

---

## 6 The Problem of Evidence in Teacher Education

---

*Mary M. Kennedy*<sup>1</sup>

The current wave of skepticism about the effects of teacher education is not new. Almost since its inception, the need for, and value of, teacher education has been doubted by non-teacher educators. Some of these doubts stem from beliefs that teachers are born rather than made, some from beliefs that the practice of teaching is not particularly difficult, some from beliefs that teaching can only be learned in the doing. For their part, teacher educators have conducted research on the effectiveness of their efforts, in part to learn how to improve their practices, but also in part to defend themselves from external skeptics.

Several years ago I reviewed these approaches to research on teacher education with an eye toward their potential to help teacher educators understand their practices (Kennedy, 1996). My aim in this chapter is to examine the same approaches, this time from the skeptic's point of view.

Five broad strategies, or genres, for research in teacher education have been and are still in widespread use. They are summarized in Table 6.1. Each offers us a glimpse of one aspect of teacher education and one aspect of its influence on teachers, but none is broad enough to capture the entire story. Indeed, if a study were sufficient to capture the entire story, it would be so hopelessly difficult, time-consuming, and expensive to complete that it will probably never exist. As I review each genre, I ask three questions of it. First, what aspects of teacher education does it look at? Second, what outcomes does it look at? And third, on what basis does the research argue that there is a causal relationship between these aspects of programs and these outcomes?

### **Multiple Regression**

The first genre, multiple regression, is most familiar to skeptics of teacher education and least familiar to teacher educators themselves. Researchers working within this genre are not testing any particular theory about teacher education. Instead, they are engaged in a relatively open-ended search for contributions to pupil learning, and one of the possible contributions is teacher education.

These studies were originally stimulated by, and are based on, the *Equality of Educational Opportunities* (EEOS) Study (Coleman, 1966), and many actually used the EEOS data. One of the earliest and best of these studies was conducted by Eric Hanushek (1971, 1972). He began by asking whether teachers differed in their

Table 6.1 *Five approaches to research on the contribution of teacher education to teaching*

Genre	Argument	Aspect of TE examined	Outcome examined
Multiple Regression	If teachers who have taken more credits in teacher education foster greater gains in student achievement than teachers with fewer credits, then teacher education has made a difference	Number of teacher-education courses or degrees held	Student achievement scores
Follow-up Surveys	If teacher-education graduates claim that their teacher education courses were valuable, then they were	Specific courses	Alumni perceptions
Comparison of credentialed and non-credentialed teachers	If teachers who have received formal teaching certificates teach differently than other teachers, then teacher education has made a difference	Presence or absence of a credential, types of credentials	Classroom behavior
Experiments	If teachers participating in one program approach improve their skill more than other teachers do, then this approach has a greater impact	Discrete course segments or course procedures	Ability to perform specified behaviors on demand
Longitudinal studies of change	If candidates express one view at the beginning of their program and another view later on, and if they refer to their courses to justify their views, then teacher education has made a difference	Whole programs as experienced by candidates	Verbally expressed attitudes and beliefs

## *The Problem of Evidence in Teacher Education*

ability to increase student achievement, after taking into account the child's initial achievement and various aspects of the child's background. Hanushek found that teachers did make a difference. That is, the teacher a child happened to have could significantly influence the child's achievement growth for the school year. Seeing that this was the case, Hanushek then tried to see which particular teacher characteristics seemed to account for these differences. Among the variables Hanushek examined were the teacher's college major, the number of hours of graduate coursework teachers had taken, and the length of time since the teachers' most recent educational experience. Hanushek found that neither college major nor the number of graduate credits teachers had taken were significantly related to student achievement. Variables that were related, on the other hand, included the teachers' general verbal ability and the recency of their last educational experience. These two variables do not necessarily reflect teacher education courses per se, although verbal ability may reflect the teachers' college education in general. The recency of the teachers' educational experiences may reflect either the nature of the experiences or the teachers' interest in continued professional learning.

Another important study that focused on teacher education was done by Murnane and Phillips (1981). Like Hanushek, they began by testing to see whether teachers made a difference, found that they did, and then tried to see what aspects of teachers seemed to account for these differences. But instead of generating a single equation which included all possible contributions to student achievement, they developed two separate equations, one of which included measures of teacher *behaviors* (e.g. circulating around the room to correct seatwork, use demonstrations, make students repeat poor work, etc.) and the second of which included measures of teacher *characteristics* (e.g. years of experience, possession of a masters degree and prestige of college attended). Their data indicated that teacher behaviors were better predictors of student achievement than were teacher characteristics. Moreover, of those characteristics that Murnane and Phillips examined, neither of their education-related variables—possession of a Masters degree or prestige of college attended—appeared to be relevant to student achievement.

Many skeptics of teacher education reside in the policy community, and they may be more aware of this body of research than of the others that I describe below. If these multiple regressions were the only studies available on the merits of teacher education, it would be easy to conclude that teacher education contributes very little to teaching. But there are some weaknesses in this genre, and these need to be considered before a conclusion is drawn.

### *Aspects of Teacher Education Examined*

The aspects of teacher education that these researchers examine are often relatively crude quantifiable indicators such as whether or not a teacher has a teaching credential, whether he or she has a masters degree, or, occasionally, the number of teacher education courses the teacher took in college. These measures may not be very meaningful, for two reasons.

First, virtually every teacher in these studies *already holds a bachelors degree* and is already certified to teach. The percentage of teachers lacking a bachelors degree was only 7 per cent in 1966, and has since fallen to less than 1 per cent (National Center for Education Statistics, 1989). Presumably, then, virtually all of these teachers have attained the minimum educational background required for teaching. The variations that are measured, therefore, are not variations in the required core of teacher education, but instead variations in what other educational courses teachers have elected to take.

