ARIZONA STATE UNIVERSITY

Colorado University of Colorado at Boulder | the School of Education

EPRU | EDUCATION POLICY RESEARCH UNIT

EPIC Education and the Public Interest Center

DOCUMENT REVIEWED:	"An Evaluation of Teachers Trained Through Different Routes to Certification: Final Report"
AUTHORS:	Jill Constantine, Daniel Player, Tim Silva, Kristin Hallgren, Mary Grider, John Deke
PUBLISHER/THINK TANK:	Mathematica Policy Research, Inc.
DOCUMENT RELEASE DATE:	February 2009
REVIEW DATE:	March 10, 2009
REVIEWERS:	Sean P. Corcoran and Jennifer L. Jennings
E-MAIL ADDRESS:	sean.corcoran@nyu.edu
PHONE NUMBER:	(212) 992-9468

SUGGESTED CITATION:
Corcoran, S. P., & Jennings, J. L. (2009). *Review of "An Evaluation of Teachers Trained Through Different Routes to Certification: Final Report."* Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved [date] from http://epicpolicy.org/thinktank/review-evaluation-of-teachers

Summary of Review

A new Mathematica Policy Research report finds that students randomly assigned an alternatively certified teacher did no worse on achievement tests than students whose teacher came through a traditional teacher-education route. Moreover, the report concludes that there is no association between greater amounts of teacher training coursework and effectiveness in the classroom. These findings are likely to be warmly received by commentators calling for the scaling-up of alternatives to traditional teacher certification. Such a reception is not warranted, however, because few if any valid conclusions about certification policy can be drawn from the study.

For reasons documented in this review, the study:

- Did not fully report and acknowledge in its conclusions the many analyses from the study finding that traditionally trained teachers outperformed alternative route teachers in both math and reading.

- Has a research design that favors finding few significant differences between groups, most notably its small sample size, sampling methods, and failure to distinguish the "treatments" that alternative certification and traditional certification teachers provided (meaning that members of the two compared groups had substantially overlapping preparation experiences).
- Is relevant only to a very limited population of teachers in schools that hire many alternatively certified teachers, and is not generalizable to most states, districts, and schools that do not allow such programs and are more selective in their hiring.

The study's primary limitations are due to the fact that it intentionally sampled from a unique subset of schools: those that routinely hire alternatively certified teachers. These schools look markedly different from the general population of U.S. schools. The average school in the study was located in a central city, was highly disadvantaged, had 91% minority enrollment, had lower-than-average achievement, and drew heavily on alternatively certified teachers. For this reason, and because the study matched alternative-certified and traditionally-certified teachers working at the same school, it is quite likely that the traditionally certified teachers who made up the comparison group in this study were substantially less qualified than the average traditionally certified teacher.

Teachers in the study were also unrepresentative of the general population of new teachers. Seventy-one percent of the fairly small number of teachers sampled were teaching in the lowest grades (K-2), and most were in just two states (California and Texas). Further, though most of the policy debate surrounding alternative certification concerns teachers in their first two years, the teachers in the study averaged more than three years of experience. And the implications of the study's findings are even further limited by the fact that teachers were not randomly assigned to training routes, but chose these routes themselves.

Taken together, none of the results found in this study can be generalized to the larger population of schools and teachers, and none can be used to meaningfully inform the broader policy debate over alternative certification. Unfortunately, the Mathematica report is quick to draw broad and unqualified implications from the study, and it neglects to properly emphasize the study's many limitations. Policy makers would do well to read it with caution.

Review[1]

I. INTRODUCTION

A new study by Mathematica Policy Research finds that students randomly assigned an alternatively certified teacher did no worse on achievement tests than students whose teacher came through the traditional teacher-education route. Moreover, the report, titled "An Evaluation of Teachers Trained through Different Routes to Certification" and funded by the Institute of Education Sciences,[2] concludes that there is no association between greater amounts of teacher training coursework and effectiveness in the classroom. These findings are likely to be warmly received by commentators calling for the scaling-up of alternatives to traditional teacher certification. As advocated by Malcolm Gladwell, "Teaching should be open to anyone with a pulse and a college degree—and teachers should be judged after they have started their jobs, not before."[3] Yet such a reception is not warranted in this case, because few if any valid conclusions about teacher certification policy can be drawn from the Mathematica study, and conclusions that can be drawn tend to favor traditional routes.

Later in this review we outline the results that we think should have been given greater attention in the report and in its executive summary. But we want to be careful here not to overstate the findings—because in reality the findings are minimal; few results meet the standard of both practical and statistical significance. This is for good reason: if one set out to design a study that would find no statistically significant differences between the achievement of students taught by traditionally and alternatively certified teachers, this is precisely the study one would have designed.

The study includes only a moderately small sample of 174 elementary teachers, with an average of three years experience, and is heavily weighted towards grades K-2, who comprise 71% of the study's teachers. Because the authors intentionally sampled from schools that routinely hire alternatively certified teachers, the average school in the study is located in the central city of an urban area, is highly disadvantaged, and is demographically distinct even from other schools found in the same district. Many of these schools are high-turnover organizations that draw heavily on alternatively certified teachers. As explained later in this review, these school characteristics are important in part because the traditionally certified teachers who are the comparison group in the study are only those employed by these disadvantaged schools. While many of these teachers are highly qualified individuals who choose to serve the most needy students, it is very likely, based on the evidence provided here and in other studies, that the teachers at these schools are, on average, the least competitive. That is, one has every reason to believe that the traditionally certified teachers in these schools tend to be less qualified than other traditionally certified teachers.

