

The Value Added by Teacher Education

Mary M. Kennedy¹
Soyeon Ahn
Jinyoung Choi
Michigan State University

Citation:

Kennedy, M. M., S. Ahn and J. Choi (in press) The value added by teacher education. In M. Cochran-Smith, S. Feiman-Nemser, and J. McIntyre (Editors). Handbook of Research on Teacher Education: Enduring Issues in Changing Contexts. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

The Value Added by Teacher Education

Mary M. Kennedy²
Soyeon Ahn
Jinyoung Choi
Michigan State University

Teaching is the sort of work that both inspires and mystifies, and many people have tried to articulate the qualities they believe make someone a good teacher. Some say a good teacher is bright, others that she is caring, others something else. These speculations take on a more practical quality when the conversation turns from what makes someone a good teacher to how to prepare good teachers, for the preparation question can lead to detailed curriculum and program specifications. So literature on the question of how to produce good teachers includes philosophical inquiries, program designs, and empirical tests of hypotheses.

Arguments about how to prepare teachers have become especially shrill in recent years, as observers become increasingly concerned about the quality of the education system as a whole. Debates have also been stimulated by a new body of work called “value-added” analyses, which examine variations among teachers’ classroom effectiveness. From these analyses, we know that teachers vary considerably in the amount their students learn, so much so that if a student had two or three consecutive weak teachers, his overall academic achievement would be seriously compromised (Aaronson, Barrow, & Sander, 2002; Rokoff, 2003; Sanders & Horn, 1998). These studies demonstrate that teachers differ substantially in their effectiveness, and raise to prominence the question of how to better prepare teachers.

The aim of this paper is to examine empirical evidence regarding the merits of the most prominent hypotheses about educational backgrounds that will improve teacher effectiveness. One hypothesis, the one that dominates most state regulations, is that teachers need specialized knowledge about issues directly pertinent to teaching—things like classroom management, techniques for teaching, the role of school in society and other educational issues. We call this hypothesis the *Pedagogical Knowledge* hypothesis. Virtually every state subscribes to this hypothesis by requiring prospective teachers to take courses from departments of teacher education whose mission is to prepare people specifically for teaching careers. But although

this hypothesis is widely represented in state regulations, it is not without its detractors. In particular, two other hypotheses are offered as alternatives. One argues that teachers need *Content Knowledge* more than pedagogical knowledge. Proponents of this hypothesis frequently note that teachers cannot teach content if they do not know it. While it is possible in principle for teachers to obtain Pedagogical Knowledge as well as Content Knowledge, advocates for Content Knowledge frequently pit the two against one another under the assumption that courses providing Pedagogical Knowledge take up too much space in the college curriculum and hence remove space for courses providing Content Knowledge that would ultimately be more beneficial to prospective teachers. There is, among those outside the teacher education community, a general skepticism about the merit and value of teacher education (Conant, 1963; Damarell, year; Hess, 2001; Kramer, 1991; Labaree, 2004; Lagemann, 1999).

In the past two decades, a third hypothesis has been put forward that suggests that teachers need a blend of pedagogical and content knowledge, something called *Pedagogical Content Knowledge*. This knowledge consists of such things as how students understand, or misunderstand, particular substantive ideas, how to present particular substantive ideas in a way that makes them more accessible to different types of students, or how to use particular resources in lessons about particular content.

All of these hypotheses focus on the college curriculum, suggesting that there are particular domains of knowledge that can make a difference, and that teachers should take courses in these domains. There is another hypothesis that challenges the validity of all of these. This fourth hypothesis argues that the best teachers are *Bright, Well-Educated* people who are smart enough and thoughtful enough to figure out the nuances of teaching in the process of doing it. For people subscribing to this hypothesis, the route to improving the quality of teaching lies in recruitment, not in specific courses that will prepare people for this work.

Advocates of the first hypothesis tend to acknowledge all of the rest as well. That is, they rarely argue *against* content knowledge, *against* pedagogical content knowledge, or *against* the value of having bright, well-educated people teaching in the nation's classrooms (see, e.g. National Commission on Teaching and America's Future, 1996). However, advocates for content knowledge and for bright, well-educated people do argue against the pedagogical knowledge and against pedagogical content knowledge hypotheses. So the debate really

focuses mainly on the merits of allocating college curriculum space to courses specifically about teaching.

We take the view that all of these hypotheses need to be examined. Therefore, we seek evidence about all of them, rather than confining ourselves to the most controversial one. For the first three hypotheses, we examine studies of teachers' college course-taking histories. We assume that courses in education represent pedagogical knowledge, courses in mathematics represent content knowledge, and courses in mathematics education represent the pedagogical content knowledge hypothesis. Because our fourth hypothesis deals more with recruitment than with college curriculum per se, we need to draw on studies that examine the selectivity of the teachers' alma maters.

In this investigation, we confine our inquiry to the content area of mathematics, and take student achievement in mathematics as our indicator of teachers' effectiveness. This limitation is necessary in part to ensure that all curricular options are tested against a common outcome. We also take measures of teachers' educational backgrounds as our independent variable. The paper tests hypotheses only about teachers' educational background, not about tested knowledge.

Studies examining the benefits of college course-taking tend to start with a population of teachers, inquire as to the number of courses teachers took in one subject or another, and then look at the relationship between these curricular experiences and their current students' achievement. Studies examining the merits of the bright, well-educated person hypothesis tend to contrast groups of teachers who enter the profession from different educational backgrounds. We located two varieties of studies that are relevant to the Bright, Well Educated Person hypothesis. One group contrasts teachers according to the selectivity of their alma maters, the other contrasts teachers who were specifically recruited by Teach for America (TFA), a program that tries to recruit bright, well-educated people and persuade them to teach for two years. Comparisons of TFA teachers and regularly-certified teachers, therefore, offer us another way to test the bright, well-educated person hypothesis.

The studies reviewed here come from a larger collection of literature addressing the role of a wide range of teacher qualifications to the quality of teaching. Below we describe our search

procedures for gathering this literature and discuss some of the methodological problems in this literature. We then review findings for college curricula and for recruitment.

Literature Search Procedures

Literature for this paper was drawn from a larger literature data base gathered as part of the *Teacher Qualifications and the Quality of Teaching* (TQQT) study. The TQQT data base includes studies that examine the relationship between at least one teacher qualification and at least one indicator of the quality of teaching. Our original search criteria defined *qualifications* to include such aspects of teachers' educational backgrounds as their college curricula, test scores, credentials, grade point averages and degrees. Indicators of teaching *quality* included such things as direct observation of classroom practice, student achievement, and principal ratings, as long as these indicators were obtained after the teacher had a full time teaching position. As we worked in this area a few other special categories of studies were added. We expanded our list of qualifications to include some things that typically are acquired *after* teachers obtain full-time teaching positions because these things may be relevant in district hiring or salary decisions. These include, for instance, teachers' years of experience, whether or not they possess an advanced degree, whether they had been certified by the National Board of Professional Teaching Standards, and how they responded to a commercial hiring interview such as the Teacher Perceiver Interview.

Excluded from this compilation were studies of student teachers, preschool teachers, and teachers of college students and adults. Also excluded were studies published prior to 1960 and studies conducted outside the United States, on the grounds that these contexts may be too different from contemporary United States to be applicable. For our tests of the four hypotheses about educational backgrounds, we also eliminate studies published before 1980, on the grounds that curricula change over time and earlier studies may no longer be applicable.

Also excluded from the larger collection are studies whose indicators of quality did not emanate directly from classroom practice. That is, we excluded numerous studies that examined a relationship between, say, teachers' college courses and their own test scores. In our view, test scores are qualifications, and do not derive directly from teaching practice. This

decision also required us to omit a handful of studies that obtained indicators of quality from exercises and simulations of classroom life. In the TQQT data base, most indicators of quality are derived either from student test scores or from observations of teaching practice.

Literature was obtained by searching the Education Resource Information Center (ERIC), PsycInfo, Dissertation Abstracts International, and EconLit. Search terms included those commonly used to define either qualifications of teaching quality, such as assessment, certification, teacher education, teacher effectiveness and so forth. In addition, we searched the bibliographies of these articles as well as those of literature reviews and policy analyses in this area and searched recent issues of entire journals whose domain encompassed this area. Studies were screened to ensure that they included at least one qualification and at least one indicator of quality, and that their analysis explicitly linked the two. Links could be established with group comparisons (e.g., a group of traditionally certified teachers versus a group of alternatively certified teachers); correlations, various multivariate strategies, or through qualitative approaches. As of this writing (Winter 2005) the data base included about 480 studies. More details about the study and the data base can be found at <http://www.msu.edu/~mkennedy/TQQT>.

To test our four hypotheses about teachers' educational backgrounds, we searched this literature for two types of studies. First, to test hypotheses about the college courses, we sought studies that tried to assess the relationship between the courses teachers took in college and the teachers' current effectiveness. We sought tallies of courses taken in education, mathematics, or mathematics education. We also sought tallies of whether teachers had majored or minored in either of these subjects or held advanced degrees in them. The bright, well-educated person hypothesis cannot readily be tested by examining college curricula, but we did find two types of studies that are relevant to that hypothesis: those that measure of the status, or selectivity, of the institution which teachers had attended for their college education, and those that looked at the effectiveness of teachers who were recruited by Teach for America. Because the TFA recruits from prestigious colleges and universities (Raymond & Fletcher, 2002b), we consider these studies to be tests of the bright, well-educated person hypothesis.