Second, since the United States does not have a centralized curriculum, and since many states give teacher educators considerable leeway in their program designs, teacher education programs can look remarkably different from one institution to the next. One recent study suggested that the number of education credits required by each state ranged from 18 to 90 for elementary teacher candidates (Council of Chief State School Officers, 1988). Not only does the number of education courses taken differ, but their content and character differ as well. These differences reflect different ideas about what teachers need to know and about how teachers learn. By failing to measure the substantive differences among programs, researchers in this genre may miss the very aspect of teacher education that is most likely to make a difference. Moreover, because there is so much variation in the content and character of teacher education programs, measures of the *amount* of teacher education do not measure of a unitary thing. Some teachers may have received extensive education in a mediocre program while others received modest education in a very good program. It should not be surprising, therefore, that these measures generally do not correlate highly with measures of student achievement gains. When a measurement scale does not measure a unitary thing, it is hard to show a relationship between it and other things.

A recent study by Ferguson and Womack (1993) corrects this problem in two ways. First, instead of simply tallying up courses, they document the actual courses students took during their undergraduate preparation and correlated the presence or absence of each course with teaching performance during student teaching. Second, instead of using presence or absence of a disciplinary major as their indicator of subject matter knowledge, they use both grade point average within the subject and score on the subject matter portion of the National Teachers Exam (NTE). With these alterations in method, the researchers found substantial relationships between the teacher education courses that teacher candidates took and the quality of the teaching they demonstrated.

### *Outcomes*

Researchers practicing within this genre take student achievement, or better still, gains in student achievement, as their primary outcome. Since teachers often aim to create benefits to students that extend far beyond those measured by achievement tests, student achievement outcomes measure some, but not all, of the important goals of education. In fact, whether such tests even measure the most important

outcomes of schooling is a contentious issue in the current reform rhetoric. Gains in student achievement, then, constitute a very narrow outcome for estimating the contributions of teacher education to teaching.

### *The Logic of the Argument*

The logic of these studies goes something like this: if teachers who have taken more credits in teacher education foster greater gains in student achievement than teachers with fewer credits (after taking into account differences in entering achievement, family background, and so forth) then these teacher education courses have made a difference. Because such differences have rarely been observed, researchers within this genre tend to be skeptics and often argue against policies that require teachers to take certain numbers of credits, or that pay teachers more if they have masters degrees, for instance (see, e.g. Murnane et al., 1991).

Studies in this genre also depend heavily on the ability of complex statistical techniques to ferret out extraneous variables that can influence student achievement -- variables such as the child's family background, for instance. The success of each study depends on whether the researcher captures all the relevant variables, and studies can differ substantially in their thoroughness. But there are always influences that cannot be measured. Take, for instance, school climate. There is ample evidence that school climate contributes to student achievement (Brophy and Good, 1986), but less is known about how and why school climate matters. But suppose that the reason school climate contributes to student achievement is because it enables teachers to teach better. If that is the case, then the effects of the teachers' educational background might be masked by variations in school climates. It may be that teachers' educational backgrounds make a difference *within each school climate*, but these effects are not apparent when a wide range of school climates are involved in the study. Or, it is possible that teacher education only makes a difference within reasonably positive school climates, and that teacher education cannot help teachers teach better when they are working in especially difficult schools.

These caveats should not be taken as an excuse for teacher education, however. No doubt a skeptic would respond to these arguments by pointing out that teacher education should have a strong enough impact that it can be seen in any school climate. Still, there are enough limitations in this research genre that it cannot be depended upon to provide unambiguous evidence for the merits of teacher education. The aspects of teacher education that it measures often are degrees, quantities of course work, or courses taken after the bachelors degree, rather than quality or content of these courses. The outcomes it measures represent only a narrow slice of the outcomes teacher educators hope to influence, and to the extent that its arguments about the relationship between teacher education and student achievement are based on incomplete lists of possible influences on student achievement lacks credibility.

### Follow-up Surveys

While multiple regression studies are most common outside of the teacher education community, the most common genre within this community is the survey of graduates. This genre is popular in part because it is relatively inexpensive and simple to do, and in part because the accreditation system devised by the National Council for Accreditation of Teacher Education (NCATE) has traditionally required some form of program evaluation. Adams and Craig (1983) surveyed teacher education programs in 1980 and found that 74 per cent claimed to be conducting some sort of follow-up of their graduates.

These surveys generally use one of two strategies to estimate the contributions of teacher education. Either they ask teachers to assess their own knowledge and skills -- that is, to assess their own ability to teach -- or they ask the teachers to assess the merits of particular courses within the program.

Pigge's (1978) survey, though old, is a good example of the genre. Pigge developed a list of 26 competencies taken from the Bowling Green curriculum on which the respondents were to rate themselves. On Pigge's five-point scale, most self-assessments were quite high. The lowest average self-assessment was close to the mid-point of the scale. Few teachers, then, viewed themselves as seriously lacking in any of these teaching competencies. Pigge also asked teachers to estimate how important these various competencies were to their work and to indicate where they learned these competencies. Generally speaking, teachers thought that the competencies *most* necessary to their work were learned on the job, whereas those considered *least* necessary were acquired in their teacher education programs.

In an interesting study by Clark, Smith, Newby and Cook (1985), teachers were observed in their classrooms and then asked where they got the ideas for what they did. The most frequently cited source for a teaching idea was that the teacher generated it him- or herself. Second most prominent was the cooperating teacher with whom the teacher had undergone student teaching. Teacher education faculty were given credit for only 17 per cent of the practices teachers were asked about.

The vast majority of follow-up surveys are intended for use only within a particular institution. Many are circulated only within their institutions, and many others are circulated only through the ERIC (Education Resource Information Center) system. Only a few are published in journals. This means that policy-makers and skeptics are not likely to have seen any of these studies. In fact, teacher educators themselves may not know much about findings from other institutions. However, occasionally, a non-teacher educator conducts a follow up survey, and these surveys tend to have wider audiences. One such survey was conducted by the National Education Association (NEA). NEA surveyed its members and asked them to evaluate the contributions of 14 different sources of knowledge about teaching, one of which was preservice teacher education (Smylie, 1989). The preservice teacher education program was ranked by these teachers 13th out of 14 sources of knowledge. The highest-rated sources of knowledge were direct experience, consultation with other teachers, and independent study and observations of other teachers, all of which are entirely in the control of the teacher him or herself. The only source of knowledge

rated less useful than undergraduate teacher education was school district sponsored inservice programs. To the extent that policy-makers and skeptics have any exposure to this genre of research, their exposure will be to studies like this one rather than to studies conducted by individual institutions. But this study, and others like it, does not help the cause of teacher education.