Yet despite this and other aspects of a study design that favors finding no differences even where differences exist, the report does find negative and statistically significant outcomes associated with alternatively certified teachers in a number of settings. For example, students in grades 2-5 with alternatively certified teachers who had

taken relatively low amounts of teacher-education coursework did substantially worse on a test of math computation. Students of alternatively certified teachers who were concurrently enrolled in coursework (43% of these teachers) also performed worse than students of their traditionally certified peers.

As explained in this review, the main concerns we have with the Mathematica report fall into three categories:

1. The study's design makes it difficult to discern what, exactly, constitutes the "treatment" in this experiment, thus limiting its internal validity.

2. The report's findings cannot be generalized beyond a highly specific population of high-needs, high-turnover classrooms of early grade students, thus limiting its capacity to inform questions about the relative success of traditional and alternatively certified teachers, and policy questions about teacher preparation.

3. The limited attention to findings of negative outcomes associated with alternatively certified teachers in the main body of the report, as well as the executive summary and accompanying press release, distorts the study's policy implications.

II. REPORT'S FINDINGS AND CONCLUSIONS

The Mathematica team compared the math and reading test scores of students taught by teachers certified through alternative routes (AC) to those of students taught by traditionally certified teachers (TC). Classroom practices of AC and TC teachers were also observed and compared. (Details on the study design are provided in Section III of this review.) From their analysis of these outcomes, they draw the following conclusions, all spotlighted in the report's executive summary:

> There was no statistically significant difference in performance between students of AC teachers and those of TC teachers … Therefore, the route to certification selected by a prospective teacher is unlikely to provide information, on average, about the expected quality of that teacher in terms of student achievement (p. xviii).
>
> There is no evidence from this study that greater levels of teacher training coursework were associated with the effectiveness of AC teachers in the classroom … Therefore, there is no evidence that AC programs with greater coursework requirements produce more effective teachers (pp. xviii–xix).
>
> There is no evidence that the content of coursework [including required hours of pedagogy instruction, or fieldwork] is correlated with teacher effectiveness (p. xix).

These are all strong statements about the relative effectiveness of traditional and non-traditional teacher training programs, and the policy community and news media have naturally been quick to take note. For example, in his column in the *New York Times,* Nicholas Kristof summarized the implications of the study this way: "The latest Department of Education study, published this month, showed again that there is no correlation between teacher certification and teacher effectiveness … The implication is that throwing money at a

broken system won't fix it, but that resources are necessary as part of a package that involves scrapping certification."[4]

Indeed, the Mathematica report's authors do not shy away from drawing broad implications from their findings. They suggest, for example, that their results on relative student achievement will be "relevant to principals faced with a choice between hiring an AC or a TC teacher" (p. xvi). Likewise, their findings on the impact of teacher training coursework will provide guidance "to policymakers and designers and administrators of teacher training programs in their efforts to identify the training characteristics and certification requirements that are related most positively to student achievement" (pp. xvi–xvii).

Mathematica's own press release presents its findings as having broad and unqualified generalizability:

> In one of the largest and most rigorous studies of alternatively certified teachers ever conducted, researchers found that students with an alternatively certified teacher did no worse on achievement tests than students whose teacher came through the traditional route. [5]

The report's lead author adds: "Our study reveals that alternatively certified teachers do not produce harmful consequences for students."[6] As we explain in Section V, such broad statements are not warranted from the study design, or from its own findings.

III. Report's Rationales for its Findings and Conclusions

As many of Mathematica's large-scale evaluations do—and often do well—this study relies on random assignment of subjects to "treatment" and "control" conditions to estimate the effect of a treatment on an outcome. In this setting, the subjects are students, while the treatment and control conditions are the classrooms of teachers certified through AC and TC routes, respectively. As in medical research, randomized control trials (RCTs) are commonly viewed as the "gold standard" for evaluating the effects of interventions, and of social and educational programs.[7] Since 2002, the U.S. Institute of Education Sciences has explicitly worked to promote RCTs in education, and this IES-funded study is one example of these efforts. In principle, the RCT is a straightforward, clean, and powerful design for making causal inferences about interventions and programs.[8] In practice, RCTs often have limited applicability outside their study sample.[9]

Mathematica designed its RCT as follows. First, a sample of 63 AC programs was selected from a total of 165 non-selective AC programs operating in 12 states.[10] Second, schools that hired from these 63 programs were recruited to participate. Only schools where AC and TC teachers were observed instructing the same grade were eligible to participate, and all participating teachers had to be "relative novices" (initially defined as three or fewer years of experience, and later re-defined as five or fewer years). A total of 87 AC teachers and 87 TC teachers in 63 schools made the cut. Third, students within the same school and grade were randomly assigned to an AC or TC teacher. Finally, outcomes in AC-led classrooms were compared with those of (matched) TC-led classrooms. The average difference in outcomes is interpreted as the average "treatment effect" of AC teacher instruction.

For the purposes of this study, AC teachers

were not considered to provide a uniform "treatment." Rather, they were differentiated into two subgroups based on the amount of class instruction and fieldwork required by the AC programs the teachers had chosen. "Low coursework" teachers were in programs requiring relatively little prior training, while "high coursework" teachers were in programs requiring relatively more instruction and fieldwork (47 AC teachers fell into the first category and 40 were in the latter). Drawing an analogy to medical RCTs, students were instructed by teachers who received different "dosages" of teacher instruction.

The experimental "treatment effect" of AC teacher preparation (or low- or high-coursework AC teacher preparation) is calculated as the average difference in outcomes between AC classrooms and their matched TC pairs. The authors also perform non-experimental analyses, using multiple regression models to examine the effects of other observable differences between AC and TC teachers and classrooms. Characteristics they control for include, among other things, student pre-test scores, teacher and student demographics, teacher education, and teacher experience. Due to random assignment, student characteristics should not vary systematically between AC and TC classrooms in the same school. But AC teacher characteristics vary according to state and program training requirements, and differential selection into the AC route. It is these AC teacher characteristics that together constitute the study's "treatment."

IV. REVIEW OF THE REPORT'S USE OF THE RESEARCH LITERATURE

The report provides an ample literature review of current studies of AC. But the report's authors do not adequately address a large literature on teacher labor markets—in particular, the process by which teachers are sorted among and within schools—that informs the context and generalizability of their study.[11] As we describe in Section V of this review, the study pays scant attention to the sorting of teachers into schools and grade levels in their sample, and thus fails to address how their findings might or might not apply in other teacher labor market conditions.

The report's literature review would have benefited from additional attention to three issues that are ultimately raised by the sample participating in the study:

The interaction between certification type and years of experience: The authors cite a number of studies on the effectiveness of AC.[12] To the extent these studies find negative effects of AC on student achievement, the authors correctly note that those effects are limited to the first two years of teaching. From the report's literature review, and from our reading of this literature as well, the debate about AC teachers appears to be entirely about what happens to students exposed to these teachers in their first two years. But there is a major disconnect here between the findings of this cited research and the sample utilized by the Mathematica study: 57% of the Mathematica sample has three or more years of teaching experience (see Section V of this review for more on this issue). What is missing from the review, then, is an express discussion of this issue. If the AC teachers studied by Mathematica are largely beyond the earlier period of time when prior research suggests harmful effects, the study's findings shed little light on the question of what happens to students who are exposed to less experienced AC teachers in their initial years.

The effects of teacher certification and the

size of teacher effects in the early grades: A large and growing literature on teacher effects in the grades tested under No Child Left Behind (in particular, grades 3-8) has inspired a fruitful debate over the role of certification in teacher quality, but no such studies exist for the lower grades (K-2), which comprise 71% of the Mathematica study's teachers. There is only one study of which we are aware that examines the magnitude of teacher effects across the early grades. Nye, Konstantopoulos, and Hedges find that teacher effects on reading are somewhat smaller for the earlier grades than for higher grades, though teacher effects on math are not dissimilar from those previously reported in grades 3-8.[13] However, without the benefit of observing teachers multiple times (as most teacher effects studies do), Nye et al. potentially overstate the size of teacher effects due to sampling variation. In short, whether teacher effects in K-2 are of similar size to teacher effects in grades 3-8—whether because there is less variation in teacher quality among early grades teachers, or because there are fewer reliable measures of skills in the early grades—remains an open question but one quite relevant for this study.

The sorting of teachers within schools: The Mathematica study assumes that there is no systematic relationship between teacher quality and the grade levels to which teachers are assigned. That TC and AC teachers are placed in the same grade is treated as random. However, in the current accountability climate, there are good reasons to believe that principals strategically deploy teachers to the tested grades (3-8), essentially placing their most effective teachers where they are most likely to benefit the school in the short term. Previous research on the implementation of high-stakes testing in New York State found support for this idea.[14] If this is the case, then one might expect the least effective teachers in the earlier grades. Since the report's sample is comprised largely of K-2 teachers, the possibility of teacher sorting within schools should have been given substantially more attention.

V. REVIEW OF THE REPORT'S METHODS

Our concerns with the Mathematica report's methods fall into three categories. The first addresses the report's internal validity—its ability to accurately draw inferences about the effects of a treatment on an outcome. As we describe below, the study's design makes it difficult to discern what, exactly, constitutes the "treatment" in this experiment. The second relates to the report's external validity—its ability to generalize beyond the unique and idiosyncratic settings of the study to other populations. We show that the findings of the Mathematica report cannot be generalized beyond a highly specific population of high-needs, high-turnover classrooms of early grade students. Finally, the third relates to the authors' selective emphasis of their findings. Throughout the report, the authors find numerous cases of negative outcomes associated with AC teachers, but more often than not choose to deemphasize these findings.

Internal validity: what is the nature of the treatment?

Mathematica's experimental analysis randomly assigned students attending the same school in the same grade to either an AC teacher or a TC teacher. Because most students in the U.S. are taught by a traditionally certified teacher, one can think of assignment to a TC teacher as the control state and assignment to an AC teacher as the treatment. So how do the study's authors

operationalize this notion of "treatment?" That is, what is it that students assigned to AC teachers are "getting" that differs from what students assigned to TC teachers receive? And how confident should readers be that this study is appropriate for identifying the effects of this treatment?

In this study, TC teachers are defined as those who began teaching only after completing their certification requirements, while AC teachers are defined as those placed in a classroom prior to completing these requirements (p. 9). Accordingly, the key distinguishing feature of TC and AC teachers is not the actual amount of "traditional" coursework and experience they had at the time of the study. Instead, the key distinguishing feature is the point at which they began classroom teaching. Of course AC teachers follow a different pathway into teaching, so the training they bring to the classroom is in some cases quite different from that brought by TC teachers. But the basic treatment is exposure to a teacher who entered the classroom prior to completing certification requirements (and who chose to pursue that route into teaching).

AC teachers are further subdivided into two subgroups: low- and high-coursework teachers. "Low-coursework" teachers attended programs requiring relatively little instruction and fieldwork, while "high-coursework" teachers attended programs requiring more hours of training. One might think of these two subgroups as "doses" of the treatment: some students were exposed to AC teachers who attended programs with low requirements, while others were exposed to teachers who attended programs with high requirements. Incidentally, these "doses" were not randomly assigned; they were determined mostly by geography and state requirements. Two-thirds of the low-coursework teachers were in Texas, while half of the high-coursework teachers were in California (p. 29-30).