Methodological Issues

Research on the relative value of different types of teacher preparation has been impeded by some very difficult methodological problems. One problem is that many of the credentials of interest don't really vary much. For example, the value of a bachelor's degree would be difficult to assess in the United States because some 99% of teachers already have a bachelor's degree (National Center for Education Statistics, 2003). A test of the value of this degree would require an extensive search to find teachers who *lacked* the degree, to serve as a comparison group for those who have it. In many important respects, the teaching population in the United States is quite homogenous, a phenomenon that is hard to reconcile with the large differences in effectiveness that we see. We know that about 75% of them are women and 84% are white (National Center for Education Statistics, 2003). We also know that 99% of teachers have a bachelors degree and that 90% of teachers are certified.

However, certificates can have quite different meanings across educational institutions and across states. States have a plethora of certifications on their books, often exceeding a hundred. And they differ in the curricula they require for any given certificate (Ballou & Podgursky, 1999; Council of Chief State School Officers, 1988; Rotherman & Mead, 2003). For this investigation, we prefer to examine teachers' actual course taking rather than formal credentials. We do this in part because there is a very uncertain relationship between college course-taking and certification status. Thus, even though the vast majority of teachers have some certification, we may find enough variation in their curricular backgrounds to inform us about the value of different curricula. This strategy helps us increase the variability in our measure of educational background, even though we still cannot know what teachers actually studied in their courses.

The second important problem affecting all studies of qualifications is that teachers decide for themselves which qualifications they will obtain. Researchers do not randomly assign teachers to colleges or universities, nor to their college majors. Thus measurements such as the number of courses taken on topic X or Y reflect not only the knowledge teachers gained from these courses but also their initial interest in those topics. This would apply when studying any teacher qualifications, because we can never separate teachers' initial interest in a subject from the knowledge they have acquired about it. Thus, if we find that courses in a certain topic seem

to lead to better-quality teaching, we cannot infer that if states *require* teachers to take the same courses, that the same results will obtain. Once the courses are required, we will find teachers who took the courses but did so without interest in them. This problem plagues the present examination. Prospective teachers take some college courses because they are required to, but they take others because they want to, and it is likely that these two reasons bear on the degree of benefit that teachers derive from their courses.

A similar self-selection phenomenon occurs when teachers seek employment, for their assignment to schools is also not at all random. In fact, there is now considerable evidence regarding the way teachers and schools become matched. We know, for instance, that most young people who choose teaching as a career are white women who come from rural and suburban communities. We also know that, upon graduation, they seek positions in schools that are similar to those they attended when they were students themselves (Boyd, Lankford, Loeb, & Wyckoff, 2003). They engage in a process of *differential migration*, such that college graduates who grew up in the suburbs seek teaching positions in the suburbs, and those who grew up in small towns seek teaching positions in small towns. We also know that school systems give preferences to their own graduates (Strauss, 1999), so that both schools and teachers are seeking to match each others' cultural and demographic fundamentals. Finally, we know that urban schools, and other schools serving lower income and non-white populations, have greater difficulty filling their vacancies, have higher turnover in their teaching staffs, and are more likely to employ teachers with fewer qualifications (Boyd, Lankford, Loeb, & Wyckoff, 2002; Lankford, Loeb, & Wyckoff, 2002; Wyckoff, 2001). As a result they tend to have more novices than other schools and to have less-qualified teachers as well. These processes, taken together, suggest that the positions teachers eventually take are likely to be *affinity assignments*, such that their social backgrounds and their qualifications match the social backgrounds and qualifications of their students.

Now imagine how these processes influence research. We want to see if teachers with different types of qualifications have different influences on their students. But teachers with different qualifications have already been matched to students with different qualifications, through these natural processes of differential migration and affinity assignments. One result of these processes is that the schools whose students need the most educational help—who are

the least qualified as students--are allocated the least qualified teachers, while those who serve the most advantaged students are allocated teachers with more qualifications. If we examine a simple correlation between, say, the prestige of the teachers' alma mater and their students' achievement test scores, we would likely see a relationship *even if the teachers have had no impact at all on their students*, simply because they have been matched to their students through these processes of differential migration and affinity assignments. The migration and assignment processes make it difficult for researchers to tell whether teachers created their students' achievement levels or whether, instead, students with different achievement levels have attracted different kinds of teachers.

Researchers often try to separate out these various influences statistically. They try to measure all the things they expect to influence teachers' effectiveness, and include them in a statistical model. Their models include not only teachers' qualifications, but also measures of student's socio-economic status, race or ethnicity, perhaps measures of the school's size, finances, or demographic makeup, and perhaps other characteristics of the teachers themselves, such as their beliefs or values. In these studies, the goal is to find a relationship between teachers' qualifications and the quality of their teaching practice *after measuring and accounting for other relevant influences*. The value of their findings depends heavily on the number and relevance of the other factors that they measure and include in their models. This is why we require our studies to use more sophisticated models: models that take into account not only the teachers' qualifications, but also important features of the school and student populations. Such models are not a guarantee that the findings yield causal influences, but they are more able to measure and take into account other aspects of teaching situations.

The need to control for numerous extraneous factors brings up yet another methodological problem, and that is that researchers often rely on pre-existing data bases to conduct their studies, rather than collecting new data specifically for the purpose of addressing this research question. This reliance on pre-existing data means that the models researchers develop are limited to those variables that happened to be included in the original data base, *whether or not these variables are the best or most relevant factors to consider*. So researchers who rely on state or district data bases are likely to include measures of teachers' years of experience but not measures of their attitudes or values. And they may include information about students'

eligibility for free or reduced lunches, but not whether their parents help them with their homework. They use these measures not because these are necessarily the best factors to include, but because the data base they are using happens to include them. This reliance on extant data bases means that many models do not take account of the main factors that a reasonable person might expect to be relevant to the question.

Finally all studies of this issue face complications that arise from numerous influences that intervene between teachers' prior college experiences and their current teaching practices. Any given sample of teachers will include some relatively new teachers, some with a few years of experience and some with many years of experience. Teachers may have been influenced by many things since they obtained their original college degrees and certifications. They may have worked in different schools, taken different kinds of professional development courses, faced parents with different expectations for their children. Any of these experiences may have influenced their practices and their effectiveness. Yet our research is ignorant of these influences and seeks evidence of only one influence, their college education, and that may have occurred years ago. Indeed, time itself is an intervening influence, for teachers who obtained their certificates 20 years ago may have experienced very different educational programs than teachers who obtained their certificates 5 years ago. And studies that were published 20 years ago may have included teachers whose preparation was quite different from that of teachers prepared today (Metzger, Qu, & Becker, 2004).

Because of these complications, we restrict our search to studies that meet three important criteria.

1. *The studies must use relatively sophisticated models*, models that allow the researcher to take account of numerous other relevant possible influences. These studies typically rely on either multiple regression or hierarchical linear models, both of which allow more opportunity to measure additional influences and take them into account than do simple correlations or group comparisons.

2. *The studies must include a pretest as one of their factors*. Measures of student achievement are like snapshots taken at a specific time. They measure student knowledge at a particular time, but that knowledge reflects the sum of everything students have learned throughout their entire lives, from parents, other teachers, other adults and from each other. Moreover, they can

present a relationship that reflects the effect of affinity assignments more than the effects of teachers' influence on student learning. We therefore prefer studies that incorporate a pretest score as one of the factors in their model, or as part of a gain score. The inclusion of some measure of prior achievement enables the researcher to separate the knowledge students gained prior to studying under the teacher who is the focus of the study.

3. *The studies must focus on individual students or teachers rather than institutional collections of teachers.* We found that many researchers relied on school averages or district averages of teachers and students, rather than using data from individual teachers. Aggregated data are often used when researchers borrow data bases that belong to school districts or states. Such data bases may not include individual-level information, or may restrict access to it to protect teachers' privacy. These aggregated data conceal all of the variations among teachers within each school or district and reveal variations among schools that may be due to many things other than instructional influences, things researchers have no knowledge of (for more on this issue, see Murnane, 1981; Hanushek, 1996). As a result, they may lead to erroneous estimates of the relationship between teachers' qualifications and the quality of teaching practices. We therefore limit our attention to studies that focus on individual teachers and students.

The studies we examine here represent what we consider to be the best evidence of the relationship between teachers' qualifications and their students' mathematics achievement. We do not limit ourselves to peer-reviewed journal publications. Instead, we rely on independent criteria of quality. Our criteria for best-evidence studies are those that (a) rely on relatively sophisticated models, usually multiple regression or hierarchical linear models, since these can more easily take account of additional influences in the system; (b) include a pretest in their model or use gain scores; and (c) use teacher or student level data rather than school, district or state averages. These criteria have been recognized by numerous researchers as indicators of better research designs (see, e. g. Goldhaber & Anthony, 2003; Hanushek, 1971; 1996; Murnane, 1981). In this respect, we follow Slavin's (1984, 1986) suggestion that reviewers should focus on best evidence, and we agree with him that what constitutes best evidence depends on the research questions and contexts.

These are not remarkably strict standards, for the studies remaining in our pool could still contain samples of teachers who are too homogeneous to reliably inform us about the merits of

particular qualifications, and they may still have poorly specified statistical models. But we have weeded out those studies with the most egregious flaws. Table 1 lists the studies that meet our criteria for best evidence. Notice that these studies still differ in the grade levels of the teachers they studied and in the number and type of other variables they were able to measure, and hence control, in their analysis.