### *Aspect of Teacher Education Examined*

Most follow-up surveys ask graduates to assess particular program offerings -- the math methods course, the student teaching experience, or the placement service. These program features are of less interest to the field as a whole than they are to the particular institutions which provide them. In fact, the more useful a survey becomes to its own institution, the less useful it will likely become to others, for local utility depends on forming questions that are highly specific to the local situation. The genre was never intended for use beyond the particular institution.

### *Outcomes*

Almost universally, follow-up surveys depend on teachers' judgments of their own knowledge or skill, or of the value of the courses they took, as their indicator of program impact. Most surveys provide the teacher with a list of knowledge or competencies or a list of program courses, and ask the teachers to rate themselves or their alma mater on something like a five-point scale.

Reliance on teacher judgment is a substantial limitation in these studies, for several reasons. First, we don't know what *criteria* teachers use when they make these assessments, or whether their criteria would be the same as an independent observer's criteria. Koziol and Burns (1986) found that teachers' self-assessments could be developed to agree more with assessments of independent observers by focusing the assessment form on specific situations (e.g. first-period social studies or the teaching of a particular body of content), rather than on teaching in general, and by using the same set of questions on repeated occasions. In this way, teachers had a chance to learn to attend to those practices that the researchers were interested in. These conditions, however, are never met in follow-up surveys.

Teacher self-assessments may also be influenced by emotional responses to the initial difficulties of developing a solid teaching practice. Gaede (1978) found that teachers' self-assessments gradually increased as teachers moved through their teacher preparation programs, but decreased substantially during their first year of teaching. Certainly these teachers did not suddenly know less than they had known when they were seniors, but just as certainly, they *felt* they knew less once they encountered the demands of real teaching.

Similarly, when a teacher claims a program has contributed to her knowledge or skill, or has not contributed to her knowledge or skill, we don't know how *accurate* these judgments are. It is highly likely that teachers do not recall what

they knew or were able to do five years earlier. Strang, Badt, and Kauffman (1997) provide some evidence that teachers cannot accurately recall their prior capability. They measured teachers' skills both before and after a program treatment, and they also asked teachers afterward to estimate the degree to which they had changed during the program. The researchers' independent assessment of teacher change showed their proficiency moving from 52 per cent to 87 per cent on their performance scale. However, the *teachers'* assessments of their own change indicated movement from 81 per cent to 85 per cent. Teachers, therefore, may not be good judges of whether they have learned from a program or not.

Generally speaking then, outcome measures employed by follow-up surveys are weakened by their heavy reliance on teachers' judgments of themselves, of their own growth, and of what their programs might have contributed to their growth.

### *The Logic of the Argument*

The logic of follow up surveys goes something like this: If teachers who choose to respond to the survey (and this is a big if, for response rates on these surveys are often very low) claim they are competent in certain areas, or if they claim they have (or have not) learned something valuable from their teacher education programs, they are probably correct. Since follow-up surveys rarely directly measure teachers' knowledge or skills, the burden of the argument falls entirely on the teachers' judgments.

These studies are also weak in that, like the multiple regressions, they usually fail to take into account the teaching context. Some teaching situations are far more challenging than others, some provide less assistance to new teachers than others, and some demand different kinds of practices than their programs prepared them for. To the extent that any of these contextual differences might influence teacher judgments, the findings are even more difficult to interpret.

Thus, to make any sense of these data, we have to assume that (a) teachers use the same criteria to judge themselves and their programs as teacher educators, policy-makers, or educational researchers would use; (h) teachers' assessments of their own knowledge and skills, and of their prior knowledge and skills, are not influenced by their emotional states; and (c) the context in which teachers are teaching has no bearing on their assessments of themselves or their teacher education programs. I argued before that these studies are of little use to policy-makers, but I also believe they are also of little use to teacher educators. In fact, the negative evaluations teachers provided in the NEA study could, when combined with multiple regression studies, enhance the skeptics' case.

### **Comparing Credentialed and Non-credentialed Teachers**

Another popular genre, used by both teacher educators and by their critics, consists of comparing teachers with different educational backgrounds. Usually, these researchers focus on a particular school district or geographic region, find all the

teachers who have provisional or emergency credentials, and compare them with teachers in the same region who are fully certified. A recent variation on this theme is to compare teachers who received their certifications through alternative routes with those who received traditional certifications. Once two groups of teachers have been identified, the researchers observe the classroom practices of both groups in search of differences. These studies offer several advantages over follow-up surveys: they rely on an independent observer to assess the teachers' practice, they compare two groups who have different kinds of educational backgrounds, and they often control for the influence of context because they sample within limited geographic regions.

The frequency of these studies waxes and wanes with the availability of non-certified teachers. Many such studies were done in the early 1960s, when school districts experienced serious personnel shortages and consequently hired a lot of provisionally certified teachers (e.g. Beery, 1960; Hall, 1964; Gray, 1962; Bledsoe et al., 1967). The recent introduction of alternative routes has spawned a new series of studies that compare traditionally certified teachers and teachers certified via alternative routes (e.g. Cornett, 1984; Brown et al., 1989; Peck, 1989).

One early study in this genre which has often been cited as evidence *against* the value of teacher education, was conducted by Popham (1971). Popham's intent was to develop a performance assessment of teaching. He devised several teaching units and then asking both certified teachers and college students to teach these units. Each unit entailed 9 hours of teaching, and pupils were randomly assigned to teachers. Teachers were not told how to teach the content, but instead were given only the instructional objectives and some materials. (This study differs from most studies in this genre by using student achievement as the outcome rather than teaching practices per se.) Popham found no significant differences between students taught by certified teachers and those taught by college students, and concluded that teacher education had not prepared teachers in a way that made them distinctly different from ordinary, inexperienced, college students.

