

Exogenous Variables and Value-Added Assessments: A Fatal Flaw

by David C. Berliner – 2014

Background: *There has been rapid growth in value-added assessment of teachers to meet the widely supported policy goal of identifying the most effective and the most ineffective teachers in a school system. The former group is to be rewarded while the latter group is to be helped or fired for their poor performance. But, value-added approaches to teacher evaluation have many problems. Chief among them is the commonly found class-to-class and year-to-year unreliability in the scores obtained. Teacher value-added scores appear to be highly unstable across two classes of the same subject that they teach in the same semester, or from class to class across two adjacent years.*

Focus of Study: *This literature review first focuses on the confusion in the minds of the public and politicians between teachers' effects on individual students, which may be great and usually positive, and teachers' effects on classroom mean achievement scores, which may be limited by the huge number of exogenous variables affecting classroom achievement scores. Exogenous variables are unaccounted for influences on the data, such as peer classroom effects, school compositional effects, and characteristics of the neighborhoods in which some students live. Further, even if some of these variables are measured, the interactions among these many variables often go unexamined. But, two-way and three-way interactions are quite likely to be occurring and influencing classroom achievement. This analysis promotes the idea that the ubiquitous and powerful effects on value-added scores of these myriad exogenous variables is the reason that almost all current research finds instability in teachers' classroom behavior and instability in teachers' value-added scores. This may pose a fatal flaw in implementing value-added assessments of teaching competency.*

Research Design: *This is an analytic essay, including a selective literature review that includes some secondary analyses.*

Conclusions: *I conclude that because of the effects of countless exogenous variables on student classroom achievement, value-added assessments do not now and may never be stable enough from class to class or year to year to be used in evaluating teachers. The hope is that with three or more years of value-added data, the identification of extremely good and bad teachers might be possible; but, that goal is not assured, and empirical results suggest that it really is quite hard to reliably identify extremely good and extremely bad groups of teachers. In fact, when picking extremes among teachers, both luck and regression to the mean will combine with the interactions of many variables to produce instability in the value-added scores that are obtained. Examination of the apparently simple policy goal of identifying the best and worst teachers in a school system reveals a morally problematic and psychometrically inadequate base for those policies. In fact, the belief that there are thousands of consistently inadequate teachers may be like the search for welfare queens and disability scam artists—more sensationalism than it is reality.*

Not long ago, I met a teacher who had just retired. As she told me about her interesting 30-year career as a classroom teacher, she said that for 10 years she had been a great teacher, for 10 years a really poor teacher, and then for another 10 years a great teacher once again. Naturally, I was intrigued and asked her how that had happened.

She explained that she taught all 30 years in the same school in East Los Angeles, a section called Boyle Heights. During that lengthy time period, she changed her teaching style and curriculum only a little, but the judgments made by administrators about her teaching changed. The reason, she thought, was that the students she served kept changing. For the first 10 years of her teaching career, she worked with a largely American-born, upwardly mobile, Jewish student body. The next decade, she worked largely with an immigrant Hispanic student body that was poor and had language difficulties. For her last decade, she served a largely Asian immigrant student body that had language problems, but were not all impoverished, nor were their families uneducated. Those students were obedient in school and received help with their studies after school. *Who* this teacher taught made her look terrific, or bad, at least in the eyes of the state of California.

Her story impressed me, since so many of my fellow Americans, as part of our near constant search for magical solutions to our perceived school problems, are great believers in the power of teachers *regardless of context*. Family, friends, and I all have similar stories to tell about how each of us was affected by some teacher. But the effect of teachers across students in large classroom groups may not be nearly so direct and as powerful as we think. If that is so, we are in danger of making an attribution error because we confuse the powerful *individual* effects teachers may have, with overall teacher effects on the achievement of the group of students they teach. These group effects may well be moderated by variables we have not always recognized, the most powerful of which may be peer and compositional effects of classrooms and schools. A simple example of the ease with which these sources of influence on achievement can be confused is found by asking American elementary school teachers the following simple question: Which class would show the most growth in achievement, a class with 18 girls and 12 boys or one with 18 boys and 12 girls? Every teacher I asked picked the class with 18 girls and 12 boys because, they said, girls are better behaved, they do their homework more frequently, many girls admire their teachers because at that age they want to be teachers themselves, their handwriting is readable, and they are more likely to study for their examinations instead of playing with their Xbox or going online to play World of Warcraft. My point is that if we judge teacher effectiveness by using some version of student growth in test scores, what are called “value-added” models of teacher effectiveness, we may be making attribution errors about effectiveness

that are enormous.

What we ordinarily think of as “teacher effects” are not purely, or perhaps not even predominantly, causally attributed to the teacher. Rather, effects on achievement may actually depend on the interactions among many variables. Among these are peer or compositional effects of classrooms and schools, curriculum effects, school leadership effects, school climate effects, technology integration and usage effects, district leadership effects, per pupil expenditure effects, professional development activity effects, and so forth. The simple model of influence, Teacher \longrightarrow Student, held so widely by the general public, and particularly by our politicians, is surely reciprocal, and more like this: Teacher \longleftrightarrow Student. And, 25-35 of such separate relationships need to be negotiated in every classroom. To add to the common misunderstandings about influence patterns, Herbert Marsh (personal communication, 2013) reported one study in which the preexisting classroom climate was unchanged by the classroom teacher, making the relationship Student \longrightarrow Teacher.

Through research, we have become aware that there are many inside-the-classroom variables that affect these individual relationships, for example, the percentage of females and males, the percentage of low- and high-ability students, the number of special education students or English Language Learners, the percentage of children whose parents are college educated, and so forth (cf. Henry & Rickman, 2007; Hoxby, 2000; Lavy, Silva, & Weinhardt, 2009; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Schindler-Rangvid, 2003; Schneeweis & Winter-Ebmer, 2005; Willms, 2006). Then, there are the many outside-the-classroom variables affecting teaching and learning inside the classroom. These include school variables such as teacher turnover rate, books in the school library, after-school programs available, percentage of time a school nurse and school counselor are available, and so forth (cf. Gysbers, 2008; International Association of School Librarianship, 2008; Little, Wimer, & Weiss, 2008; National Association of School Nurses, 2012; and Ponfeldt, Loeb, & Wycoff, 2012). Classroom processes are also affected by outside-the-school variables such as local violence rates, poverty rate, hunger rate, single-parent family rate, mobility rates of families, percentage of home ownership and rentals in the community, day care attendance rates, and so forth (cf. Attar, Guerra, & Tolan, 1994; Belsky et al., 2007; Berliner, 2006, 2009, 2012).

Many of these outside-the-classroom and outside-the-school variables, as well as the more proximal inside-the-classroom variables are hidden from us, lurking as unexamined exogenous variables. In economics, an exogenous variable is one that is outside the statistical model being used, and is unexplained by the model. Exogenous variables with a reasonable likelihood of affecting the teaching and learning process in the classroom probably number in the hundreds, if not more. Even when we specify a few of these variables, we often measure them quite imperfectly and, thus, they are poorly controlled. Regardless of whether these variables are measured well or poorly, these variables are rarely analyzed as part of second-order, third-order, or higher order interaction terms. If you don't look for interactions, or cannot look for them because of the limitations of the statistical methods used, that does not rule out the occurrence of second- and third-order interactions! We should remember that one reason for the rise in qualitative educational research over the last few decades was the recognition that quantitative research cannot find reliable measures for all that occurs in schools that is worthy of study. Nor can quantitative research ever capture all the interactions that occur.

The complexity of the teaching and learning process in classrooms just described leaves us with a validity problem, a problem of inferring the “causes” of students' achievements in school. Student achievement appears to be only partially attributable to teachers, and that contribution to achievement may be much less than is commonly believed. Yet, the goal of all current value-added models is to infer a teacher's contribution to a student's test score *independent* of these other, often powerful and often unaccounted for, variables. In these circumstances, it is too easy to draw an invalid inference about a teacher's competence from the data we collect unless we pay attention to the myriad of variables that have an effect on what goes on in classrooms and schools. Because the number of such variables may be impossible to specify and measure, value-added approaches to judging teacher effectiveness may be fatally flawed.

The advocates of value-added evaluation think that a teacher is a teacher is a teacher, that no matter where you drop them, in whatever classroom, with whatever grade level, with whatever subject matter, in whatever school, with whatever population that attends the school, with whatever resources, a good teacher will still be a good teacher. This is nonsense. But, it is a basic premise of those promoting value-added models in our country. This unfounded belief that we have superb and bad teachers, without respect to the kinds of schools they teach in and the kinds of students they teach, is an accepted part of life in contemporary America. But, this belief is factually wrong, though still influencing national policy, and may well do serious harm to the profession.

A belief in strong and independent teacher effects is based on a form of bizarre logic as well. My fellow professors of higher education, including those who support value-added approaches to teacher evaluation, do not usually feel responsible for classroom learning. Their college students are. In medicine, physicians aren't responsible for health, their patients are. But, we in the United States have bought the bizarre notion that our public school teachers do not just affect student learning by their provision of opportunities to learn, they are also *responsible* for student learning (Cuban, 2012). