors actually contributes to educational outcomes and creates the impression they are all important even when that is not the case. This is of particular concern given that the policy implications are not uniform. What the nation should do to recruit teachers of higher academic ability is not the same as what it should do if the need is for more teachers with master's degrees. By failing to distinguish the effects of general intelligence or experience from teacher training, the commission makes it difficult to assess which of its proposals are worthy of adoption.

In addition, the commission is far too uncritical in its discussion of research on the effects of teacher education. It fails to note that findings in this area are often statistically insignificant or are sensitive to specifications of the model or a selective approach to the evidence.

a. The commission describes Ferguson's 1991 research on student achievement in Texas schools as follows:

In an analysis of 900 Texas school districts, Ronald Ferguson found that teachers' expertise—as measured by scores on a licensing examination, master's degrees, and experience—accounted for about 40% of the measured variance in students' reading and mathematics at grades 1 through 11, more than any other single factor. (DWMM, p. 8)

This is an extraordinary result. It appears to controvert a long-standing belief among researchers, dating back to the work of James Coleman in the 1960s, that students' family backgrounds have far more explanatory power as predictors of achievement than schooling variables.6

However, the commission's account of this research is incorrect and misleading. Consider first the claim that teacher expertise explains "40% of the variance" in student achievement or "more of the variance" than any other factors. This claim is less meaningful than it seems. First, Ferguson's study analyzes student test scores aggregated to the district level. Many studies have shown that most variation in student test scores occurs within rather than between districts. Thus, even if it were the case that variation in teacher qualifications accounted for 40 percent of explained variation between districts, the same teacher qualifications may account for a very different share of explained variation in individual student achievement.7

In fact, however, the commission's statement that teacher qualifications account for 40 percent of the measured variance in student scores is flatly
incorrect; indeed, it is a statistical solecism. There is no sound way to apportion the measured variance of test scores among various explanatory factors. It is possible to speak of the percentage of variance explained by all factors taken together. But there is no acceptable way to assign “43 percent of the variance” to some of them and “57 percent of the variance” to the rest. A full explanation of the errors in the commission’s claim, some of which is rather technical, appears in the appendix to this paper.

Finally, the commission has constructed its “teacher expertise” category in such a way that it masks how much of the effect is due to factors that can be affected by the NCTAF’s reforms, for example, teacher training. Teacher expertise comprises three variables: the mean score earned by a district’s teachers on a competency examination, the proportion of district teachers holding a master’s degree, and the average experience of a district’s faculty. By Ferguson’s own account, the most important of these was the score on the competency test, the Texas Examination of Current Administrators and Teachers (TECAT). The other teacher expertise variables were much less important than the TECAT in accounting for student achievement. By referring to all three together as “teacher expertise,” the commission creates the impression that this study provides strong evidence that graduate education training is associated with higher student achievement. In fact, the proportion of teachers holding an MA was statistically significant in only three of six grade levels. Its estimated impact on achievement was actually negative (though insignificant) in grades 9 and 11.

b. The next study considered by the commission was Ferguson and Ladd’s (1996) investigation of student achievement in Alabama schools. Here is what the commission has to say about this report:

Ferguson and Helen Ladd repeated [the Texas] analysis with a less extensive data set in Alabama that included much rougher proxies for teacher knowledge (master’s degrees and ACT scores instead of teacher licensing examination scores), and still found sizable influence of teacher qualifications and smaller class sizes on student achievement gains in mathematics and reading. These influences held up when the data were analyzed at both the district and the school level. (DWMM, pp. 8–9)

What in fact did Ferguson and Ladd find? Using student-level data, they estimated regression models to explain three dependent variables: reading, math, and combined reading and math scores. The proportion of teachers at the student’s school who held a master’s degree was one of the explanatory variables. It had a statistically significant effect only in the math regression. Ferguson and Ladd also estimated other regressions using district-level data for eighth and ninth grade mathematics. When they controlled for third and fourth grade math scores (approximating a longitudinal or
"value-added" estimate) the proportion of teachers in the district with a master's degree was significant in only one model specification, and then only at the 10 percent significance level.