Dewalt and Ball's (1987) recent study is more typical of the genre, in that it consisted of observing teachers in their regular classrooms. Observers checked an observation form each time they saw a particular behavior, and the behaviors they watched for were taken from research literature on effective teaching strategies. In addition, the observers specifically asked their teachers to demonstrate these competencies. Thus, their observations do not reflect what teachers might normally do in their classrooms, but instead reflected their ability to do these specific things on demand. One group of teachers had taken no teacher education courses, the other had taken at least 12 credit hours in teacher education but had not done student teaching. So the comparison really asks whether taking at least 12 credits in teacher education makes a difference. The two groups were found to differ on several variables, but the differences did not always favor the same group. Behaviors that were more often demonstrated by teachers who had taken at least 12 credits of teacher education were those having to do with creating a non-punitive classroom climate and accommodating individual differences. Those that favored teachers who had taken no courses in teaching had to do with holding students accountable

for their work and asking a wide range of questions about the material. These researchers also found, incidentally, a wider variation in practices among the non-prepared teachers than among the prepared teachers.

The studies in this genre are remarkably diverse in the outcomes they measure and in the way they select groups for contrast. Not surprisingly, given this diversity, their findings are mixed as well. Most reviewers of this genre (e.g. Haberman, 1985; Evertson, Hawley and Zlotnik, 1985) perceive the overall pattern of differences as indicating that teacher education does make a difference, though they also point out that these studies do not take account of the content or character of teacher education programs. They share this weakness with multiple regression studies. Now let's consider our three questions.

### *Aspects of Teacher Education Examined*

Instead of examining the particular courses teachers took, researchers in this genre usually focus on entire programs, defining 'teacher education' as whatever set of courses the teacher took in order to receive a certificate. The aspect of teacher education that is of interest to them is the complete program relative to the incomplete, or non-existent, program. Since skeptics are particularly critical of policies that require teachers to complete programs in order to receive certificates, these studies are relevant to their concern. And comparisons among different types of programs -- for instance, alternative routes versus traditional programs -- are also of interest, particularly in policy climate that encourage alternatives.

Many of these researchers also document the number of undergraduate, as opposed to graduate, education courses taken by teachers in the non-certified group. An important finding from this research is that very few provisionally certified or emergency certified teachers have had absolutely no exposure to teacher education. Instead, they have taken a few courses, but not enough to become certified. Thus, many comparisons are actually between teachers who have taken everything that is required to become certified and teachers who have taken some portion of the requirements. Still, these contrasts are relevant, since both policy-makers and teacher educators expect the completed undergraduate program to make a difference.

A significant weakness of these studies, however, is their failure to document which courses were actually taken, or from which institution. As I noted above, the content and character of teacher education can vary substantially from one institution to the next, so that differences in conclusions among these studies could reflect differences in local teacher education programs from which different samples of teachers are drawn.

### *Outcomes*

Although a few comparison studies use tests of knowledge, such as the NTE or a state-required test (e.g. Cornett, 1984), most depend on observations of teacher for

their outcomes. For such observations to be valuable, we would need wide agreement that the practices they document are indeed valued. In fact, because ideas change over time regarding what counts as good teaching, the observation systems change as well, with each reflecting what is fashionable at the time the study was conducted. One could argue, of course, that even though the criteria change over time, each criterion reflects views of good teaching that would also have guided teacher education programs at the time the studies were done. If this is true, it is not unreasonable to expect certified teachers to perform better than non-certified teachers in these studies.

### *The Logic of the Argument*

Studies in this genre actually represent a two-sided argument. On one side, skeptics occasionally conduct such studies to see if teacher education *hinders* teaching. On the other side, advocates conduct such studies to learn whether teacher education *facilitates* teaching. But there is a serious limitation to both sides of this argument, for neither side can be certain that its two groups of teachers did not already differ before they enrolled in their college programs. If those who enroll in teacher education are qualitatively different from those who do not, these differences may still be contributing to teachers' observed pedagogies later on. And there is reason to believe that such prior differences exist. Kennedy (forthcoming), for instance, found that different types of teacher education programs tended to enroll teacher candidates who held different views, even upon entry into the programs, and Skipper and Quantz (1987) found differences between freshmen enrolled in liberal arts programs and those enrolled in teacher education programs.

In fact, even a finding of *no difference* does not avoid this dilemma, for it is possible that different kinds of people enroll in different programs and that the programs washed out the initial differences. Skipper and Quantz (1987) illustrate this point. They followed a group of arts and sciences students and a group of teacher education students from their freshman year through their senior year. They found that substantial differences existed between the two groups as freshmen, but that these differences had disappeared by the time the groups were seniors. No difference at the end of a program, then, means no evidence that teacher education has hindered teaching, no evidence that teacher education has contributed to teaching, and no evidence that different kinds of people enrolled in different programs to start with.

Grossman's (1990) comparisons of six novice teachers with different educational backgrounds improves over many comparison studies in three important ways. First, Grossman interviewed her teachers as well as observing them, and asked them where they got their ideas for teaching. Many referred to specific teacher educators. Second, Grossman visited the teacher education program attended by her sample of certified teachers so that she could test the idea that the program in fact provided content that was consistent with the ideas the novice teachers held. Finally, one of the certified teachers in Grossman's study had actually tried

teaching earlier, and, based on this experience, returned to college to obtain a teaching credential. This teacher's recollections of the first teaching experience provided yet another source of evidence to justify the inference that credentialed teachers taught differently because of what they had been taught in their teacher education program.

Overall, then, comparisons of credentialed and non-credentialed teachers focus on a more relevant aspect of teacher education -- completed programs versus alternatives or incomplete programs -- than do follow-up surveys, and their outcomes are more relevant as well, at least within the era in which they are conducted. But comparison studies suffer a logic problem in that they do not enable researchers to separate the impact of the courses or programs teachers took from the original beliefs or dispositions which may have motivated some teachers to take formal course work in the first place, and motivated others to enter the field later on.

### **Experiments in Teacher Education**

The fourth way to find out whether teacher education makes a difference is to experimentally test the effects of specific approaches to teacher education. This genre has been especially popular among people interested in microteaching, but experiments have also been used to study the effects of hypermedia (Goldman and Barron, 1990), video demonstrations (Winitzky and Arends, 1991), direct instruction (Klesius, Searls and Zielonka, 1990) and a variety of other teacher education strategies. Generally, researchers working within this genre contrast two or more approaches to teacher education in an effort to discern the relative merits of each.