While there is no commonly accepted definition of an alternate certification teacher (or a low- or high-coursework AC program), this report's definition results in a very broad and inclusive range of "treatments." For instance, while timing of entry into the classroom differs for these teachers, the instructional training required of AC and TC teachers overlapped considerably. The Mathematica report's executive summary states, "the total hours required by AC programs ranged from 75 to 795, and by TC programs, from 240 to 1,380. Thus, not all AC programs require fewer hours of coursework than all TC programs" (p. xvii). It continues, "in California, the range of coursework hours required was similar for AC and TC teachers" (p. xviii).[15] In other words, many pairs of AC and TC teachers brought very similar training and experiences to the classroom. This was especially true for high coursework AC teachers (Exhibit III.11). Some AC and TC teachers in the study were trained in the same institutions, and may have taken the same courses (p. 25-27).[16]

Further muddying the waters, teachers in the study are also categorized into low- or high-coursework programs based on "the requirements of the programs they attended and the amount of coursework required for certification, *not the amount actually completed at the time of the study*" (p. xxiii, emphasis added). So very little is known about how much training the teachers in the study actually received by the time they were observed. Transcripts would have provided a much clearer picture of the actual coursework completed at the time of the study. Given the report's description of the categorization of AC and TC teachers and of

the experiences and training they brought to the classroom, there does not seem to be a unique "treatment" that students assigned to AC teachers were really receiving.[17]

Interestingly, one dimension on which the AC teachers in the study did differ markedly from TC teachers was the availability of a mentor, master teacher, or supervisor during their first year of teaching. According to the report, 93.5% of low-coursework AC teachers worked with a mentor during their first year, compared with 78.3% of their TC counterparts (p. 47-49). The difference between high-coursework AC teachers and their TC counterparts was even larger. AC teachers also reported more professional development and administrative support than TC teachers (p. 48-50). None of these differences are unexpected, as some AC programs involve these supplemental services. But they do illustrate another component of the "treatment" that students assigned to AC teachers receive—additional classroom support during the school year—that TC counterparts did not. Measured outcomes might be picking up effects of mentoring, professional development, or administrative support, rather than (or in addition to) other aspects of alternative certification.

Internal validity: cooperation and interference

A central assumption of randomized controlled trials is that there is no interference between units in the experiment. In other words, a subject's outcome must depend only on the subject's own treatment assignment, not the treatment assignments of other subjects.[18] In settings like elementary schools where teachers often work together in grade-level teams, this assumption is often violated. Strictly speaking, in the Mathematica study *students* are the subjects randomly assigned to treatment and control states, not teachers. However, to the extent TC and AC teachers interact with each other, support each other, and plan together, we would expect their students' outcomes to look more similar to each other than they would in the absence of such cooperation, thus increasing the likelihood that the study would show—as it mainly did—no significant effects. Nowhere in the study do the authors discuss the incidence or likelihood of these interactions.

Internal validity: The effects of teacher-student race matching

A number of studies have found that African American students perform better academically when they are taught by African American teachers,[19] which is of interest in the Mathematica study because AC teachers are 2.7 times more likely to be African American (36.1% versus 13.6%), and the study's teachers are predominately in schools with high proportions of African American students. Recognizing this issue, the Mathematica authors test for an interaction effect of teacher and student race, and indeed find that African American students perform better when they are matched to African American teachers. These effects are substantial in size,[20] but the authors dismiss this issue as not relevant to their estimates of the effects of AC teachers because these effects are not statistically significant at the .05 level. Considering the substantial size of these race matching effects, the Mathematica study would have been strengthened if the authors reported the effects of AC and TC teachers net of teacher race and of the interaction between teacher and student race. By failing to do so, the report only raises additional questions about the nature of the AC "treatment."

External validity: to what populations do these results apply?

By design, the Mathematica report restricted its analysis to AC and TC teachers in schools that (a) regularly hired a large number of AC teachers, and (b) had "relatively novice" AC and TC teachers providing instruction to the same grade level. From a design point of view these choices were practical, for two reasons. First, absent the ability to force AC teachers on unwilling schools, the authors were limited to schools that already hire from this pool of teachers. Second, in a desire to compare "apples to apples," the authors sought to compare teachers instructing the same grade in the same school. Most readers surely (and justifiably) would find fault with a study that opted to compare, say, a 3rd grade AC teacher in one school to a 6th grade TC teacher in another.

But such decisions necessarily place strict limitations on the population of schools, grades, and teachers for which this study has relevance. In turn, these decisions place strict limitations on the population to which the study's results can ultimately be generalized. Below, we describe how Mathematica's research design restricts its generalizability with respect to districts and schools, grade levels, and teachers.

Selection of schools and districts

Most importantly, schools and districts that hire AC teachers—especially teachers from the kinds of non-selective training programs studied here—look markedly different from the general population of schools and districts in the United States. Alternative certification programs emerged in large part in response to staffing shortages, particularly in hard-to-staff schools and subject areas. In particular, urban schools serving high concentrations of poor and minority students historically have found it difficult to recruit and retain certified teachers.

The report provides descriptive statistics for schools and districts included in the study, and it compares student characteristics in participating schools to those in non-participating schools in the same districts (p. 20-22). The authors acknowledge that these statistics "provide a context for understanding the settings and students for which the study findings are most relevant" (p. 20). But little more is said about how closely these schools and districts resemble the general population of schools and districts, or how these choices should inform inferences drawn from the study.

Exhibits II.4 and II.5 illustrate some striking differences between study schools and the general population of schools. Fourteen of 20 districts and the vast majority of schools in the study are in central cities of urban areas.[21] The study's schools have an average of 79% eligibility for free and reduced-price lunch, and an average of 93% non-white enrollment. By contrast, 38% of students in the nation at large are eligible for free or reduced-price lunch, and 45% are nonwhite.[22]

Striking dissimilarities exist even when comparing participant and non-participant schools in the same districts (Exhibit II.5). Schools in the study had much higher rates of poverty and non-white enrollment than other schools in their same district. In some cases—including two large urban districts in California, urban districts in Georgia and Wisconsin, and a rural district in Louisiana—the differences are substantial, at 10 to 40 percentage points. These differences are consistent with the existing literature on the distribution of non-traditionally certified teachers across schools.

Taken together, we can conclude that districts and schools hiring AC teachers face much different circumstances than districts and schools that do not—a conclusion that holds even among schools in the same labor market (i.e., a particular school district). This observation is critically important, for two reasons. First, these differences highlight the limited population to which this study can be generalized. Perhaps even more importantly, they offer useful insight into the TC teachers that serve as the study's counterfactual, or "control" group.

TC teachers in schools staffed with AC teachers are unlikely to be representative of TC teachers in the general population, and they may not even be representative of TC teachers in their own districts. Schools that hire AC teachers typically do so out of need. If these schools are troubled, high-turnover organizations, they do not have the luxury of selecting their hires from among many qualified applicants. Instead, it is likely that even their TC staff suffers from lower-than-average quality. In fact, most of the existing literature confirms this: schools with high concentrations of poor or minority students are disproportionately staffed with less effective, less experienced, and less academically talented teachers.[23]

Selection of grades

The sampling design of the Mathematica report also yields an unusually skewed distribution of teachers over grade levels. As the report's Exhibit II.3 shows (and the below table summarizes), nearly 56% of all teachers in the study were Kindergarten and 1st grade instructors; 71% were concentrated in grades K-2.

There are a number of plausible explanations for this over-representation of early grade teachers. As the report shows, schools hiring from AC programs have higher shares of poor and minority students, greater turnover, and fewer qualified teaching staff.[24] In a high-stakes testing environment—such as that under No Child Left Behind—an under-resourced school may rationally assign its most capable and effective teachers to the tested grades (3-8). If schools behave in this way, we will observe more AC teachers and—even more importantly—weaker-than-average TC teachers in the early grades.[25]

Table 1. Distribution of teachers by grade

	# of teachers	% of matched pairs
K	20	22.2
1	30	33.3
2	14	15.6
3	9	10.0
4	11	12.2
5	6	6.7

There are several reasons why the over-reliance on early grade teachers is important. First, as was true for the non-random sample of schools, the study's method of selecting eligible teachers further restricts the population to which its results can be generalized. Second, if TC teachers assigned to early grades are among the least qualified or effective in the school, comparisons between AC and TC teachers in the earlier grades will be least likely to find differences in effectiveness.[26] Third, as noted in Section IV, the existing literature has little to say about teacher effects in the early grades. Finally, the nature of instruction differs markedly between the early and middle grades. If early grade educators are more likely to follow rote lesson plans or otherwise have less control over their curriculum, few outcome differences between AC and TC classrooms might be expected.

Selection of teachers

The Mathematica report sought to compare "relatively novice" AC and TC teachers providing instruction to the same grade level within a school. The report's sampling design also sought to include roughly half "low-coursework" and half "high-coursework" AC teachers. As it turned out, finding a sufficient number of teacher pairs who met these criteria proved to be quite difficult. In the first year of the experiment, when "relatively novice" was defined as 3 or fewer years of experience, only 25 AC teachers and 24 TC teachers were available for inclusion (p. 14-15). In order to obtain a sufficient sample size, the authors in the second year retained as many teachers as possible from the first year, and broadened its definition of "relatively novice" to 5 or fewer years of experience. This allowed for many more matched pairs; over the two years combined, 87 AC and 87 TC teachers participated.

A consequence of this expanded sampling procedure was a relatively high average experience level among participating teachers. As shown in the table below (calculated using Exhibit 4 and Exhibit III.14), the average teacher in the study had about 3.1 years of "study-eligible teaching experience."

Table 2. Average teacher experience

	AC	TC	Combined (weighted average):
Low coursework	2.7	3.3	3.0
High coursework	3.3	3.0	3.2
Combined (weighted average):	3.0	3.2	3.1

Given random assignment of teachers to students, why should the average experience of participating teachers matter? We offer two reasons. First, AC teachers may have a higher rate of turnover than TC teachers in their first few years of teaching. To the extent that exiting AC teachers are less effective than the ones who stay, the average quality of remaining AC teachers will be higher. If this was the case, the study would overlook the potential negative effect that AC teachers have on students in their first year of teaching. Second, much of the policy discussion surrounding alternate certification relates to the potential risk of hiring under-prepared and inexperienced teachers who have not completed their formal training. (The authors cite this as one of their central motivations for the study: see p. xv and our discussion in Section IV.) But as the above table shows, participating teachers already have demonstrated longevity in the classroom. The bulk of teachers were recruited in the second year under the less stringent definition of novice teachers, and 14 of the study's teachers were retained and observed for two years in a row (p. 14). Even these teachers were a select sample, given that many teachers in the first year of the study were unavailable for year two.

Recent empirical evidence finds that less effective teachers are in fact more likely to transfer schools or exit teaching than more effective teachers in their first few years.[27] Further, a more limited body of research shows that AC teachers have a higher rate of turnover than TC teachers in their first few years.[28] The latter result should not be too surprising—TC teachers have already revealed a potentially greater commitment to the profession (by investing in a more time-consuming educational program) than AC teachers. The same argument may differentiate low- and high-coursework AC teachers.