Nonetheless, in describing these findings, the NCTAF groups teacher education (the proportion of teachers with an MA) with other explanatory variables (ACT scores, class size, teacher experience). Yet the effect of the MA variable itself is very small.

Here is how the authors of this study characterize their student-level findings:

Although the fraction of teachers with master's degrees appears to have little or no effect on reading scores, it exerts a small positive effect on student math scores: a one-standard deviation increase in the fraction of teachers with a master's degree (0.53 points) would increase student test scores by 0.026 standard deviations, about one-quarter the effect of a standard deviation increase in teacher test scores [ACT scores]. (Ferguson and Ladd, 1996, p. 278)

In the same study, Ferguson and Ladd find that the difference in math score gains between the top and bottom quartile of districts is 2.77 standard deviations. Thus, a gain of 0.026 standard deviations is a trivial effect.

In sum, the Ferguson and Ladd study provides relatively weak support for a relationship between graduate training in education and student achievement. They find much stronger support for a link between teachers' general academic achievement (as measured by the ACT) and student test scores.

c. One of the studies on which the commission relies is a meta-analysis conducted by Greenwald, Hedges, and Laine. The NCTAF describes this as a review of "sixty production function studies," and summarizes the GHL findings in a chart that purports to show the gain in student test scores per $500 invested in various educational reforms: reducing class size, raising teacher pay, increasing teacher experience, and investing in teacher education. The largest "effect" is the impact of spending on teacher education: according to the commission, for every $500 invested in teacher education, student test scores will rise by more than two-tenths of a standard deviation. As the standard deviation of student achievement is large, this is a very impressive rate of return. But there is much less here than meets the eye.

First, the commission's description of the GHL study is inaccurate. The relevant table in GHL (Table 6, p. 378) shows that the findings on teacher education are based not on sixty studies, but eight. Moreover, "teacher education" as used by GHL refers merely to whether a teacher has earned an advanced degree. The "investment" in teacher education described by
the commission is simply the extra salary teachers typically receive for holding a master's degree. The legend accompanying the chart in Doing What Matters Most is therefore misleading. As a reading of the GHL study shows, it is not a one-time investment of $500 in teacher training that is at issue here, but rather an ongoing expenditure of $500 per student per year, GHL's estimate of the salary cost associated with an increase of five standard deviations in the proportion of teachers holding an MA.8

Thus the claim that we can significantly raise student achievement by spending $500 on teacher education turns out to be merely the familiar argument that teachers with master's degrees are better. We have already discussed the reasons one should be skeptical about research supporting this claim. This skepticism is not allayed by a close reading of the GHL study itself. The same issue of the journal that published the GHL paper contained a sharply critical review of the study by economist Eric Hanushek that concluded that “ . . . [GHL's] manipulations and interpretations systematically distort the conclusions that should be drawn from the evidence” (Hanushek, 1996, p. 397). Not a criticism of meta-analysis per se, Hanushek's complaint was based on the selective use of evidence by the investigators. In the full set of 46 estimates with which the investigators began, 7 found a significant positive effect of teacher graduate education on student performance, while 6 showed a significant negative effect. The remaining 33 estimates found no statistically significant effect, with approximately equally many positive and negative point estimates.

To arrive at an estimate that showed a large positive effect for teacher education, GHL excluded most of these studies from their meta-analysis and relied on a subset of eight studies that measured teacher education as a simple dichotomous variable (1 for a master's degree, 0 otherwise). According to the investigators, they regarded the dichotomous measure as more reliable than years of postsecondary education or number of credits (used in the remaining studies), although they provided no evidence for this belief. Their results, however, are highly sensitive to this choice. Had they used all twelve of the available studies to compute the impact of a $500 "investment" in teacher education, they would have found an increase of just .0015 standard deviations—basically, a zero effect.9