Experiments avoid many of the limitations of the first three genres. They always contrast two or more clearly defined program variations, rather than leaving teacher education undefined; they often assess teachers' knowledge or skill *prior to* their participation in the program as well as after it, so that they can be more sure that whatever differences are eventually observed are due to program differences rather than pre-existing differences; and they usually assess the outcome of interest directly, rather than asking teachers to judge their own performance. In addition, they often randomly assign teacher candidates to program variations to insure that groups receiving different variations do not differ in their prior knowledge or motivations prior to participating. The combination of these features gives experimental researchers a tremendous advantage over those using follow-up surveys, those comparing credentialed and non-credentialed teachers and those looking for predictors of student achievement. Experimental researchers can ascertain how teachers differed both before and after program participation, they know what the program actually did, and they often directly assess the outcomes the program intended to influence.

Recently, as the popularity of microteaching has waned, experiments with other forms of teacher education have begun to appear. One example is a series of studies reported by Nancy Winitzky and Richard Arends (1991). These researchers first contrasted the effects of visiting exemplary classrooms with that of observing videotapes,

and found both to be equally effective in helping teachers use cooperative grouping in their own microteaching. In a second study, they contrasted two methods of developing novices' intellectual schemata regarding cooperative grouping, and in the third, they contrasted learning in the exemplary classrooms with learning via microteaching, and found them to be equally effective. Like many such studies, though, these studies did not follow the students into their own student teaching experiences to see the extent to which they carried their new skills into their own teaching practice.

Now let's consider my three questions.

### *Aspects of Teacher Education Examined*

More than any of the other studies, these studies tend to focus on specific approaches to teacher education. Studies in this genre are especially relevant to teacher educators, but are less relevant to policy makers and skeptics, for they do not address the value of entire programs. Even for teacher educators, many of these studies suffer because they are too short in duration. They may contrast relatively small program units -- three weeks of one approach versus three weeks of another, or even three hours of one versus three hours of the other. They do this, of course, in part because smaller projects are easier to manage. But evidence from such studies cannot be used to make larger scale changes in the structure of teacher education programs. Nor is it realistic to expect such brief experiences to have lasting effects on teacher candidates.

### *Outcomes*

With respect to outcomes, most of these researchers focus on teacher's ability to perform specific skills. The outcomes are directly relevant to teacher educators who want to develop those skills, but often are limited in that they do not consider the teachers' affective response to the skill. That is, it is possible for teachers to learn what the program intends, and to perform it on demand, but to disapprove of it and never use it again once the teachers are practicing independently. Knowing how to perform does not assure a desire to perform, and many experimental studies have found that teachers learned the target skills but did not practice them once in their own classroom.

### *The Logic of the Argument*

The logic of these studies is persuasive: if teachers participating in one approach improve their skill more than other teachers do, then this approach has a greater impact. Because researchers have assessed their candidates' skills both before and after the candidates participated in these approaches, and because candidates are

often randomly assigned to approaches, experimenters can be more sure than other researchers that the outcomes they observe do reflect program influence rather than sampling differences.

Overall, then, these studies are more relevant to teacher educators in the aspects of teacher education they examine, more relevant in the outcomes they assess, and more powerful in their ability to draw unambiguous findings regarding the relative merits of one program approach over another. They do not, however, address the skeptics' concerns, for they do not address the question of whether teacher education as a whole is worthwhile, and they often fail to show whether even the unit they were experimenting with had an influence on teaching practice once teachers graduated and began independent teaching.

### **Longitudinal Studies of Change**

The fifth genre, a relatively new addition to the field, involves following teacher candidates as they proceed through their college education, gathering data on them at several points along the way, to see whether and how their ideas about teaching change over time. Researchers working within this genre want to learn what students are like when they enter their programs, how they change over time in response to their programs, and what they are like when they finish. Like experiments, these studies offer us the advantage of being able to document *change* so that, if differences exist at the end of the study, we can interpret these differences relative to differences that may have existed at the outset. And like experiments, longitudinal studies usually involve looking at the details of the programs in which students participate so that the relationship between student thinking and program content can be examined. Unlike experiments, though, these studies rarely allow us to compare students who participate in different kinds of programs. While we learn more about how they change as they encounter particular aspects of their programs, we cannot say with any confidence how they might have changed if they had participated in some other kind of program.

Longitudinal studies have been especially popular among researchers who are interested in the student teaching component of teacher education (Hodges, 1982; Silvernail and Costello, 1983; Tabachnik and Zeichner, 1984; Goodman, 1986). In the typical study, researchers contrast student-teachers' beliefs or knowledge or skills before and after they participate in student teaching. While these studies are valuable, and have increased our understanding of student teaching, they leave untouched the center of the enterprise—the large, diffuse, complicated web of courses and other events that we call preservice teacher education.

One of the earliest and best examples of an effort to move this genre into the university program is Feiman-Nemser and Buchmann's (1989) study of teacher candidates participating in two different teacher education programs. They followed six students participating in two preservice teacher education programs, interviewing them on several occasions about their understanding of what they were learning and about their views of teaching. They also observed the teacher

## *The Problem of Evidence in Teacher Education*

education courses these students took. Through their descriptions of these students, they were able to demonstrate gradual shifts in views and to demonstrate ways in which program content was sometimes misinterpreted by the candidates. The study demonstrated the importance of candidates' initial assumptions and the ways in which they combined their own childhood experiences with the lessons they were being taught to form their ideas about teaching and learning.

Another good illustration of this genre is Hollingsworth's (1989) study of MAT (Masters of Art in Teaching) candidates. She followed candidates through both university courses and teaching internships. Through her investigation, she showed not only the role that prior beliefs played in these teachers' learning, but also how university learning connected to practical experience. She found that candidates' prior beliefs influenced their receptivity to the program, and that they went through several distinct phases in their practice as they tried to accommodate what they had learned in the program with their classroom experiences.