Selective emphasis of results

A careful read of Mathematica's report reveals a large number of relevant findings that received little or no attention in the executive summary, or in press coverage of this study. We briefly highlight several that caught our attention:

1. The authors reported that there was no difference between the math performance of AC and TC teachers, and they also reported that there was no variation in the effects of AC versus TC teachers across grade levels. Yet the authors excluded scores from half of the math tests administered to students in grades 2-5 ("Math Computation") from the analyses reported in the body of the study.[29] Excluding the Math Computation scores from these primary analyses represents a surprising design choice, given that Exhibit A.9 demonstrates that *the students of alternatively certified teachers from low coursework programs scored significantly lower on math computation, and the magnitude of this effect is substantial in size (Effect size = -.18 of a standard deviation).* This finding not only suggests that students in grades 2-5 are harmed by exposure to alternatively certified teachers, but that there is important variation in the effects of alternative certification across grades.

2. A central concern with AC programs is how the timing of teacher-education coursework affects student outcomes. The Mathematica study found that the *students of alternatively certified teachers currently taking coursework—43% of all alternatively certified teachers— performed worse in math (ES = -.09).*

3. The report's findings are also sensitive to the inclusion or exclusion of students and teachers who exited during the study. In analyses reported in Exhibit A.10, the authors find that after excluding students and teachers who left during the year from the analysis, the students of *alternatively certified teachers from high-coursework programs scored lower on math, and these differences were statistically significant (ES= -.08).*

4. There are potentially important geographic variations in the effects of alternative certification that were not attended to in the press release or executive summary. The authors found that overall that *the students of AC teachers in California performed worse in math than the students of TC teachers (ES=-.13).* They further determined that this effect is driven by the 62% of California AC teachers currently enrolled in coursework (ES = -.16).

5. The authors underplay the finding that on all instructional dimensions observed in the classroom, *high-coursework AC teachers were rated substantially worse than high-coursework TC teachers.* The reason for this de-emphasis was because many of these large effects do not reach statistical significance. But this lack of statistical significance is not surprising, since the study had a very small sample size, allowing the researchers to detect only what were (by educational research standards), enormous differences between groups. The difference between high coursework TC and AC teachers ranges from .22 to .37 standard deviations in reading, and it ranges from .27 to .33 standard deviations in math—but the only difference between TC and

AC high-coursework teachers that reached statistical significance was on literacy culture (and it registered a remarkable difference of .40 standard deviations).

6. Finally, the authors underplay that AC teachers received lower principal ratings on every dimension, and some of these differences are quite large in magnitude. Again, because the sample size is small, even large differences—for example, the .42 standard deviation TC advantage over high-coursework AC teachers in classroom management—do not reach statistical significance.

Regarding points 5 and 6, these results do not reach statistical significance because the comparisons are (appropriately) made at the level of the teacher (n=188) rather than the level of the student. However, when a study using relatively small sample sizes produces results showing large effect sizes that fall short of statistical significance, it is important to bear in mind what tests of statistical significance are intended to do—provide readers with an estimate of the probability of a difference between two groups occurring simply by chance. The smaller the sample, the more likely it is that even differences of practical significance—differences that are substantively important for education policy—will not be statistically significant. As many other researchers have argued,[30] discussing both the effect size and the statistical significance of effects is particularly important in the case of "low-power tests"—tests of statistical significance where the sample size is small.

VI. REVIEW OF THE VALIDITY OF THE FINDINGS AND CONCLUSIONS

Throughout the Mathematica report, the authors argue that their findings on student achievement will be "relevant to principals faced with a choice between hiring an AC or a TC teacher." They add that their results on low- or high-coursework training programs will provide guidance "to policymakers and designers and administrators of teacher training programs in their efforts to identify the training characteristics and certification requirements that are related most positively to student achievement" (pp. xvi-xvii, 3-4, 12). It is unclear to us that the report's findings are useful to either of these ends.

The latter of these two claims is discussed further in the final section of this review. With respect to the first (helping to inform the hiring of teachers), even if this study were perfectly executed, these findings would apply only to very select population of districts, schools, grades, and geographic regions of the country. The schools that provided the study's sample are disproportionately poor, heavily minority, low-performing schools located in central cities of urban areas, and the schools have histories of hiring at least moderate numbers of AC teachers. Further, the vast majority of teachers in the study were K-2 teachers.

If a California or Texas principal working in a central-city, urban school facing similar conditions has a choice of an AC or TC teacher for Kindergarten, 1st, or 2nd grade, he or she could use this study to assess the costs and benefits of hiring an AC teacher. (Teachers from California and Texas comprised 71% of the sample.) The findings of this study cannot, however, be generalized to the overall population of schools and teachers, particularly the kinds of schools that do not have prior histories of hiring AC teachers.

Moreover, as we described in Section V of this review, it is unclear what information

the study can provide even to the select population of districts, schools, and grades for which it is relevant. The AC "treatment" as operationalized here is not easily differentiated from the "control," and the training routes themselves were not randomly assigned to teachers but governed by self-selection and state requirements. The study assumes that AC and TC teachers in the same grade and the same school do not cooperate, and thus do not influence each others' practice and outcomes, which we find highly implausible.

Unfortunately, despite these limitations on the report's external and internal validity, the report's authors elected not to provide the necessary cautions to their readers.

VII. REPORT'S USEFULNESS FOR GUIDANCE OF POLICY AND PRACTICE

Policymakers and other readers of the Mathematica report will surely be looking for a simple answer to a simple question: "Is alternative teacher certification a *bad thing* or a *good thing?*" They will also surely be interested in an important corollary to this question: "If alternative certification does no harm, is traditional teacher certification even *necessary*?" Notwithstanding suggestions to the contrary in the report's press release or executive summary, this report is unable to provide a satisfactory or general answer to either of those questions.