Finally, the research considered by GHL comprised a mix of student- and district-level studies. Nearly all relied on cross-section data. However, most researchers consider studies based on longitudinal data, which provide both a pretest and follow-up achievement score for each student, to furnish more reliable information on the relationship of schooling inputs to student outcomes. Only two of the teacher education studies considered by GHL used longitudinal data. If their $500 calculation had been based on these studies alone, they would have found a small negative effect of a teacher MA on student achievement (−.05 standard deviations).
Teacher Qualifications and NAEP Scores

The evidence reviewed in *Doing What Matters Most* is not limited to the four studies we have considered in detail. Some of the additional discussion is perfunctory, consisting of by-now-familiar assertions that "hundreds of studies" support the commission's claims, with citations to the education literature. We will not consider such remarks here, which we examined in an earlier paper (Ballou & Podgursky, 1997). However, *Doing What Matters Most* also contains discussion of statistical evidence assembled by the commission, notably a claim that the commission's reform proposals receive support from results of the National Assessment of Educational Progress (NAEP), the periodic examination of American students' learning in various core subjects conducted by the U.S. Department of Education. The commission writes:

The National Assessment of Educational Progress has documented that the qualifications and training of students' teachers are also among the correlates of reading achievement. Students of teachers who are fully certified, who have master's degrees, and who have had professional coursework in literature-based instruction, whole language approaches, study strategies, and motivational strategies do better on reading assessments (see Table 1). (*DWMM*, pp. 11–12)

We reproduce a portion of Table 1 from the commission's report below. The entries in the table are the mean scores on the fourth grade reading test, contrasting students whose teachers have low or substandard qualifi-

<table>
<thead>
<tr>
<th>Correlates of Reading Achievement</th>
<th>Teacher Qualifications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower Scores</td>
</tr>
<tr>
<td>Level of Certification</td>
<td>Substandard or none</td>
</tr>
<tr>
<td></td>
<td>214</td>
</tr>
<tr>
<td>Levels of Education</td>
<td>Bachelor's degree</td>
</tr>
<tr>
<td></td>
<td>215</td>
</tr>
<tr>
<td>Course work in literature-based</td>
<td>No coursework</td>
</tr>
<tr>
<td></td>
<td>214* (216)</td>
</tr>
<tr>
<td>language instruction</td>
<td></td>
</tr>
<tr>
<td>Course work in whole language</td>
<td>No coursework</td>
</tr>
<tr>
<td></td>
<td>214* (216)</td>
</tr>
<tr>
<td>instruction</td>
<td></td>
</tr>
</tbody>
</table>

Reproduced from *Doing What Matters Most*, p. 12. Entries marked with an asterisk are misreported; correct numbers appear in parentheses.
cations (by the commission’s reckoning) with those whose instructors have had superior training.

In several respects, the commission’s claims do not stand up to scrutiny. First, the NAEP results in the commission’s table differ from those reported by the U.S. Department of Education. The average 1992 score for teachers who had no course work in literature-based instruction was not 214, as the commission reports, but 216.4. (All of our statistics come from the National Center for Education Statistics NAEP website: nces.ed.gov/nationsreportcard/y25alm/almanac.shtml) For teachers with no course work in whole language instruction, the mean student score was 215.8, not 214.

Second, the commission reports results for the 1992 NAEP, overlooking data from the more recent 1994 assessment. It turns out the 1994 results are considerably less favorable to the commission. We present the 1994 results in Table 2 below. Observe that in the 1994 results there is no difference between teachers with bachelor’s degrees and those with master’s degrees. Moreover, the result for whole language instruction has been reversed—teachers with course work in whole language approaches have lower student test scores than do those who have had no such training.

Results are still less favorable to the NCTAF in the 1994 assessment of eighth grade reading achievement. Eighth grade teachers with a bachelor’s degree have higher-achieving students than teachers who hold a master’s degree (240.4 to 236.3). Teachers with substandard or no certificates outscore teachers with the highest level of certificate (263.2 to 261.2). Students whose teachers lack whole language training do better than those whose teachers have had it (262.2 to 259.8).

Finally, we have included in Table 2 the standard errors for NAEP proficiency scores as computed by the National Center for Education Statistics—

<table>
<thead>
<tr>
<th>Correlates of Reading Achievement</th>
<th>Teacher Qualifications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower Scores</td>
</tr>
<tr>
<td>Level of Certification</td>
<td>Substandard or none 212 (3.2)</td>
</tr>
<tr>
<td>Levels of Education</td>
<td>Bachelor’s degree 215 (1.2)</td>
</tr>
<tr>
<td>Course work in literature-based instruction</td>
<td>No coursework 214 (2.4)</td>
</tr>
<tr>
<td>Course work in whole language instruction</td>
<td>No coursework 218 (2.9)</td>
</tr>
</tbody>
</table>
information that the NCTAF did not provide in its report. These standard errors are so large that it would be unwise to make much of the contrasts in Tables 1 or 2, since differences of two or three points are generally not statistically significant. Had the NCTAF reported standard errors in the first place (along with correct proficiency scores), informed readers would have seen at once that there was no meaningful difference between teachers the commission deems less qualified and those it regards as more highly qualified. The fragility of the comparisons in Table 1 is all the more evident given (as we have seen) that these differences do not hold up in the 1994 NAEP or in the eighth grade assessment.\textsuperscript{11}

To summarize, the NCTAF is selective in its use of NAEP evidence, choosing the year and grade level that appear to support its thesis and ignoring the year and grade level that do not. Some of the numbers the commission presents are simply wrong. Standard errors are omitted that, had they been reported, would have shown how little basis there is for the commission’s conclusions.

\textit{NAEP Scores and State-Level Reforms}

\textit{Doing What Matters Most} devotes several pages of discussion and three charts to an analysis of NAEP math and reading test score changes. Rather than plotting all of the available data, however, the commission plots outcomes for a subset of states only. For example, for fourth grade math only 11 of 36 potential cases are displayed. (There are only 36 potential cases because state-level scores are not reported for all 50 states on each NAEP administration.) For reading, only 9 of 37 cases are shown.

The discussion that accompanies these charts begins with states whose efforts are deemed praiseworthy.

The critical importance of investments in teaching is demonstrated by states’ experiences over the past ten years. . . . Notable among them for the size and scope of investments were North Carolina and Connecticut. Both of these states coupled major statewide increases in teaching salaries with intensive efforts and initiatives to improve pre-service teacher training, beginning teacher mentoring, and ongoing professional development. Since then North Carolina has posted among the largest student achievement gains in mathematics and reading of any state in the nation, now scoring well above the national average in 4th grade reading and mathematics, although it entered the 1990’s near the bottom of state rankings. . . . Connecticut has also posted significant gains, becoming one of the top scoring states in the nation in mathematics and reading. . . . (\textit{DWMM}, p. 11)
For some praiseworthy states with flat or below-average test score gains, the commission switches to a discussion of levels rather than changes.

Meanwhile, there are a number of states that repeatedly lead the nation in achievement, each of which has made longstanding investments in the quality of teaching. The three long-time leaders—Minnesota, North Dakota, and Iowa—have all had a long history of professional teacher policy and are among the 12 states that have state professional standards boards. . . . (DWMM, p. 13)

Other states are chastised.

By contrast state reform strategies during the 1980’s that did not include substantial efforts to improve teaching have been much less successful. For example, the first two states to organize their reforms around a student testing strategy were Georgia, with its Quality Basic Education Act of 1985, and South Carolina, with its Education Improvement Act of 1984. . . . As figures 7–9 show, student achievement in mathematics has been flat in these states while achievement in reading declined since 1990. (DWMM, p. 14)

Unfortunately, this type of ad hoc exercise lacks validity. Researchers must consider all the data, not just those observations that fit their theories. As noted, the commission makes much of recent trends in North Carolina. Governor (and NGTAF chairman) James Hunt has made his state a showcase for the commission proposals. However, had the commission plotted data for all states, the reader would have seen that there were math gains almost as great in Texas. The 1992–1994 fourth grade math score gains of North Carolina and Texas were identical and first in the nation (+11). The 1990–1996 eighth grade test score gains were +17 in North Carolina (rank 1) and +12 in Texas (rank 2). (The changes in fourth grade reading scores were statistically insignificant for both states.) The Texas case is noteworthy because in 1996 the commission rated Texas dead last in efforts to assure a high quality, professional teaching workforce. (On a ten point scale, Texas received a score of zero, tying it with two other states for the lowest score in the nation.) How is it that a state that in the commission’s judgment devotes so little attention to teacher quality has posted gains that essentially match those of the showcase state and that exceed 47 other states? If the North Carolina experience sheds light on the hypothesis, so does that of Texas.

Moreover, even for the states that are plotted, the commission is very selective about what part of a state’s experience counts and what part does not. Consider Connecticut, the other showcase state. In 1996 the commission gave it a rating of 3 (on a ten-point scale) for its efforts to assure a high-quality, professional workforce. This below-average rating reflected
the state's poor performance with respect to several of the commission's criteria. The proportion of teacher education programs accredited by NCATE was just 13 percent, one of the lowest in the nation. As of 1996 Connecticut provided no incentives for teachers to obtain National Board certification and had not established an independent professional board. In its 1997 report, however, the commission has ignored all of this, lauding the state for enacting very large increases in teacher pay combined with "performance-based" teacher exams. Thus, the Connecticut experience is taken to support the NCTAF's agenda. If Connecticut test scores had fallen, would they have been used as evidence in favor of NCATE accreditation, professional boards, and National Board incentives?

On the other hand, the commission faults Georgia, where test score changes were below-average. Yet Georgia's 1996 professionalization score was above Connecticut's (4 versus 3). The state's proportion of NCATE-accredited programs is 53 percent. Georgia has created an independent professional board and provides incentives for National Board certification. Yet we are told that Georgia did not make the right investments in teacher professionalization.

We could give many other examples of selective use of the evidence and ex post rationalization of the data. However, a more scientific approach is to examine all of the state-level NAEP data to determine whether there is any relationship between the commission's proposals and student achievement.

As noted, in its 1996 report the NCTAF rated each of the 50 states on its efforts to assure a high-quality workforce. The resulting state report card, entitled "Indicators of Attention to Teacher Quality," purports to measure how much each state has done to implement reforms the NCTAF considers important for professional quality. Like other features of the commission's reports, these indicators have been taken up by the media. For example, many of these "professionalization" indicators are now used by Education Week in its widely publicized Quality Counts annual survey. Given that the NCTAF has rated all states' efforts, we examine whether there is a relationship between these ratings and student achievement when all states are included in the analysis.

First, however, a caveat. Many of the state-level policies on which these ratings are based have only recently been implemented and will take some time to affect aggregate test scores. We do not believe that comparing state-level NAEP scores to NCTAF ratings of state reform efforts constitutes a powerful test of the hypothesis that these reforms are valuable. We undertake the analysis solely because the commission has claimed that evidence from NAEP supports its recommendations. Our goal is to determine whether this conclusion holds when data from all states are considered rather than from a select group of states.
In Table 3 we present results from a regression of state average NAEP scores on the ratings in the NCTAF state report card. Column (1) reports simple regression coefficients and column (2) reports the estimated regression coefficients in a model that controls for the student poverty rate in the state. The dependent variables in the models are reading and math test score levels (rows 1–3) and changes between various administrations of these tests (rows 4–6). In none of the twelve sets of estimates do the NCTAF’s ratings of states have a statistically significant relationship with NAEP outcomes.