Longitudinal studies range from cases of individuals (e.g. Valli and Agnostonelli, 1992) to large-scale quantitative studies (Galluzzo, 1984). On the quantitative side, Galluzzo gave candidates the National Teaching Examination (NTE) each year during their college program. Interestingly, he found that scores did not change on the general studies portion of the NTE, but did change on the professional knowledge portion. Quantitative studies are particularly popular for assessing the impact of the entire college experience, and Pascarella and Terenzini (1991) provide a remarkably thorough summary of this literature.

Kennedy (forthcoming) presents an example of a large-scale, but qualitative, longitudinal study, the Teacher Education and Learning to Teach (TELT) study. TELT researchers interviewed teachers and teacher candidates multiple times as they participated in different teacher education programs. Her data allow us to see the initial views of teachers entering these different programs, and to see how programs with different content move teachers' thinking in different directions. They also show, however, that different types of programs tend to attract teachers with different initial beliefs. Two findings from that study are especially important. First, nearly every group of candidates changed views as they moved through these programs, and these changes were virtually always in the direction of the program's substantive orientation. However, the changes were not as substantial as teacher educators would probably want to see. Moreover, because different programs had different substantive orientations, the net effect across all programs could appear as if teacher education had not made a difference. That is, some teachers moved in one direction while others moved in the opposite direction.

And now to my three questions.

### *Aspects of Teacher Education Examined*

The aspect of teacher education that longitudinal researchers tend to focus on is the whole program *as it is perceived by the candidates themselves*. That is, longitudinal researchers tend to follow teacher candidates through their coursework, noting how

the candidates interpret and respond to the various courses that they take. Further, instead of allowing official program rhetoric to define the courses students take, they often actually attend courses with their sample students. This aspect of teacher education is more relevant than that of experiments, since longitudinal studies tend to examine entire programs. It is not clear whether this aspect of teacher education programs is of interest to skeptics or policy-makers.

### *Outcomes*

With respect to outcomes, longitudinal researchers tend to be more interested in teachers' beliefs and attitudes than in their tested knowledge or their teaching skills. The importance of beliefs has become apparent in microteaching experiments, when teachers learn skills but do not apply them. And the literature indicating the relationship between beliefs and actions is growing as well. Often, however, these studies are limited because they do not follow candidates into practice, to learn whether the changes observed during a program extend into the teachers' classroom practices. Thus, they document changes during preservice teacher education, or during student teaching, but do not tell us whether these changes are sustained later on or whether they actually influence teaching practices.

### *The Logic of the Argument*

Most longitudinal studies are based on the assumption that teacher candidates enter their college programs with a set of initial beliefs that will influence their responses to the courses they take. As candidates participate in teacher education courses, they incorporate some new ideas, but they also reinterpret many of them to make them more consistent with what they already believed.

Studies of change share with multiple regression a tendency to document a wide range of variables, most of which are defined in advance. But whereas multiple regression researchers are measuring various influences on student achievement, longitudinal researchers are measuring a complex of attitudes, values and beliefs and the multitude of experiences that influence these. The success of longitudinal studies, therefore, can depend on the theory of learning that guides data collection.

One difficulty that some such studies encounter is a confusion between change due to normal maturation among college students and change due to program impact. The fact that students are changing and developing over time does not necessarily mean that these changes are a result of a program. College students are still in a highly formative stage in their lives, and may be changing in several ways that have little to do with the particular courses or curricula they encounter as students. Thus, the credibility of longitudinal studies depends either on comparison groups, which permit changes to be contrasted across program types, or on documenting the content of the program so that a clear relationship can be shown between what teachers think and what they have been taught.

Another difficulty that can arise in longitudinal studies derives from the number of observations made on students. Students are often interviewed, or respond to questionnaires, on numerous occasions, and it is highly likely that, over time, they learn not only what they will be asked but also how they are supposed to respond. Thus there is a chance that the researchers themselves are at least partly responsible for the changes they describe.

Overall, then, these studies focus on relevant aspects of teacher education -- undergraduate programs and components within those programs, and on relevant outcomes -- changes in knowledge and beliefs about teaching. The logic is also sound, provided that attention is given to sorting out natural maturation from program effects. To the extent that longitudinal studies attend to whole programs, and particularly to differences among programs, they may be relevant to policy makers and skeptics. The TELT study, for instance, suggests that programs of teacher education do make a difference, and that different programs make different differences.

### **Conclusions**

Skeptics of teacher education seek evidence that teacher education is sufficiently beneficial to justify that it be required. All five of these research genres are intended to document whether or how teacher education makes a difference, but they examine different aspects of teacher education, they document different types of outcomes, and they employ different forms of logic to establish their case for whether teacher education makes a difference.

Of the five genres reviewed here, findings from multiple regressions are more likely to find their way into the policy community than other findings, for these studies are designed to inform policy decisions. Taken together, these studies do not make a strong case for teacher education. Moreover, the only follow-up surveys to enter the policy making community are those conducted outside of teacher education, such as the NEA study, and that study also does not make a case for teacher education. Some comparisons of credentialed and noncredentialed teachers have entered the policy arena, and these tend to be more favorable toward teacher education. However, even these studies have mixed results, depending on when and where they are done and depending on what outcomes are measured.

All of these genres also have significant weaknesses, however. The first three, for instance, cannot show with certainty that the behaviors, knowledge, or attitudes they document resulted from teacher education programs, rather than from prior dispositions of those who choose to enter teacher education programs. The two remaining genres -- experiments and longitudinal studies -- offer the strongest possibilities for causal inferences. Both of these genres document changes in teachers, rather than limiting themselves to how teachers respond after they have completed their programs, and both document the content of the program so that they can show the relationship between the changes that were observed and program content. Each, however, is limited in its potential to persuade skeptics and policy makers. Experiments do not address whole programs, and tend to be of very short duration,

and longitudinal studies are often case studies, limited to a handful of teacher candidates. And they frequently are limited only to attitudes and beliefs, and do not include teaching practice. So far, neither of these genres has shown that a large group of teachers change in desired ways as a result of participating in teacher education, and continue to practice in improved ways once in the field.