Table 1 here

Findings are presented in two sections. In the first, we summarize research that examines the influence of teachers' curricula on their students' mathematics achievement. These studies allow us to evaluate hypotheses about the relative importance of pedagogical knowledge, content knowledge, and pedagogical content knowledge. In the second section, we examine two sets of literature that are relevant to the bright, well-educated person hypothesis. These include multiple regressions that measure the status or selectivity of the institutions that teachers attended, and a handful of studies that contrast TFA teachers with other teachers.

One final note is needed about the findings we present here. A synthesis of this sort requires us not only to define our criteria for including studies, but also to define criteria for selecting specific findings from each study. Researchers frequently report multiple models, varying the characteristics of their statistical models. They sometimes provide findings for multiple subgroups (e.g., high-achieving and low-achieving students) or for multiple achievement sub-tests. To avoid over-representation of any one study, we selected a single model from each grade or sample examined. Our decision rules were to first seek models that met our criteria of including a pretest and focusing in individual teachers rather than aggregations. If an author presented one model using a gain score and another using the pretest as a covariate, we chose the covariate model. Next we sought findings that applied to the total group, rather than to specific subgroups and that used total test scores rather than sub-test scores. Along the same line, we avoided models that included interaction terms. If different models from the same sample are presented in different publications, we selected the most recent version. Otherwise we selected the model that included the largest number of curriculum measures or, absent differences in this, the one that was most complete with respect

to other variables. If, however, a researcher reported one-year findings for each of two grade levels, and then also reported a two-year gain for the group as a whole, we eliminated the two-year finding in favor of the two one-year findings, since our goal is to attach student learning as closely as possible to the years in which teachers actually taught the students.

Evidence for the Influence of College Courses

When people suggest that teachers need more knowledge of, say classroom management, they really mean more knowledge of this *relative to other areas of knowledge*. And when people argue that teachers do not need knowledge in an area, they rarely mean that this knowledge is absolutely lacking in value, but rather that it is not as valuable as some other knowledge would be. Defined in this way, the empirical question then becomes one of identifying the value added by different parts of the teachers' total college education, where one part might be courses in education, another might be courses in subject matter (in our case, mathematics), and a third might be courses in math education, or how to teach mathematics³.

The first problem we face in our effort to summarize this literature is that outcomes in these studies can depend heavily on the combination of other variables the researcher has chosen to measure and control in his model. Thus it is not possible to directly compare one study's estimate of the effect of courses in education with another study's estimate of the effect of courses in mathematics. We approach this problem in two ways. First, we examine studies that test more than one hypothesis within a single statistical model. These studies allow a direct comparison of different course content under the same methodological conditions. Our second strategy arrays findings from studies conducted under more varying statistical conditions. We have already made these studies somewhat more comparable by requiring that they all fall within our "best evidence" boundaries. We also increase comparability by examining all effects relative to student achievement on that particular test.

The second problem we face has to do with defining the meaning of an "effect" on student test scores. We want to see whether teachers who took more courses have students with higher scores than other students. But how much higher? Certainly the apparent influence of a college course will look different when student achievement is measured on a 50-

point scale than when it is measured with a 10-point scale. The conventional method for solving this problem is to divide all effects by the groups' standard deviation, thus creating a scale that puts all effects on a common metric. But this approach can only work when all researchers derive their standard deviations in the same way (Hedges, 1986). If one researcher provides a pooled within-school estimate of variation while another provides an estimate that includes between-school variation, then the standard deviations will not convert effect sizes to a comparable metric. Even if the standard deviations were comparable, we face another problem when studies rely on national data bases and their associated tests, for these tests are not aligned with local curricula. They will likely measure some content that was not taught and fail to measure some content that was taught. Because they miss some areas the teacher has influenced, and measure some areas teachers have not tried to influence, we expect these tests to underestimate teachers' actual instructional influence on students. Hence their effects are not comparable to effects that are based on state or local tests which are designed to reflect the local curriculum. In fact, this under-estimation is apparent in the average annual gains shown in national data bases, which are often just one or two points for the entire school year.

Our solution to these problems is to define all outcomes relative to the average amount that students gain during a given school year. That is, instead of standardizing effects by dividing them with the standard deviation, we standardize by dividing them with the average annual gain. For instance, if majoring in mathematics adds 1 point to students' achievement scores, and if students typically gain 3 points during the year, then we would say that the effect of the major was equivalent to 1/3, or 33%, of students' annual gain. Our estimates of effects, then, are based on the following equation:

$$\frac{b_i}{\overline{Y}_{post} - \overline{Y}_{pre}}$$

where

b_i is the unstandardized regression slope for a particular qualification taken from a particular sample; and

$\bar{Y}_{post} - \bar{Y}_{pre}$ is the gain in mathematics achievement for the population from which the sample was taken⁴.

This strategy cannot solve all problems associated with variations in test designs, of course. But it does offer certain advantages over other standardizing strategies. We do not have to eliminate studies because they calculate their standard deviations in different ways, for instance. And we do not have to worry about differences among outcome tests in how closely aligned they are to local curricula because these differences are accounted for by the use of average gains.

A. Within-study comparisons of hypotheses.

Here we examine two studies that meet our best-evidence standards and also test more than one type of curriculum content. One (Monk, 1996) examines secondary mathematics teachers and the other (Rowley, 2004) examines kindergarten teachers.

Monk used multiple regression, a statistical technique that allows researchers to estimate the contributions of several different potential influences within a single equation. Monk's equations included student background variables as well as several measures of teachers' educational backgrounds. He estimated the influence of individual courses teachers took as well as the influence of majoring in mathematics. Since Monk studied both sophomores and juniors, we review two equations, one for each group. Figure 1 displays his findings.

Insert Figure 1 here

Each hatch-mark in Figure 1 represents the amount that students' test scores increased (or decreased) with each additional course their teachers took in math or math education. For example, the top line indicates that, among teachers of sophomores, each additional course taken in math education—that is, methods of teaching mathematics--appears to have raised student achievement by an amount equivalent to about 10-15% of sophomores' average achievement gains (the actual percentage was 11%). Courses in mathematics also added to student achievement, but added a smaller amount. In this study, the average math teacher took about 8 courses in mathematics but only 2 in mathematics education. If a teacher had taken two courses in mathematics education, would her students could gain 22% more than

other students? If a teacher took 8 courses in mathematics would her students gain 8 times as much as from a teacher who had taken one mathematics course? We cannot expect these patterns to generalize to larger numbers of courses.

Such speculations are also tempered by the fact that when teachers majored in mathematics, their students actually gained less than other teachers' students. Moreover, there were substantial differences between findings for sophomores and those for juniors, so that majoring in mathematics appears to have a negative effect on student learning. Monk suspected that there might be a limit to the benefits of taking additional courses in mathematics, and tested this idea with his sample of juniors. He found that the benefits of additional courses in mathematics tapered off after about 5 courses. In other words, the first five courses in mathematics show distinct, additive benefits to student learning, but additional courses after these did not show any apparent benefit. In fact, they appear to have a negative effect.

Here we face one of the difficulties of squaring research findings with our common sense. It is hard to imagine that additional study in any topic could actually make someone a *less able* teacher. Instead, we can seek some other explanation for this pattern. One possibility is that the negative effects are an artifact of the statistical model itself. Another is that additional courses aren't really harmful, but merely not helpful. The third is that people who chose to take numerous courses are already different from other college students even before they take the courses. Perhaps college students who choose to major in mathematics have different personalities, values or beliefs that render them less effective as teachers, and it is these initial differences that account for the apparent negative effects, not the courses themselves. This pattern of findings reminds us of the two important interpretive difficulties in this type of research: The effects we see in these studies must be interpreted in the context of the entire statistical model, and these measures of course-taking reflect *both* the knowledge gained from the courses *and* the teachers' original interests and dispositions that motivated her to take these courses in the first place.

Now let's consider Rowley's (2004) findings, shown in Figure 2. Rowley examined kindergarten teachers and looked for influence from several different types of courses. Her outcome was a test of mathematics designed for kindergarten-level students and she employed a statistical technique called Hierarchical Linear Modeling (Raudenbush & Bryk, 1988). This

technique based on different assumptions, but yields regression coefficients that can be interpreted in much the same way.

Figure 2 here

Each hatch mark in Figure 2 represents the amount of increase or decrease in student achievement that is associated with courses teachers took in these various domains of knowledge, and each is expressed as a percentage of students average gains for the entire year. Notice that the scale is much narrower than the scale for Monk's findings. To display Monk's findings, we needed a range from -75% to +100% to include all our data, but here we need a range from only -6% to +6%. Figure 2 suggests that additional courses in child development were least beneficial— in fact, they had negative effects—and that additional courses in early education and in elementary education were most beneficial. But even the most beneficial courses were adding only 1% to students' annual achievement gains, effects much smaller than those Monk found at the secondary level.

The apparent negative effect shown in Figure 2 suggests that each additional course taken in child development further diminishes a teachers' ability to actually teach young children. The finding reminds us again that measures of course-taking reflect both the influence of the course and the prior dispositions of those who choose to take it. It is possible that the negative hatch mark we see in Figure 2 reflects the predispositions of the teachers who took these courses, rather than the effects of the courses themselves.