In this review we have addressed some of the ways in which the Mathematica study provides a weak test of the AC "treatment effect." We have also emphasized the very limited population to which the study's findings can be generalized. And we have pointed to de-emphasized results that point to negative AC outcomes, in conflict with the report's broad conclusions.

In truth, the study's limitations are not completely disregarded by the report's authors. The executive summary contains the following key observation:

> An important distinction of this design is that because certification routes are not randomly assigned to teacher trainees, the estimates of the effects on student achievement and classroom practices of teachers who were trained through different routes to certification pertain to those *who chose to participate in these programs*. Because of likely differences in the types of people who attend various certification programs, *the results cannot be used to rigorously address how a graduate of one type of program would fare if he or she had attended another type* (p. xxi, emphasis added).

Teacher candidates self-select into training programs, which are traditional or alternative, and which are low-coursework and high-coursework. Absent random assignment of teachers to training routes, we cannot determine whether AC programs add value, or not, or if lower-coursework programs are as effective as higher-coursework programs. All one can say is that—for those types of schools and grade levels studied—the students of teachers who decided to go through AC routes may have performed no worse in reading and somewhat worse in math (in some instances) than those who were taught by teachers who selected traditional routes. The study is fundamentally unable to provide evidence about a counterfactual world in which traditional certification ceases to exist, or is not the default.

But the report fails to recognize other

limitations. For instance, another critically important source of non-random variation is the certification requirements imposed by states, but the report's introduction states:

> The increased variation in the teacher preparation approaches created by the existence of various AC and TC programs offers an opportunity to examine the effect of different components of training on teacher performance … We can exploit this type of variation to examine whether the form of training is associated with differences in teacher performance (p. xv).

But, of course, state requirements for AC programs are not randomly assigned. States designate requirements for a reason. A state with particularly dire teaching shortages may elect to set a very low bar for alternative certification. On the other hand, a state with exceptionally high standards for teacher quality may set a higher bar. Either way, these requirements will be related to the average quality of schools and teachers.

Finally, this study can make no claims about the long-run effects of a wholesale movement away from traditional teacher certification. The fact that a small number of AC teachers in a select population of schools, grades, and states performed only somewhat worse than TC teachers in the same schools cannot help us learn about the systemic changes in teacher quality, selection, and instruction that would arise under a wholly different system. Unfortunately for fans of randomized control trials, no amount of randomized assignment will answer such questions.

Notes and References

[1] The authors would like to thank Aaron Pallas and Kevin Welner for helpful comments and suggestions.

[2] Constantine, J., Player, D, Silva, T., Hallgren, K., Grider, M. & Deke, J. (2009, Feb.) *An Evaluation of Teachers Trained Through Different Routes to Certification: Final Report*. Princeton: Mathematica Policy Research Inc. Retrieved March 9, 2009, from http://www.mathematica-mpr.com/publications/PDFs/Education/teacherstrained09.pdf

[3] Gladwell, M. (2008, Dec. 15), "Most Likely to Succeed," *The New Yorker*.

[4] Kristof, N. (2009, Feb. 14), "Our Greatest National Shame," *The New York Times.* Retrieved March 4, 2009, from http://www.nytimes.com/2009/02/15/opinion/15kristof.html.

[5] Mathematica Policy Research Inc. (2009, Feb. 9). "Study evaluates different routes to teacher certification." Press release by Author. Retrieved March 4, 2009, from http://www.mathematica-mpr.com/press%20releases/alternativecertifcation_2_9_09.asp.

[6] Mathematica Policy Research Inc. (2009, Feb. 9). "Study evaluates different routes to teacher certification." Press release by Author. Retrieved March 4, 2009, from http://www.mathematica-mpr.com/press%20releases/alternativecertifcation_2_9_09.asp.

[7] Shadish, W.R., Cook, T.D., and Campbell, D.T. (2003) *Experimental and Quasi-Experimental Designs for Causal Inference.* Boston: Houghton-Mifflin.

Schneider, B. et al. (2007) *Estimating Causal Effects: Using Experimental and Observational Designs.* Washington, D.C.: American Educational Research Association.

[8] Randomized control trials are valued for their ability to remove the confounding effects of non-random selection into treatment and control groups. Because subjects are randomly assigned to these groups, we can assume that "all else is held constant," and infer that differences between groups (short of randomness) are due solely to the treatment.

[9] Murnane, R.J. and Nelson, R.R. (2005, Dec.), "Improving the Performance of the Education Sector: The Valuable, Challenging, and Limited Role of Random Assignment Valuations," *National Bureau of Economic Research Working Paper* #11846

Deaton, A.S. (2009, Jan.), "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development," *National Bureau of Economic Research Working Paper* #14690.

[10] This excluded some high-profile but selective AC programs like Teach for America, based on the rationale that the vast majority TC comparison teachers are not put through comparable selective screens.

[11] For example, see

Lankford, H., Loeb, S., and Wyckoff, J. (2002) "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis* 24(1).

Clotfelter, C.T., Ladd, H.F., and Vigdor, J. (2005) "Who Teaches Whom? Race and the Distribution of Novice Teachers," *Economics of Education Review* 24(4).

Clotfelter, C.T., Ladd, H.F., and Vigdor, J. (2006) "The Academic Achievement Gap in Grades 3 to 8," *National Bureau of Economic Research Working Paper* #12207.

[12] Teach for America and New York City Teaching Fellows teachers (both selective AC programs).

[13] Nye, B., Konstantopoulos, S., & Hedges, L. (2004). "How large are teacher effects?" *Educational Evaluation and Policy Analysis.* 26(3), 237-257. Retrieved March 9, 2009, from http://www.sesp.northwestern.edu/docs/publications/169468047044fcbd1360b55.pdf

For a cross-sectional study of teacher effectiveness in the early grades, based on the Early Childhood Longitudinal Study, see

Croninger R.G. et al. (2007), "Teacher Qualifications and Early Learning: Effects of Certification, Degree, and Experience on First-Grade Student Achievement," *Economics of Education Review*, 26(3).

[14] Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2008) "The Impact of Assessment and Accountability on Teacher Recruitment and Retention," *Public Finance Review* 36(1).

[15] California accounts for 22% of the sample used in the study (p. 15).

[16] Eight of 28 sponsoring institutions of AC programs "also operated TC programs whose graduates were in the same study" (p. 25-27). Surprisingly, the authors "did not systematically explore potential connections or similarities between AC and TC programs operated by the same institution such as the extent to which they required the same courses or the extent to which they shared instructors" (p. 27).

[17] One report highly critical of existing AC programs argues that there is fundamentally no difference between TC and AC programs in the United States. In the forward to this report, Chester Finn and Michael Petrilli write, "[This report's] findings confirm our fears and suspicions. Two-thirds of the [AC] programs that they surveyed accept half or more of their applicants. One-quarter accept virtually everyone who applies. Only four in ten programs require a college GPA of 2.75 or above—no lofty standard in this age of grade inflation. So much for recruiting the best and brightest. Meanwhile, about a third of the [AC] programs for elementary teachers require at least 30 hours of education school courses—the same amount needed for a master's degree. So much for streamlining the pathway into teaching; *these programs have merely re-ordered the traditional teacher-prep sequence without altering its substance, allowing candidates to take this burdensome course load while teaching instead of before*" [emphasis added]. See

Walsh, K. and Jacobs, S. (2007), *Alternative Certification Isn't Alternative,* Thomas B. Fordham Institute and National Council on Teacher Quality.

[18] Rubin, D.B. (1980) "Discussion of 'Randomization Analysis of Experimental Data in the Fisher Randomization Test," *Journal of the American Statistical Association,* 75.

[19] Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2006) "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources* 41: 778-820.

Dee, T.S. (2004) "Teachers, Race and Student Achievement in a Randomized Experiment." *Review of Economics and Statistics*, 86(1): 195–210.

Ehrenberg, R.G., & Brewer, D.J. (1994) "Do School and Teacher Characteristics Matter? Evidence from High School and Beyond?," *Economics of Education Review* 13: 1-17.

Ehrenberg, R.G., & Brewer, D.J. (1995), "Did Teachers' Verbal Ability and Race Matter in the 1960s? Coleman Revisited," *Economics of Education Review* 14: 1-21.

[20] Normal Curve Equivalent differences for black students are 3.45 in math, p=0.09, and 2.35 in reading, p=0.17.

[21] It is not possible using Exhibits II.4 and II.5 to determine the exact number of schools in the study that are in urban areas. However, 15 of the 63 schools are in California, where all of the study's districts are in urban areas, and 27 of the 63 are in Texas, where all but one of the study's districts is in an urban area. All of Georgia and Wisconsin's study districts are in urban areas, as are two of New Jersey's three sampled districts.

[22] Authors' calculations using the Common Core of Data 2004-05.

[23] For example, see

Lankford, H., Loeb, S., and Wyckoff, J. (2002) "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis* 24(1).

Clotfelter, C.T., Ladd, H.F., and Vigdor, J. (2005) "Who Teaches Whom? Race and the Distribution of Novice Teachers," *Economics of Education Review* 24(4).

Clotfelter, C.T., Ladd, H.F., and Vigdor, J. (2006) "The Academic Achievement Gap in Grades 3 to 8," *National Bureau of Economic Research Working Paper* #12207. Clotfelter and his colleagues show that over two-thirds of the black-white gap in exposure to novice teachers can be attributed to within- rather than between-district differences.

[24] See Exhibits II.4 and II.5, and page 21 of the report.

[25] There is very little empirical evidence on schools' staffing responses to testing and accountability, but one recent paper found that teachers assigned to teach 4th grade in New York State were less likely be inexperienced once mandatory 4th grade tests were implemented. See

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2008) "The Impact of Assessment and Accountability on Teacher Recruitment and Retention," *Public Finance Review* 36(1).

[26] It is possible, of course, that these schools assign the most effective AC teachers to the tested grades, in addition to the most effective TC teachers. In that case, there's an assignment bias issue on both sides of the equation.

[27] For example, see Goldhaber, D., Gross, B. and Player, D. (2007) "Are Public Schools Losing Their 'Best'? Assessing the Career Transitions of Teachers and Their Implications for the Quality of the Teacher Workforce," *Center on Reinventing Public Education Working Paper* #2007-2.

[28] There is very thin empirical evidence on this pattern. But one legislative study from Texas on this subject can be found. See

Herbert, K.S. & Ramsay, M.C. (2004, September). Teacher Turnover and Shortages of Qualified Teachers in Texas Public School Distrists, 2001-2004: Report to the Senate Education Committee. State Board for Educator Certification. Retrieved March 5, 2009, from http://www.sbec.state.tx.us/SBECONLINE/reprtdatarsrch/ReportforSenateEducationCommittee.pdf

The turnover rate among teachers was high in the Mathematica study: 7 of 93 AC teachers in the study (7.5%) left *during the school year*, while 5 of 95 TC teachers did (5.3%).

[29] The math test administered to students in grades 2-5 included two components, "Math Concepts and Applications" and "Math Computation." However, because the Math Computation test was not available for students in K-1, the authors used only the "Math Concepts" score in the analyses reported in the primary body of the report, "for comparability across grades" (p. A-4).

[30] McCloskey, D. & Ziliak, S. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ann Arbor: University of Michigan Press.

The Think Tank Review Project is made possible by funding from the Great Lakes Center for Education Research and Practice.