The sad fact is that poorly designed studies are not merely noninformative. Often, they are misinformative: by failing to document the content and character of teacher education programs, they confuse quantity with quality. By failing to consider what teacher candidates know or think prior to participating in teacher education, they may over- or under-estimate the contributions of teacher education. By failing to consider the context in which teachers are teaching, they may confuse the effects of the current teaching context with the effects of the earlier teacher preparation. When they study only handfuls of teachers, they cannot tell us how widespread their observed changes are likely to be. To maximize the potential of their studies, researchers need to address all of these concerns. Until such time as an adequate body of such knowledge exists, teacher education will continue to have a hard time defending itself from skeptics.

### **Note**

- <sup>1</sup> An earlier version of this paper was presented at the Annual Meeting of the Association of Teacher Educators, February, 1991. Preparation of the paper was supported by the National Center for Research on Teacher Learning under a grant from the US Department of Education, Office of Research. However, the views expressed here are the authors and no official endorsement should be inferred.

### **References**

- ADAMS, R. D. and CRAIG, J. R. (1983) 'A status report of teacher education program evaluation', *Journal of Teacher Education*, 34, 2, pp. 33-6.
- ARCH, E. C. (1989) 'Comparison of student attainment of teaching competencies in traditional preservice and fifth-year master of arts in teaching programs', Presented at the annual meeting of the American Educational Research Association, San Francisco, March.
- BEERY, J. R. (1960) *Professional Preparation and Effectiveness of Beginning Teachers*, Coral Gables: Graphic Arts Press.
- BLEDSON, C., COX, J. V., and BURNAM, R. (1967) *Comparison between Selected Characteristics and Performance of Provisionally and Professionally Certified Beginning Teachers in Georgia*, ERIC No. ED 015 553.
- BROPHY, J. and GOOD, T. L. (1986) 'Teacher behavior and student achievement,' in WITTROCK, M. (ed.) *Handbook of Research on Teaching (Third Edition)*, New York: MacMillan, pp. 328-75.
- BROWN, D., EDINGTON, E., SPENCER, D. A., and TINAFERO, J. (1989) 'A comparison of alternative certification, traditionally trained, and emergency permit teachers,' *Teacher Education and Practice*, 5, 2, pp. 21-3.

## *The Problem of Evidence in Teacher Education*

- BROWNE, D. and HOOVER, J. H. (1990) 'The degree to which student teachers report using instructional strategies valued by university faculty,' *Action in Teacher Education*, 12, 1, pp. 20-4.
- CLARK, D. C., SMITH, R. B., NEWBY, T. J., and COOK, V. A. (1985) 'Perceived origins of teaching behavior,' *Journal of Teacher Education*, 36, 6, pp. 49-53.
- COLEMAN, J. S. (1966) *Equality of Educational Opportunity*, Washington DC: US Department of Health, Education, and Welfare.
- COPELAND, W. D. (1975) 'The relationship between micro-teaching and student teacher classroom performance,' *Journal of Educational Research*, 68, pp. 289-93.
- COPELAND, W. D. (1982) 'Laboratory experiences in teacher education,' *Encyclopedia of Educational Research* (V. 2, Fifth Edition), pp. 1008-19.
- CORNETT, L. (1984) 'A comparison of teacher certification test scores and performance evaluations of graduates in teacher education and liberal arts and sciences in three southern states,' Atlanta, GA: Southern Regional Education Board.
- CORNFIELD, I. R. (1991) 'Microteaching skill generalization and transfer: Training pre-service teachers in introductory lesson skills,' *Teaching and Teacher Education*, 7, 1, pp. 25-56.
- COUNCIL OF CHIEF STATE SCHOOL OFFICERS (1988) *State Education Indicators, 1988*, Washington, DC: Author.
- CRUICKSHANK, D. R. and METCALF, K. K. (1990) 'Training within teacher preparation,' in HOUSTON, R. W., HABERMAN, M. and SIKULA, J. (Eds) *Handbook of Research on Teacher Education*, New York: Macmillan (pp. 469-97).
- DEWALT, M. and BALL, D. W. (1987) 'Some effects of training on the competence of beginning teachers,' *Journal of Educational Research*, 80, 6, pp. 343-47.
- EVERTSON, C. M., HAWLEY, W. D., and ZLOTNIK, M. (1985) 'Making a difference in educational quality through teacher education,' *Journal of Teacher Education*, 36, 3, pp. 2-12.
- FEIMAN-NEMSER, S. and BUCHMANN, M. (1989) 'Describing teacher education: A framework and illustrative findings from a longitudinal study of six students,' *Elementary School Journal*, 89, pp. 365-77.
- FERGUSON, R. (1990) *Racial Patterns in How School and Teacher Quality Affect Achievement and Earnings*, Cambridge, MA: Kennedy School of Government mimeo.
- FERGUSON, P. and WOMACK, S. T. (1993) 'The impact of subject matter and education coursework on teaching performance,' *Journal of Teacher Education*, 44, 1, pp. 55-63.
- GAGE, N. L. (1977) *The Scientific Basis of the Art of Teaching*, New York: Teachers College Press.
- GAGE, N. L. (1985) *Hard Gains in the Soft Sciences: The Case of Pedagogy*, Bloomington, IN: Phi Delta Kappa.
- GALLUZZO, G. (1984, March) 'An evaluation of a teacher education program,' Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- GALLUZZO, G. and CRAIG, J. R. (1990) 'Evaluation of preservice teacher education programs,' in HOUSTON, W. R., HABERMAN, M. and SIKULA, J. (Eds) *Handbook of Research on Teacher Education*, New York: Macmillan (pp. 599-616).
- GAEDE, O. F. (1978) 'Reality shock: A problem among first-year teachers,' *The Clearinghouse*, 51, 9, pp. 405-9.
- GLIESSMAN, D. H. (1987) 'Changing complex teaching skills,' *Journal of Education for Teaching*, 13, 3, pp. 267-75.

- GLIESSMAN, D. H. and PUGH, R. C. (1987) 'Conceptual instruction and intervention as methods of acquiring teaching skills,' in TELLEMA, H. and VEENMAN, S. A. M. (Eds) 'Developments in training methods for teacher education,' *International Journal of Education*, 11, pp. 555-63.
- GOLDMAN, E. and BARRON, L. (1990) 'Using hypermedia to improve the preparation of elementary teachers,' *Journal of Teacher Education*, 41, 3, pp. 21-31.
- GOODMAN, J. (1986) 'What students learn from early field experiences: A case study and critical analysis,' *Journal of Teacher Education*, 36, 6, pp. 42-8.
- GRAY, H. B. (1962) 'A study of the outcomes of preservice education associated with three levels of teacher certification,' Doctoral Dissertation, Florida State University.
- GROSSMAN, P. L. (1990) *The Making of a Teacher: Teacher Knowledge and Teacher Education*, New York: Teachers College Press.
- HABERMAN, M. (1985) 'Does teacher education make a difference? A review of comparisons between liberal arts and teacher education majors,' *Journal of Thought*, 20, 2, pp. 25
- HALL, H. O. (1964) 'Professional preparation and teacher effectiveness,' *Journal of Teacher Education*, 15, 1, pp. 72-6.
- HANUSHEK, E. A. (1971) 'Teacher characteristics and gains in student achievement: Estimation using micro data,' *American Economic Review*, 61,2, pp. 280-S.
- HANUSHEK, E. A. (1989) 'The impact of differential expenditures on school performance,' *Educational Researcher* 18, 3, pp. 45-5 1, 62.
- HENRY, M. (1987-88) 'The effect of increased exploratory field experiences upon the perceptions and performance of student teachers,' *Action in Teacher Education*, 9, 4, pp. 93-7.
- HODGES, C. (1982) 'Implementing methods: If you can't blame the cooperating teacher who can you blame?' *Journal of Teacher Education*, 33, 6, pp. 25-9.
- HOLLINSWORTH, S. (1989) 'Prior beliefs and cognitive change in learning to teach,' *American Educational Research Journal*, 26, 2, pp. 160-89.
- KENNEDY, M. M. (1991) 'Some surprising findings on how teachers learn to teach,' *Educational Leadership*, 49, 11, pp. 14-7.
- KENNEDY, M. M. (1996) 'Research genres in teacher education,' in MURRAY, F. (ed.) *The Teacher Educator's Handbook: Building a Knowledge Base for the Preparation of Teachers*, San Francisco: Jossey Bass.
- KENNEDY, M. M. (forthcoming) *Learning to Teach Writing: Does Teacher Education make a Difference?*, New York: Teachers College Press.
- KLESIOUS, J. F., SEARLS, E. F., and ZIELONKA, P. (1990) 'A comparison of two methods of direct instruction of preservice teachers,' *Journal of Teacher Education*, 41, 4, pp. 34-43.
- KOZIOL, S. M., JR., and BURNS, P. (1986) 'Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory,' *Journal of Educational Research*, 79, 4, pp. 205-9.
- LUPONE O. J. (1961) 'A comparison of provisionally certified and permanently certified elementary teachers in selected school districts in New York State,' *The Journal of Educational Research*, 55, 2, pp. 53-63.
- MCINTYRE, D. J. and KILLIAN, J. E. (1986) 'Students' interactions with pupils and cooperating teachers in early field experiences,' *The Teacher Educator*, 22, 2, pp. 2-9.
- MCINTYRE D. J. and KILLIAN, J. E. (1987) 'The influence of supervisory training for cooperating teachers on preservice teacher development during early field experiences,' *The Journal of Educational Research*, 80, 5, pp. 277-82.

## *The Problem of Evidence in Teacher Education*

- MURNANE R. J. and PHILLIPS, B. R. (1981) 'Learning by doing, vintage and selection: Three pieces of the puzzle relating teaching experience and teacher performance', *Economics of the Education Review*, 1, 4, pp. 453-65.
- MURNANE R. J., SINGER J. D., WILLET, J. B., KEMPLE, J. J. and OLSEN, R. S. (1991) *Who Will Teach? Policies that Matter*, Cambridge: Harvard University Press.
- NATIONAL CENTER FOR EDUCATION STATISTICS (1989) *Digest Of Education Statistics 1989*, Washington, DC: Author.
- PASCARELA, E. and TERENCEZINI, P. (1991) *How College Affects Students: Findings and Insights From Twenty Years of Research*, San Francisco: Jossey Bass.
- PECK, H. I. (1989) 'The effect of certification status on the performance of mathematics teachers: A pilot study,' Paper presented at the annual meeting of the American Educational Research Association, San Francisco, March.
- PIGGE, F. L. (1978) 'Teacher competencies: Need, proficiency, and where proficiency was developed,' *Journal of Teacher Education*, 29, 4, pp. 70-6.
- POPHAM, W. J., (1971) 'Performance tests of teaching proficiency: Rationale, development, and validation,' *American Educational Research journal* 8, 1, pp. 105-17.
- SILVERNAIL, D. L. and COSTELLO M. H. (1983) 'The impact of student teaching and internship programs on preservice teachers' pupil control perspectives, anxiety levels, and teaching concerns,' *Journal of Teacher Education*, 36, 4, pp. 32-6.
- SKIPPER, C. E. and QUANTZ, R. (1987) 'Changes in educational attitudes of education and arts and sciences students during four years of college,' *Journal of Teacher Education*, 38, 3, pp. 39-44.
- SMYLIE, M. A. (1989) 'Teachers' views of the effectiveness of sources of learning to teach,' *The Elementary School Journal*, 89, 5, pp. 543-58.
- STRANG, H. R., BADT, K. S., and KAUFFMAN, J. M. (1997) 'Microcomputer-based simulations for training fundamental teaching skills,' *Journal of Teacher Education*, 38,1, pp. 20-6.
- STRAUSS, R. P. and SAWYER, E. A. (1986) 'Some new evidence on teacher and student competencies,' *Economics of Education Review*, 5, 1, pp. 41-5.
- TABACHNIK, B. R. and ZEICHNER, K. M. (1984) 'The impact of student teaching experience on the development of teacher perspectives,' *Journal of Teacher Education*, 35, 6, pp. 28-36.
- VALLI, L. and AGNOSTINELLI, A. (1992) 'Teaching with and without formal preparation,' Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- WINITZKY, N. E. and ARENDS, R. (1991) 'Translating research into practice: The effects of various forms of training and clinical experiences on preservice students' knowledge, skills and reflectiveness,' *Journal of Teacher Education*, 42, 1, pp. 52-65.