B. All Best-Evidence Studies

Our second approach to reviewing studies of teachers' course-taking gathers together all the research findings that meet our "best evidence" standards into a single graph. Because our synthesis includes far more hatch-marks than the earlier figures, we separate out the effects of courses in each subject from the effects of majoring in a subject or taking an advanced degree in it. Figure 3 summarizes all evidence regarding the number of courses teachers took, and Figure 4 summarizes all the evidence regarding intensive study in a particular area. In these figures, each hatch mark represents a single influence found in a single study. Some studies, such as Rowley's and Monk's, are represented by several hatch marks because they tested multiple influences. Other studies provide us with just a single hatch mark. In the case of Rowley's

study, three of her four course content areas are collapsed here into the domain of pedagogical knowledge, or education, and they are the only hatch marks shown in Figure 3 for education courses taken by elementary teachers.

Viewing the pattern of hatch-marks, we need to remember that the size of these coefficients depends in part on what else is in the original equations, so to the extent that these studies devised different equations, and controlled for different aspects of students, we would not expect their coefficients to be exactly the same. This is one reason why multiple estimates are informative.

Insert Figure 3 here

Figure 3 reveals several important patterns. First, most of the influences of teachers college courses, regardless of their content, are small but positive. Each course adds to student achievement an amount less than 5 percent of the students' average annual gain, with the exception of Monk's unusually large estimates. Second, the estimates for secondary teachers are more various than those for elementary teachers. Even if we remove the one outlier—the Monk estimate that each math course adds the equivalent of 95% of students' average annual gain, the remaining estimates are more spread out than estimates for elementary teachers. And third, the domain of pedagogical knowledge, represented here by courses in education, tends to reveal smaller effects than the other two domains.

These patterns can be contrasted with those in Figure 4, where we summarize the effects of obtaining a full major in one of these subjects, or of obtaining an advanced degree in either subject. Before turning our attention to Figure 4, we should mention that there is one study we were unable to display in Figure 4 because we could not convert its findings to a metric that was comparable to our other findings. This study (Rowan, Correnti and Miller, 2002) examined elementary teachers and found that holding an advanced degree in mathematics (in contrast to no degree in mathematics) had a negative effect on students' achievement. It is the only study that examined the influence of majoring in mathematics on elementary teachers.

The pattern in Figure 4 is limited to secondary teachers and suggests far more variability in estimates of influence than we saw in Figure 3. Since these hatch-marks represent the effects of concentrated study in one of these domains, and probably also reflect the inclusion of more

advanced courses, we might expect their influences to be substantially larger than those of individual courses summarized in Figure 3. Indeed, most of these effects are larger than those in Figure 3, but not by much. Moreover, many are negative as well as large, something that rarely occurred in Figure 3. Figure 4 suggests that perhaps Monk is right, and that after some point, additional courses yield diminishing returns which are not accurately reflected in models that assume uniform additional benefits from each additional course.

Insert Figure 4 here

Only at the level of advanced degrees do we begin to see a difference across domains of knowledge. Advanced degrees in education are more likely to have a negative influence on student achievement in mathematics, whereas advanced degrees in mathematics are more likely to have a positive influence on student mathematics achievement. However, we need to add one caveat to this observation: All of these estimates of advanced degrees come from a single data base, the National Educational Longitudinal Study of 1988. This data base is weaker than others in that the students took their pretest two years prior to their post test, and therefore probably received instruction from other mathematics teachers during this time interval. Moreover, Ludwig & Bassie (1999) have argued that this data base produces biased estimates. They are convinced that the pretest is not adequately controlling for non-random assignment of students to their teachers. Since all of our estimates of the influence of advanced degrees comes from this one data base, we may want to consider these findings especially cautiously.

Generally speaking, Figure 4, along with the Rowan, Correnti and Miller (2002) study, does raise some questions about the role of intensive study in any one subject. The fact that intensive study in a subject can have a negative effect on students' achievement when individual courses in that same subject appear to have a positive influence presents a conundrum. As before, though, we are chary to attribute negative consequences to the direct influence of the courses themselves, and instead inclined to suspect that different types of people seek out intensive study in the first place. Or perhaps the relationship is an artifact of differential migration and affinity assignment processes, such that teachers with advanced degrees in education are more likely to be placed in secondary mathematics classes with less qualified students.

Summary: In this section, we have examined evidence regarding three hypotheses about the kind of knowledge teachers need. We are interested in the difference between pedagogical knowledge, pedagogical content knowledge, and content knowledge. We used coursework in education to test the hypothesis that pedagogical knowledge makes a difference, coursework in mathematics education to test the hypothesis that pedagogical content knowledge makes a difference, and coursework in mathematics to test the hypothesis that content knowledge makes a difference. We have also reviewed and considered a variety of methodological problems that complicate research in this area.

Three conclusions are consistent with the patterns we have examined so far: First, at the individual course level, all three domains of knowledge appear to improve teachers' effectiveness. The pattern suggests that each additional course adds small benefits, worth up to 5% of students annual achievement gains. Second, at the individual course level, the domain of pedagogical knowledge, as represented by courses taken in education, has a smaller effect than the other two domains. Third, the effects of intensive study in these domains look different. In each subject, estimates of the effects of intensive study are more widely dispersed, ranging from large negative effects to large positive effects. These findings are complemented by another study (Rowan, Correnti and Miller, 2002) that found negative effects of advanced study in mathematics among elementary teachers. We also see that, at the advanced degree level, intensive study in education appears to have a negative influence on student learning and that advanced degrees in mathematics appear to have a positive influence. However, each of these findings comes from a single data base.

We are inclined to believe that these apparent negative effects are not due to advanced knowledge per se, but instead are due to differences in the personalities, interests or values people who chose to pursue advanced study in the first place. Perhaps those who seek advanced study in mathematics have characteristics that don't suit them to the elementary classroom, and perhaps those who seek advanced degrees in education have characteristics that don't suit them to the secondary classroom.

Evidence for the Bright, Well-Educated Person hypothesis

We rely on two very different types of studies to test the bright well-educated person hypothesis. One group of studies examines the influence of the status, or selectivity, of teachers' alma maters on their current students' achievement. The other examines the effectiveness of teachers who were recruited into the profession via the Teach for America (TFA) program, a program that explicitly recruits people presumed to be bright and well educated, but who have not necessarily studied teaching.

A. Institutional Status

Two studies examined some indicator of institutional status, or selectivity, in their attempts to account for teacher effectiveness. They differ in how they define and measure institutional status, however, so each must be examined separately. Figure 5 and 6 show the pattern of changes in students' mathematics achievement that are associated with the status of their teachers' alma maters.

Insert Figure 5 here

Insert Figure 6 here

The first study (Clotfelter, Ladd and Vigdor, 2002a, 2002b) is the only study in our data base to explicitly recognize that teachers and students may already be matched by their social class and qualifications even before any data are collected. These authors checked individual schools to see how students were assigned to teachers. They found that many schools had assignment policies that placed less advantaged students with different teachers than more advantaged students. In Figure 5, we see the difference in the apparent benefit of the status of teachers' alma maters, as classified by Barrons' College Ranks, when examined with the full sample and again when examined only in those schools that relied on random assignment practices⁵. The figure suggests that higher alma mater status is indeed associated with higher student achievement in the full sample, where affinity assignments occur, but that the opposite pattern appears when the study is restricted to schools that randomly assign students to teachers. For this sub-set of schools, Figure 5 suggests that teachers from the most prestigious alma maters may actually be less able to foster learning in their 5th grade students.

Figure 6 summarizes findings from the second study (Aaronson, Sander and Barrow, 2003), which focuses on 9th grade mathematics teachers and defines institutional status with a scale based on US News and World Reports classifications. We might expect affinity assignments to be more of a problem in secondary schools than in elementary schools, since secondary schools tend to engage in more tracking, but Aaronson and colleagues tested for this possibility by comparing their observed classroom assignments with a variety of simulated assignment possibilities. Based on these analyses, the authors argue that their data did not suggest systematic classroom sorting. With respect to outcomes, the pattern of influence revealed by Aaronson's study suggest that institutional status does make a difference, and that students' mathematics achievement tests scores were higher when their teachers had attended higher-status institutions. But the pattern is also uneven and suggests that only the level 5 institutions stand clearly apart from the others. This may be due to unusually small samples in some of the status categories. We are also skeptical of the author's claim that students and teachers were not systematically assigned, and suspect that the pattern may reflect some processes of affinity assignments. In either case, the combination of this study with the one above does not provide strong support for the bright, well-educated person hypothesis.

B. Teach for America Recruits

The second type of study that is relevant to the BWE hypothesis examines teachers who were recruited into the profession by Teach for America (TFA). On its website, TFA defines its recruitment strategy as follows:

Each year, Teach For America launches an aggressive effort to recruit the most outstanding graduating college seniors and recent college graduates - people who will be the future leaders in fields such as business, medicine, politics, law, journalism, education, and social policy. We seek a diverse group - socio-economically, racially, ethnically, politically, and in every other respect. We seek leaders who can describe significant past achievements and who operate with an exceptional level of personal responsibility for outcomes. Because our corps members face such tremendous challenges, we seek applicants who have demonstrated determination and persistence when confronted with obstacles in the past. Lastly, we seek people with the specific skills - from critical thinking to organizational ability - that we

have seen characterize our most successful teachers. [retrieved June 27 2005 from <http://www.teachforamerica.org/looking.html>]

TFA does not assume its recruits will take up teaching for their entire careers, but does ask for a two-year commitment and most recruits apparently honor that commitment.

Once admitted into the program, TFA enrolls candidates in a 5-week summer institute that requires them to teach summer school students while concurrently taking courses in such topics as classroom management, learning theory, student diversity and so forth. To test the benefits of this program, researchers usually contrast TFA recruits with other teachers.