CONCLUSION

The NCTAF proposes an agenda that would shift considerable regulatory power out of the public domain and into the hands of private professional education organizations. Unlike markets for medical or other professional services, most education consumers (parents and children) have little choice

Table 3. Regression coefficients: Student test scores on NCTAF state scores

\[ p \text{-values in parentheses} \]

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>1996 State Score</th>
<th>1996 State Score*</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAEP 1994 Reading</td>
<td>1.053</td>
<td>.329</td>
</tr>
<tr>
<td>Grade 4 ((n = 39))</td>
<td>(.20)</td>
<td>(.62)</td>
</tr>
<tr>
<td>NAEP 1996 Math</td>
<td>1.190</td>
<td>.362</td>
</tr>
<tr>
<td>Grade 4 ((n = 44))</td>
<td>(.16)</td>
<td>(.60)</td>
</tr>
<tr>
<td>NAEP 1996 Math</td>
<td>1.725</td>
<td>.508</td>
</tr>
<tr>
<td>Grade 8 ((n = 41))</td>
<td>(.11)</td>
<td>(.51)</td>
</tr>
<tr>
<td>NAEP Reading</td>
<td>.373</td>
<td>.397</td>
</tr>
<tr>
<td>Grade 4, Change</td>
<td>(.20)</td>
<td>(.19)</td>
</tr>
<tr>
<td>94–92 ((n = 37))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAEP Reading</td>
<td>.101</td>
<td>.049</td>
</tr>
<tr>
<td>Grade 8, Change</td>
<td>(.68)</td>
<td>(.85)</td>
</tr>
<tr>
<td>96–92 ((n = 36))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAEP Math</td>
<td>.348</td>
<td>.323</td>
</tr>
<tr>
<td>Grade 8, Change</td>
<td>(.32)</td>
<td>(.39)</td>
</tr>
<tr>
<td>96–90 ((n = 31))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*These regressions control for student poverty rate in the state.

Sources: NAEP test data from various volumes of National Report Card, published by the National Center for Education Statistics (U.S. Department of Education), NCTAF teacher quality state score from Doing What Matters Most, appendix A.
as to their teachers or schools. Given that it is a captive market, the potential for harm from such a policy cannot be ignored. The burden of proof is on the commission to make a convincing case that such a change will improve the performance of schools, rather than simply promoting the interests of education producers.

This burden has not been met. The commission’s latest report, *Doing What Matters Most*, reviews several studies from the education production function literature to make the case that teacher expertise matters. The commission also turns to the NAEP for findings that contrast teachers who are “highly qualified” with those whose qualifications are “substandard.” Yet on close inspection, it turns out that the evidence on expertise and qualifications offers little support for the commission’s specific proposals. In particular, the evidence that teachers’ effectiveness is enhanced by advanced degrees earned in schools of education is very weak. By contrast, the data establish more clearly that it is important to recruit teachers of above average general intelligence and academic ability. This pattern is at odds with the NCTAF’s reform agenda, which would lengthen teachers’ preservice training but offer little to attract more capable people into the profession.

**APPENDIX: DECOMPOSING EXPLAINED VARIATION IN A MULTIPLE REGRESSION MODEL**

In his study of Texas school districts, Ferguson estimates the following regression model (we suppress the subscripts for each district and express all variables in deviations from their means):

\[
T5 = b_1 T3 + b_2 HFF + b_3 CS + b_4 TQ + e
\]  

(1)

where \(T5\) is the fifth grade math test score for the school district, \(T3\) is the third grade math score, \(HFF\) is a vector of home and family factors (including community characteristics), \(CS\) is average class size in the district, \(TQ\) is teacher quality (average literacy test score for teachers and average teacher experience and education in the district), and \(e\) is the residual.

Total variation is decomposed as follows:

\[
V(T5) = b_1^2 V(T3) + b_2^2 V(HFF) + b_3^2 V(CS) + b_4^2 V(TQ) \\
+ 2b_1 b_2 C(T3,HFF) + 2b_1 b_3 C(T3,CS) \\
+ 2b_1 b_4 C(T3,TQ) + 2b_2 b_3 C(HFF,CS) \\
+ 2b_2 b_4 C(HFF,TQ) + 2b_3 b_4 C(CS, TQ) + V(e)
\]  

(2)
where $V(.)$ denotes variance and $C(.)$ covariance. The sum of the right hand terms except $V(e)$ equals the explained variation. The presence of the covariance terms accounts for the familiar problem, mentioned in the text, that it is not possible to uniquely decompose explained variation into a part due to HFF and a part due to CS, etc. Only in the special (and empirically irrelevant) case where all of the sample covariances are zero will this be possible.

What, then, does the commission report? Professor Ferguson has provided a copy of the printout that the commission used in its calculations. In that printout, he calculated the following terms: $b_1 \text{SD}(T3)$, $b_2 \text{SD}(HFF)$, $b_3 \text{SD}(CS)$, $b_4 \text{SD}(TQ)$, where $\text{SD}(.)$ refers to the standard deviation of the regressor (hence, $\text{SD}(.) = V(.)^{5}$). Ferguson correctly labels these products as the effect of a one standard deviation change in the regressor on fifth grade test scores. The commission took these estimates, ignored the T3 value, summed the values for HFF, CS, TQ (apart from an error, noted below) and computed the ratio of each term to the totals. This is what they report as “Proportion of Explained Variance in Math Test Score Gains (from Grades 3 to 5).” For example, they compute the contribution of TQ as follows:

$$\frac{\text{Contribution of TQ to Total}}{\text{Explained Variation}} = \frac{b_4 \text{SD}(TQ)}{b_2 \text{SD}(HFF) + b_3 \text{SD}(CS) + b_4 \text{SD}(TQ)}$$

There are numerous errors in this procedure. First of all, the denominator of the preceding expression omits the most important explanatory variable of all—students' scores on tests administered the preceding year. Because scores are highly correlated over time, by omitting this variable from the exercise the commission dramatically increases the percentage of the "total" due to teacher expertise. Moreover, by failing to make clear that a pretest score was included among the explanatory variables, the commission obscures a principal reason that student's family backgrounds are not more important in the analysis. Since family background is largely stable from one year to the next, its impact on student achievement is captured already in the pretest score. Had the original regression equation omitted the pretest score, the measured effect of family background would have been substantially greater.

Possibly, however, the commission assumed that the regression in equation (1) can be reinterpreted as a gain score regression (i.e., with $T5 - T3$ as the dependent variable) simply by dropping terms involving $T3$ from the right side of the equation. Yet this does not save the commission from error. First, respecifying the dependent variable as $T5 - T3$ would change the coefficient estimates ($b_1$, $b_2$, etc.). Converting to gain scores is not done as simply as the commission seems to believe.
Reforming Teacher Preparation and Licensing

In addition, the variance of T5 - T3 is not simply V(T5) + V(T3) but involves C(T5,T3), which has not been accounted for. Thus, even if the regression model had been run with T5 - T3 as the dependent variable (a gain score), the above calculation would be incorrect as it ignores the covariance terms in equation (2). Moreover, even if the covariance terms were all zero, the ratio should be based on variances and not standard deviations.

Finally, apparently due to an oversight, in performing these calculations the proportion of teachers with an advanced degree was omitted from the set of teacher quality variables (TQ) and included instead with the home and family factors (HFF).

Notes

1 This paper, published in the Government Union Review, is available on-line at www.psr.org/doc/v174_art.html.

2 A similar situation arises in the pharmaceutical industry, for example. The Food and Drug Administration routinely relies on firms that develop new drugs to test the efficacy and safety of those drugs. Any sign that the industry deviates from recognized scientific procedure (e.g., carefully controlled clinical trials) in this process immediately calls the wisdom of this policy into question and leads to proposals that the FDA itself assume testing as part of its regulatory functions. A still closer analogy is the self-regulation of the medical profession. If the AMA and the medical specialty boards were not careful and even-handed in their treatment of data that underlie medical protocols, it is doubtful that elected officials would (or should) continue to entrust these organizations with the accreditation of medical schools and the development and administration of licensure and certification tests for physicians.

3 An examination of ACT and NTE scores by institution in Missouri, PRAXIS II scores in Pennsylvania, and state mandated teacher exams of basic verbal and communications skills in Massachusetts finds that many institutions in the lower quartiles are NCATE accredited. See Balkou and Podgursky (1999).

4 To be fair to the commission, some of the studies it describes have concerned teaching methods. For example, reviewing Cohen and Hill's (1998) report on California's new mathematics assessment, the commission writes: "[T]eachers who participated in sustained professional development based on the curriculum they were teaching were much more likely than those who engaged in other kinds of professional development to report reform-oriented teaching practices. These practices and this professional development participation were, in turn, associated with higher mathematics achievement for students on the state assessment..." But even here the policy implications are narrow. The new assessment differed substantially from the old. Teachers who had more training to prepare their students for it altered their teaching methods and saw their students do better. Students who were not so prepared were more likely to be bewildered by the new exam. Thus the research shows that professional development can be efficacious. This is not a trivial finding, and it supports the commission's contention that professional development can be better designed than at present. But it does not establish that the "reform-oriented teaching practices" would have been superior had there been no change in the exam or that the new assessment was a better measure of student accomplishment than the old. Indeed, California has since abandoned both the new test and the mathematics curriculum on which it was based.

5 The claim that teacher attributes explain more than 90 percent of the variance in student test scores should not be accepted at face value, however. The investigators did not
have individual student test scores for their study, and relied instead on average scores in each of the schools. Since there were only eight schools in the study, the aggregation of scores to the school level removed most of the initial variation in the data. Moreover, by matching each teacher with the average test score for the school, the investigators produced a textbook case of an overfitted model: eight distinct test scores regressed on the characteristics of 186 different teachers. This explains why the explanatory variables account for such an extraordinary proportion of the variation in test scores.

6 As noted above, this 40 percent figure has been widely cited. For example, it appears in an article on NCTAF research in the Educational Researcher (Darling-Hammond, 1998). The School Board News reports: "[D]oing What Matters Most] quotes a Texas study that found that teacher’s expertise accounted for 40 percent of the difference in mathematics and reading achievement—more than any other factor." (School Board News, November 25, 1997, p. 2)

7 We are indebted to an anonymous referee for this point.

8 About 50 percent of all teachers now hold master’s degrees. The standard deviation of the proportion with an MA in the GHL data was 10 percent. The authors have apparently calculated that by spending an additional $500 per pupil per year, it will be possible to pay for an MA for every teacher. This represents an increase of five standard deviations in the independent variable. Hence, they simply multiply the coefficient on the standardized regressor by five. This is a dangerous procedure. They are extrapolating far outside the observed data and assuming that the estimated linear relationship still holds. In fact, the data tell us very little about the impact on student test scores that would result from such an extreme increase in one independent variable.

9 This equals the “full analysis” coefficient from GHL (Table 6, p. 978) times an increase of five standard deviations in the percentage of teachers with master’s degrees (.0005 × 5). The cost of this increase in teacher education, on GHL’s reckoning, is an additional $500 per student in teacher salaries. See note 8 above.

10 We have not reproduced the part of the table that compares NAEP scores on the basis of teaching practices, since practices used in the classroom are influenced by the ability level of the students. A simple comparison of means without controls for student ability therefore tells us little about the efficacy of these teaching methods.

11 Even if the NAEP results had been as favorable to the commission as the NCTAF’s report claims, it is far from clear what a comparison of means would actually establish. Tables 1 and 2 include no controls for student background or ability. If it is the poorest schools that have to hire teachers with substandard credentials, it is wrong to attribute the whole difference in scores to the credential per se. In addition, the NAEP is a measure of cumulative student learning through the fourth and eighth grades, whereas the information collected on teachers pertains to the instructor in the year in which the assessment was administered. It is therefore far from clear how much weight should be placed on the answers these teachers give.

References


DALE BALLOU is an Associate Professor of Economics at the University of Massachusetts at Amherst. His research concerns regulation and incentives in public education as they affect teacher training, licensing, recruitment, and compensation.

MICHAEL PODGURSKY is Professor and Chairman in the Department of Economics at the University of Missouri–Columbia. His research in labor economics includes studies of worker displacement and also of labor markets for teachers. In addition to research and testing, he frequently addresses governmental and professional organizations on teacher training and licensure, school choice, and the role of markets in school reform.