There is considerable muddle about what one is learning from such a comparison. Some researchers think that this comparison informs them about the value of *pedagogical knowledge*, assuming that the comparison group of traditionally-certified teachers has taken courses in pedagogy that TFA recruits have not taken. But these groups may not differ in their pedagogical knowledge. TFA teachers are not completely lacking in pedagogical knowledge because TFA does provide some of this content to its teachers. In addition, TFA candidates must also meet state requirements for certification and many states require that teachers who enter the field without a traditional certificate must work toward obtaining one during their first few years of practice. In these cases, differences in pedagogical knowledge between TFA recruits may be small at the outset and may disappear completely as TFA recruits take the courses required for traditional certificate.

There is another reason why these comparisons don't inform us about pedagogical knowledge: in the studies we review here, most teachers in the "other" category also frequently lack a full traditional certificate. The non-TFA teachers in these studies represent an eclectic mix of emergency certified, alternatively certified and traditionally certified teachers. For all these reasons, then, contrasts of TFA teachers with other teachers are not very informative about the value of pedagogical knowledge. We cannot be sure how much pedagogical knowledge either group in the comparison actually has.

But the contrast does inform us about the value of different *recruitment strategies*, since TFA focuses its recruitment on more selective institutions. So an examination of TFA graduates offers us a good opportunity to test the bright, well-educated person hypothesis in situations

where pedagogical knowledge may not be substantially different between these groups of teachers.

We found three studies that meet our evidence standards and that examine the influence of TFA recruits on student math achievement scores⁶. In each of these studies, the number of TFA teachers available for study was relatively small. The first two studies were both conducted in Houston, Texas, using data that belonged to the school district. These two studies both used multiple cohorts of teachers and students, and were not able to control students classroom assignments. They both used multiple regression to try to statistically control for other differences among students.

The first Houston study was conducted by Raymond and Fletcher (Raymond & Fletcher, 2002a, 2002b; Raymond, Fletcher, & Luque, 2001). These authors used 1996-2000 data from upper elementary and middle school grade levels. During this period, Houston hired between 350 and 420 new teachers each year, but only about 20-25 per year, or less than 10%, were TFA teachers. The researchers note that TFA teachers were placed in schools with higher percentages of Latinos and with higher percentages of students receiving free or reduced-price lunches. Test scores were also lower in the schools where TFA teachers taught.

The second Houston study (Darling-Hammond, Holtzman, Gatlin, & Hellig, 2005a) was explicitly designed to replicate and extend the first by paying more attention to the specific credentials of the comparison group. These authors found over 100 different certification categories in Houston's data base and, after studying the state's code, reduced these to seven categories: Standard, Alternative, emergency/temporary, certified out of field, certified no test, uncertified, and unknown.

The third study of TFA recruits was an experiment in which students within each grade level and school were randomly assigned to either an TFA teacher or another teacher (Decker, Mayer, & Glazerman, 2004). Because experiments entail random assignments of students to programs (in this case, to teachers with different kinds of credentials) they remove the confounding caused by affinity assignments that tends to occur in natural settings and they increase the chances of seeing an unbiased estimate of the effect of the program of interest.

Because the Decker et al. study is an experiment, it is stronger than the other two, but there are certain problems shared by all three studies. In all cases, the number of TFA teachers is

rather small, so that researchers had to combine grade levels to obtain reasonable group sizes. In all cases, the non-TFA teachers represented an eclectic mix of credentials. For example, in their comparison group of novice teachers, Decker et al found that 31% had a full traditional certificate, 28% had a temporary certificate, and another 25% had emergency certificates. Still, the contrast remains a reasonable test of the bright, well-educated person hypothesis, since the main distinction between these two groups lies in TFAs recruitment strategies. Less than 4% of the novices in the Decker et al's "all other" group graduated from a competitive college or university, whereas 70% of TFA teachers graduated from such schools.

Figure 7 shows the effects of TFA recruits relative to other teachers in each of these three studies⁷. Each hatch mark represents a particular estimate of the influence of TFA recruits on their pupils' mathematics achievement. They are all positive, ranging from 3% of annual growth to 38%. These findings are encouraging for the bright, well-educated person hypothesis, in part because we are more confident that the normal processes of differential migration and affinity matching have been disrupted—in one case by an experiment and in the other two by the TFA recruitment and placement strategy. So these findings are more persuasive than those about institutional status that were displayed in Figures 5 and 6. However, the most persuasive findings, those from Decker et al, also yielded the smallest effect.

Insert Figure 7 here

Summary. Evidence regarding the bright, well-educated person hypothesis comes both from studies of the institutional status of teachers' alma maters and from studies of TFA, which can be considered as more of a recruitment strategy than a different approach to curriculum. Studies in the first group are more susceptible to biases associated with affinity assignments, but those in the second group correct for these biases by systematically drawing teachers from elite institutions and placing them in schools with less qualified students. The second group, however, is limited to very small samples of bright, well-educated teachers compared with large, amorphous samples of teachers with highly variable backgrounds. If we can assume that more selective institutions tend to produce brighter, more well-educated people, then both sets of studies provide support for the hypothesis that bright, well-educated people do help

students learn more. The strongest study in our set suggests that bright well-educated people add about 3 or 4 percent to students' average annual achievement gains.

Discussion

Since teachers are in the business of educating students, we want to believe that their own education is relevant to their effectiveness. The empirical evidence suggests that this is the case, though none of it shows strong and unambiguous patterns of influence. Many factors are likely to reduce the apparent relationship, making it difficult to use data such as these to make curriculum policy decisions. For instance, most studies rely on a generic test of student achievement that does not necessarily measure the content teachers are actually teaching. If, say, half the content is actually taught and the other half is not, we would not expect to see any changes in the second half of the test. We try to correct for this problem by looking at effects relative to the average amount students in the sample gained *on this test*. Often, these gains were only 1 or 2 points. Another problem plaguing these studies is that teachers generally choose their own curriculum. The number of courses they take in a particular domain tells us how much they may have learned, but it also tells us how interested they are in that domain. All estimates of the influence of courses and majors are confounded with the influence of personalities, values and interests. Keeping this in mind, we assume that any apparently negative "effect" is not really an effect of the courses per se, but instead an effect of differences in personalities and interests among those who take different configurations of courses. A third problem plaguing research in this area is that teachers and students in schools are often matched through processes of differential migration and affinity-matching assignments. Even though the studies we examine have tried to take account of many other influences on students' achievement, we cannot know that they have included the right set of factors in their models, so virtually all of these findings, with the possible exception of the experiment, may be distorted by the influence of other, unmeasured influences on students.

These caveats aside, we suggest that individual courses do indeed add to teachers effectiveness, and that this is true for all three domains of knowledge--mathematics, education, and mathematics education. However, the effects we see when we examine intensive study are

more various, ranging from large negative effects to large positive effects. We identify two possible reasons for this phenomenon. First, it is possible that the relationship between the number of courses teachers take in a domain and their later teaching effectiveness may be curvilinear, as Monk suggested. If this is the case, the first courses teachers take would have a relatively larger impact on their future effectiveness, but more advanced courses would not add benefit. Since all the studies in our sample other than Monk's assume a linear relationship, such that each course adds an equal increment to student learning, their estimates may be skewed either positively or negatively. The second explanation for this more varied pattern is that the larger apparent effects reflect the influence of differences in personalities or interests between those who choose intensive study in a subject and those who do not.

Finally, the evidence we have reviewed suggests that students learn more from teachers who graduate from more selective institutions. We base this conclusion mainly on TFA studies, since the two regression studies yielded inconsistent findings. Still, the effects of TFA's selectivity recruitment are not substantial, especially when compared with the effects of individual courses, which are measured without regard for institutional quality.

It is also likely that teachers' educational backgrounds make larger contributions to student achievement than any of these studies would suggest. The education system is extremely "noisy," with numerous events and circumstances interfering with any relationship we seek to identify. For example, there is tremendous turnover in the teaching population, and substantial migration as well. There are also vast differences among schools that can influence teachers' ability to be effective. And, of course, time is always passing, so that the distance between a teachers' education and her current practice is continually expanding. Throughout this time interval, teachers are influenced by other things, and are learning from other sources. This number and variety of potential intervening variables makes it difficult to ferret out the impact of teachers' educational backgrounds, even when these backgrounds matter. In light of these complications, the findings summarized here should be encouraging for those who provide these courses to prospective teachers.

We suggested at the outset that debates about the education of teachers tend to focus more on the merits of courses in education specifically than on the merits of courses in other domains. Advocates for teacher education are persuaded that they have meritorious and important

content that teachers need to know. Skeptics suggest that, on the contrary, these courses are rarely beneficial and that teachers would be better educated if the space currently reserved for education courses were released. Skeptics also fear that requirements for such courses may actually discourage bright well-educated people from entering teaching.

Our evidence provides little ammunition for either side. It does not suggest that education courses are harmful; indeed it suggests that they are beneficial, but that there may be a limit to their value. In fact, our evidence also suggests that intensive study in mathematics may not be beneficial either. Evidence regarding the benefits of intensive study even at the undergraduate level within mathematics, is mixed. Finally, our evidence suggests that teachers who are bright and well-educated tend to be more effective than their more ordinary counterparts. However the they bring to students are not substantial, relative to the benefits of individual courses taken by prospective teachers in less prestigious institutions.

The debate over ways to optimize the college education we provide for future teachers has intensified in recent years as evidence reveals how important teachers can be in fostering, or failing to foster, growth in student achievement. Participants on all side of this debate are anxious to improve the quality of education this nation provides its students. Our evidence suggests that additional teacher courses in content, content pedagogy, and pedagogy, all benefit students. But this evidence also raises questions about the limits of knowledge within each domain. One study suggests that the benefits of additional courses in mathematics tapers off even before students complete a major in mathematics. Another suggests that the benefits of additional courses in education taper off at the conclusion of the bachelors degree. If improvements are needed in the curriculum we offer to prospective teachers, perhaps we should reconsider the need for intensive study in any domain, rather than debating the relative merits of the domains as a whole.

Table 1: Best-evidence studies examining teachers' college education in relation to student achievement in mathematics

<i>Citation</i>	<i>Hypothesis</i>	<i>Educational Background of interest</i>		<i>grade Level</i>	<i>Sample</i>		
		<i>Metric</i>	<i>Content area</i>		<i>N of Teachers</i>	<i>N of statistical controls</i>	<i>N of estimates of influence</i>
Aaronson, Barrow, & Sander (2003)	CK	Major in:	math;	9	52991	35	2
	PK		education				
	BWE	Level of:	institutional status	9	52991	35	6
Brewer and Goldhaber (1996)	CK	Major in:	math;	10, 12	5113, 3768	41	3
	PK		education			37	
Brewer, (1996a, 1996b; 1997; 1999; 2000)	SMK	MA in:	math;	10, 12	5113, 3768	41	4
			education			37	
Chiang (1996); Rowan, Chiang and Miller (1997)	CK	Double degree in:	math	12	4751	16	1
		Any degree in:	math	10	5381	23	1
Clotfelter, Ladd, & Vigdor (2004)	BWE	Level of:	Institutional status	5	60656	25	6
Darling-Hammond, Holtzman, Gatlin, & Hellig (2005a; 2005b)	BWE	program entry:	TFA	4&5	Range from 10488 to 105511	10	3
Decker, Mayer, & Glazerman (2004)	BWE	program entry:	TFA	1-5	1715	6	2
Eberts & Stone (1984)*	CK PK	N of courses in:	math teaching	4	Over 14,000	8	1
Hill, Rowan, & Ball (2005)	CK+ PCK	N of courses in:	math OR math education	1, 3	1190, 1773	17	2
Fagnano (1988)	CK PK	N of courses in:	math; math education; education	8	211	~100	16
Monk (1994); Monk & King (1994)	CK	N of courses in:	math;	10, 11	1492, 983	10, 11	4
	PCK		math education,				
	CK	More than 5 Math courses in:		11	983	11	1
Raymond & Fletcher, (2002a; 2002b); Raymond et al. (2001)	BWE	Program entry:	TFA	4-7	Range from 11321 to 96279	9	4
Rowan, Correnti, & Miller (2002)**	CK	BA/MA in:	Math	1-6	NI	NI	2
Rowley (2004)	PK PCK	N of courses in:	elementary ed; early ed; child dev; math teaching	K	2408	31	4
Taddese (1997)	CK	Major in:	Math	12	20840	22	1

* This is the only study we used whose original outcome metric was measured via gain scores

** Not included in Figure 4 because we could not convert the data into a comparable metric.

References

- *Aaronson, D., Barrow, L., & Sander, W. (2003). Teachers and student achievement in the Chicago public high schools. Chicago: Federal Reserve Bank of Chicago.
- Ballou, D., & Podgursky, M. (1999). Teacher training and licensure: A layman's guide. In M. Kanstoroom & C. Finn (Eds.), *Better teachers, better schools*. (pp. 31-82). Washington DC: Thomas B. Fordham Institute.
- Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2003). The draw of home: How teachers' preferences for proximity disadvantage urban schools. Cambridge MA: National Bureau of Economic Research.
- Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2002). Analyzing the determinants of the matching of public school teachers to jobs. Albany: SUNY Albany.
- *Brewer, D. J. and D. D. Goldhaber. (1996). Educational achievement and teacher qualifications: New evidence from microlevel data. In B. S. Cooper and S. T. Speakman (Eds.), *Optimizing educational resources* (pp. 243-264). Greenwich, CT: JAI.
- *Cavalluzzo, L. (2004). Is National Board Certification An Effective Signal of Teacher Quality? Washington DC: The CNA Corporation.
- *Chiang, F.-S. (1996). Ability, motivation, and performance: A quantitative study of teacher effects on student mathematics achievement using NELS:88 data. Unpublished Dissertation, University of Michigan.
- *Clotfelter, C., Ladd, H., & Vigdor, J. (2004). Teacher quality and minority achievement gaps. Durham NC: Duke University Sanford Institute of Public Policy.
- Council of Chief State School Officers. (1988). *State Education Indicators, 1988*. Washington, D.C.: Author.
- Conant, J. B. (1963). *The Education of American Teachers*. New York: McGraw Hill.
- Cunningham, G. K., & Stone, J. E. (2005). Value-added assessment of teacher quality as an alternative to the national board for professional teaching standards: What recent studies say. *Education Consumers Clearinghouse*.
- Damarell, R. G. (year). *Education's Smoking Gun: How teachers colleges have destroyed education in America*: publisher.
- *Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Hellig, J. V. (2005). Does teacher certification matter? Evidence about teacher certification, Teach for America, and teacher effectiveness.

Stanford University. Retrieved April 16, 2005, from the World Wide Web:
<http://www.schoolredesign.net/sm/server.php?idx=934>

- *Darling-Hammond, L., D. J. Holtzman, S. J. Gatlin and J. V. Heilig (2005b). Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness. *Education Policy Analysis Archives* **13** (42).
- *Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). The Effects of Teach For America on Students: Findings from a National Evaluation. Princeton NJ: Mathematica Policy Research.
- *Eberts, R. W., & Stone, J. A. (1984). Unions and public schools. Lexington, MA: Lexington Books.
- *Fagnano, C. L. (1988). *An investigation into the effects of specific types of teacher training on eighth grade mathematics students' mathematics achievement*. Unpublished Dissertation, University of California at Los Angeles.
- *Goldhaber, D. D. and D. J. Brewer. (1997). Evaluating the effect of teacher degree level on educational performance. In W. Fowler (Ed.), *Developments in School Finance, 1996* (pp. 197-210). Washington DC: US Department of Education National Center for Education Statistics.
- *Goldhaber, D. D. and D. J. Brewer. (1999). Teacher licensing and student achievement. In M. Kanstoroom and C. F. J. Finn (Eds.), *Better teachers, better schools* (pp. 83-102). Washington, DC: Thomas B. Fordham Foundation.
- *Goldhaber, D. D. and D. J. Brewer (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis* **22** (2): 129-45.
- Goldhaber, D. D., & Anthony, E. (2003). *Teacher quality and student achievement*. New York: ERIC clearinghouse on urban education.
- Hanushek, E. (1971). Teacher characteristics and gains in student achievement: estimation using micro data. *American Economic Review*, *61*(2), 280-288.
- Hedges, L. V. (1986). Issues in meta-analysis, *Review of Research in Education* (Vol. 13, pp. 353-398). Washington DC: American Educational Research Association.
- Hess, F. (2001). *Tear down this wall: The case for a radical overhaul of teacher certification*: Progressive Policy Institute.
- *Hill, H., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371-406.

- Kramer, R. (1991). *Ed School Follies: The Miseducation of America's Teachers*. New York: The Free Press.
- Kupermintz, H. (2002). Value-added assessment of teachers: The empirical evidence. In A. Molnar (Ed.), *School reform proposals: The research evidence*.
- Labaree, D. F. (2004). *The trouble with Ed Schools*. New Haven CT: Yale University Press.
- Laczko-Kerr, I. I. (2002). The effects of teacher certification on student achievement: An analysis of Stanford Nine achievement for students with emergency and standard certified teachers. Paper presented at the American Educational Research Association, New Orleans, LA.
- Laczko-Kerr, I. I. (2002). *Teacher certification does matter: The effects of certification status on student achievement*. Unpublished Dissertation, Arizona State University.
- Laczko, I. I., & Berliner, D. C. (2001). *The effects of teacher certification on student achievement: An analysis of the Stanford Nine*. Paper presented at the American Educational Research Association, Seattle, WA.
- Laczko-err, I. I., & Berliner, D. C. (2002). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case of harmful public policy. *Education Policy Analysis Archives*, 10(37).
- Lagemann, E. C. (1999). Whither Schools of Education? Whither Educational Research. *Journal of Teacher Education*, 50(5), 373-376.
- Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Education Evaluation and Policy Analysis*, 24(1), 37-62.
- Ludwig, J., & Bassie, L. J. (1999). The puzzling case of school resources and student achievement. *Educational Evaluation and Policy Analysis*, 21(4), 385-403.
- Metzger, S. A., Qu, Y., & Becker, B. J. (2004). An Examination of Literature on Teacher Qualifications: Influence of Ill-defined Constructs on Synthesis Outcomes. Paper presented at the American Educational Research Association, San Diego.
- *Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review*, 13(2), 125-145.
- *Monk, D. H., & King, J. A. (1994). Multilevel teacher resource effects on pupil performance in secondary mathematics and science: The case for teacher subject-matter preparation. In R. G. Ehrenberg (Ed.), *Choices and consequences: Contemporary policy issues in education* (pp. 29-58). Ithaca, NY: ILR Press.

- Murnane, R. J. (1981). Interpreting the evidence on school effectiveness. *Teachers College Record*, 83(1), 19-35.
- National Commission on Teaching and America's Future. (1996). *What Matters Most: Teaching for America's Future*. New York: National Commission on Teaching & America's Future.
- National Center for Education Statistics. (2003). Digest of Educational Statistics. author. Retrieved June 20, 2005, from the World Wide Web: <http://nces.ed.gov/programs/digest/d03/tables/dt067.asp>
- Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. K. Rothkopf (Ed.), *Review of Research in Education* (Vol. 15, pp. 423-475). Washington DC: American Educational Research Association.
- *Raymond, M., & Fletcher, S. (2002a). Education Next Summary of CREDO's Evaluation of Teach for America. Education Next. Retrieved, from the World Wide Web:
- *Raymond, M., & Fletcher, S. (2002b). The Teach for America Evaluation. *Education Next*(Spring), 62-68.
- *Raymond, M., Fletcher, S. H., & Luque, J. (2001). Teach for America: An evaluation of teacher differences and student outcomes in Houston, Texas. Houston TX: CREDO.
- Rokoff, J. E. (2003). The impact of individual teachers on student achievement: Evidence from panel data.
- Rotherman, A., & Mead, S. (2003, October 23-24). Back to the Future: The history and politics of state teacher licensure and certification. Paper presented at the A qualified teacher in every classroom: Appraising old answers and new ideas, Washington DC.
- *Rowan, B., Chiang, F.-S., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, 70(October), 256-284.
- *Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of elementary schools. *Teachers College Record*, 104(8), 1525-1567.
- *Rowley, K. J. (2004, April). Teacher experience, certification, & education: Characteristics that matter most in kindergarten math achievement. Paper presented at the American Educational Research Association, San Diego, CA.
- Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) data base: Implications for Educational Evaluation and Research. *Journal of Personnel Evaluation in Education*, 12(3), 247-256.

- Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System: Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8(1), 299-311.
- Slavin, R. E. (1984). Meta-Analysis in Education: How has it been used. *Educational Researcher*(October), 6-15.
- Slavin, R. E. (1986). Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educational Researcher*, 15(9), 5-11.
- Strauss, R. P. (1999). Who gets hired? The case of Pennsylvania. In M. Kanstoroom & C. E. Finn, Jr. (Eds.), *Better Teachers, Better Schools* (pp. 103-130). Washington DC: The Thomas B Fordham Foundation.
- *Taddese, N. (1997). The impact of teacher, family and student attributes on mathematics achievement. Unpublished Dissertation, University of Cincinnati.
- Wykoff, J. (2001). *The Geography of Teacher Labor Markets: Implications for Policy*. Albany: SUNY.
- * Included in Table 1.

EndNotes

1. Work for this paper was supported by a grant from the U. S. Department of Education, Institute for Educational Sciences, and by the National Science Foundation, Program on Research, Evaluation and Communication. Responsibility for the content and quality of the paper reside solely with the authors.
2. Work for this paper was supported by a grant from the U. S. Department of Education, Institute for Educational Sciences, and by the National Science Foundation, Program on Research, Evaluation and Communication. Responsibility for the content and quality of the paper reside solely with the authors.
3. The term “value added” has two meanings in the literature. It sometimes refers to a very specific analytic approach (Kupermintz, 2002; Sanders & Horn, 1994) and sometimes to the general approach of using pretests to control for prior achievement in a statistical model (Cunningham & Stone, 2005). We use it here in the more general sense.
4. There were a few cases in which authors did not provide this information. These all involved public data bases and we were able to obtain the information from other sources.
5. Since these authors did not provide information on average students’ change during the year, we present their findings using raw scores. We are unable to say whether these differences are substantial relative to average gains or not.
4. Another recent study, by Laczko-Kerr (2002a; 2002b; Laczko and Berliner, 2001; Laczko-Kerr and Berliner, 2002) was rejected because it did not employ a pretest.
7. The Darling-Hammond et al study actually includes effects as measured by several different tests, and the authors argue that, in the Houston situation, the other tests may be better indicators of student learning because there have been some instances of cheating on the TAAS. However, the TAAS is the assessment that is aligned with the state curriculum and that teachers are accountable for, so we focus on that test. Since, in all cases, we take only one model per study sample, to avoid data dependency among effects, this is the only estimate we take from that study.

Figure 1:
Effects of additional coursework in different content areas
(From Monk, 1994)

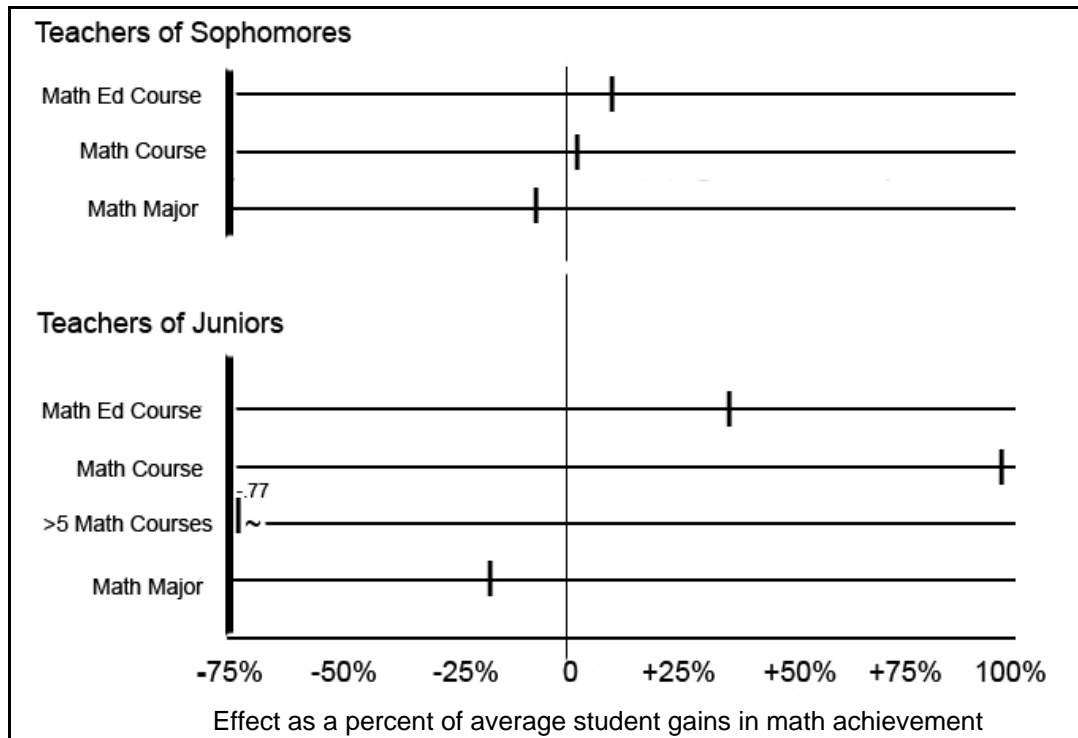


Figure 2:
Effects of additional courses in different content areas
(From Rowley, 2005)

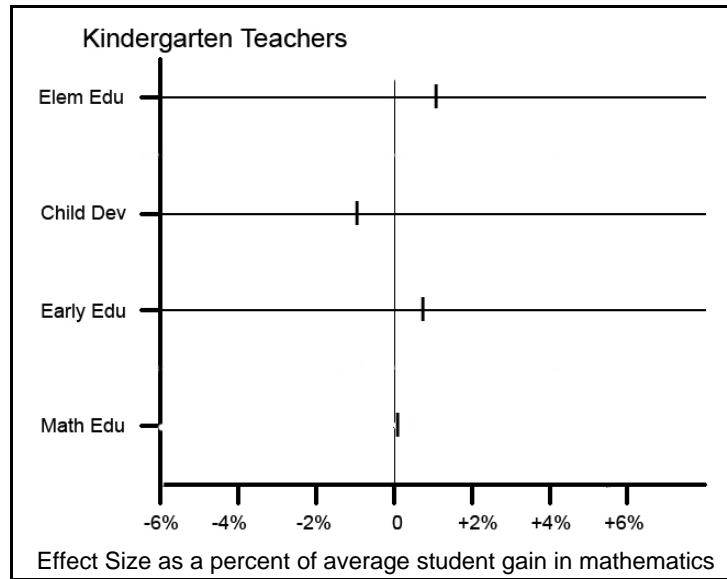
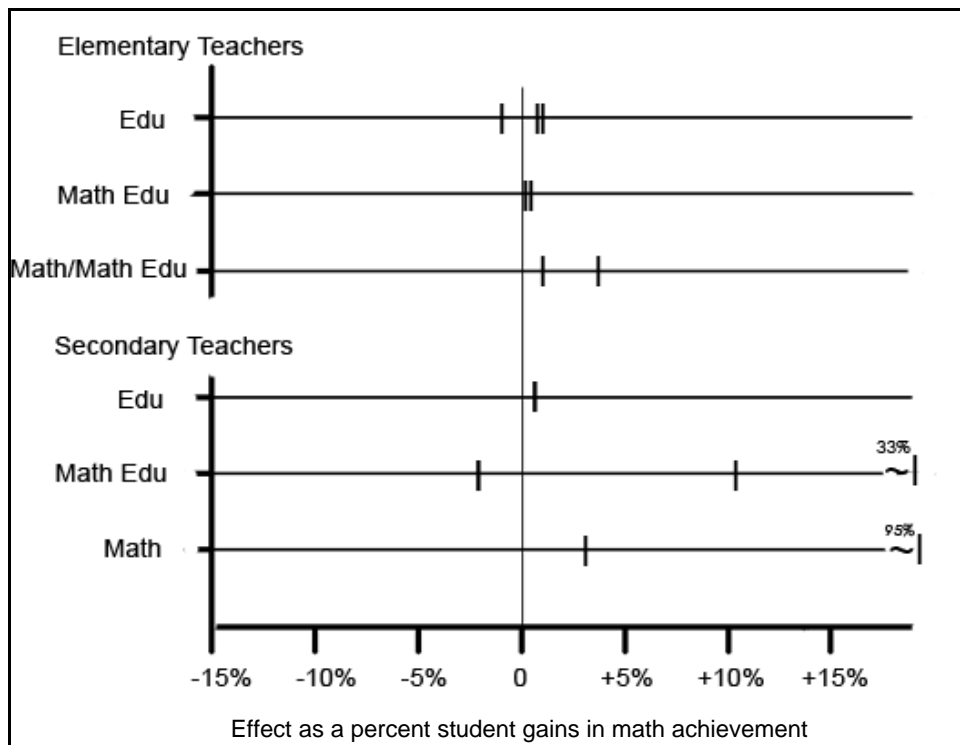


Figure 3
All estimates of effects of additional courses



Note: We count grades K-6 as elementary grads and 7-12 as secondary.

Figure 4
All estimates of the effect of intensive study in one content area

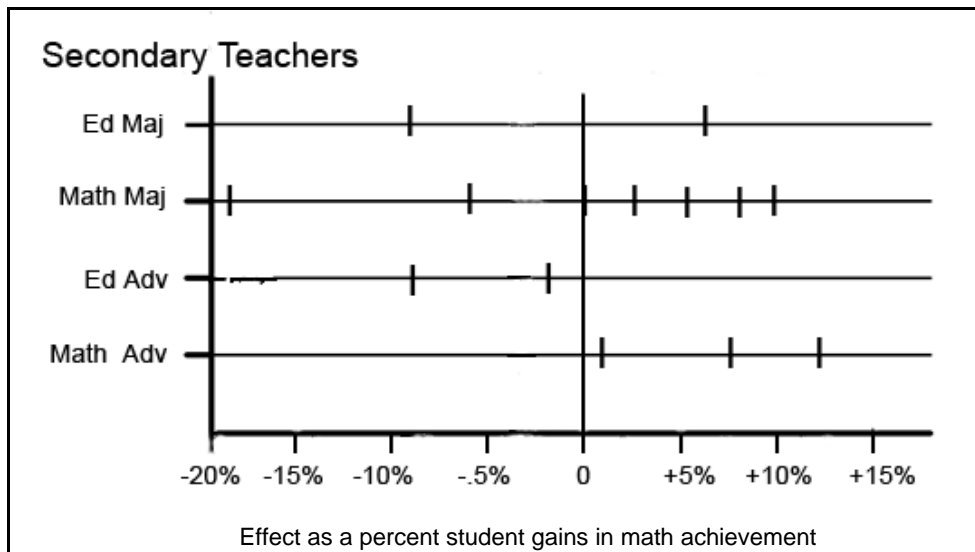


Figure 5
Effects of Institutional Status on Students' Achievement
(From Clotfelter, 2004)

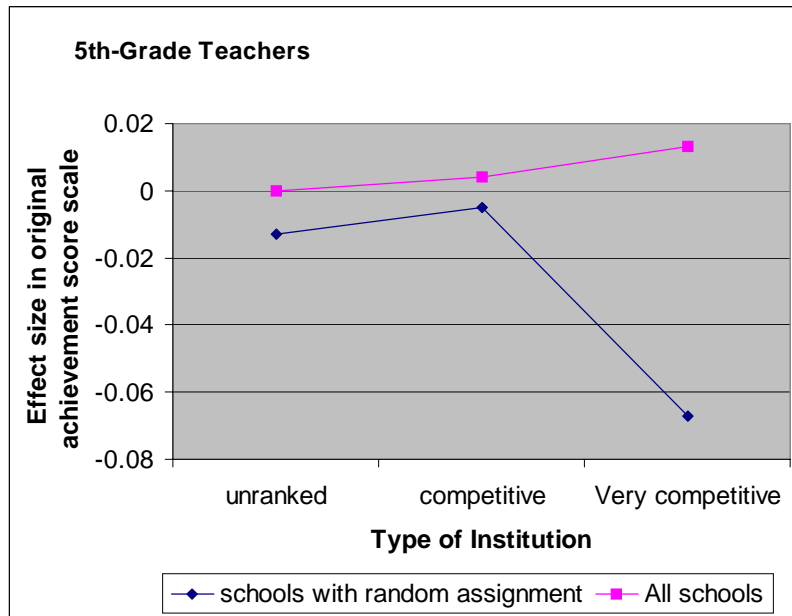


Figure 6
Effects of Teachers' Alma Mater's Institutional Status
on Student's mathematics Achievement
(From Aaronson, 2003)

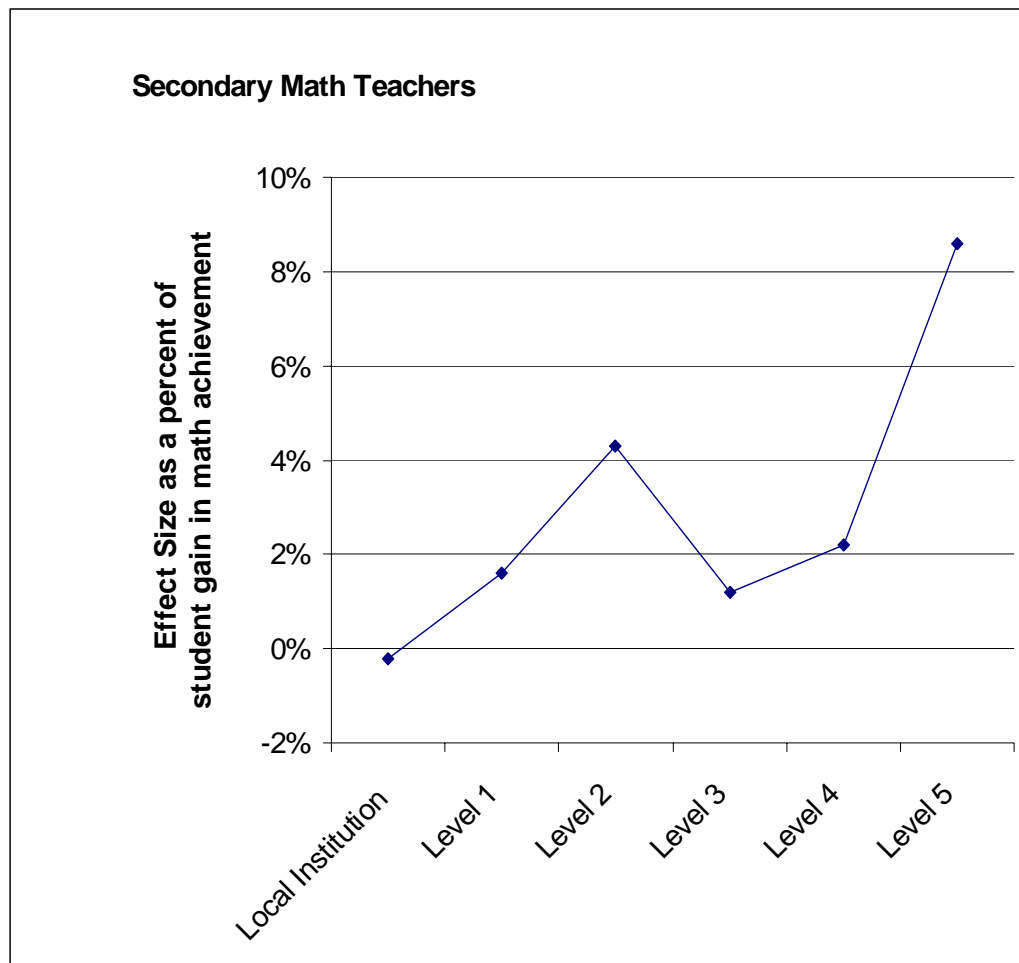
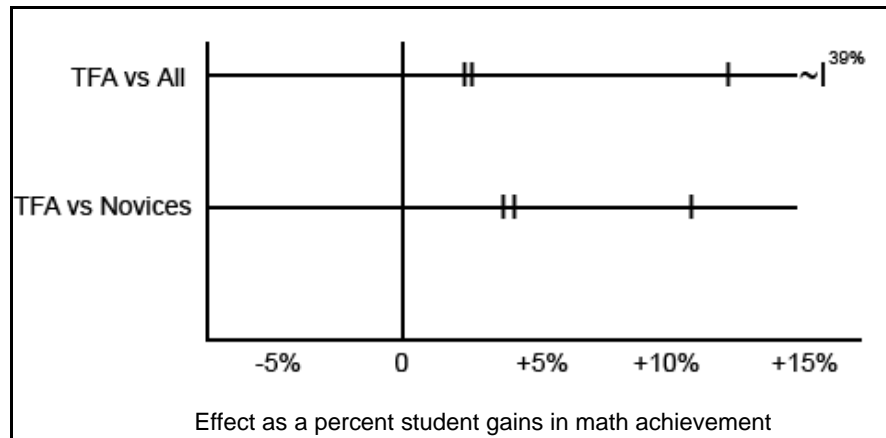


Figure 7
All TFA group comparison data



Each hatch mark reflects one estimate of change in student test scores associated with TFA teachers relative to other teachers. Note that estimates are based on original test score metric and so are not necessarily comparable.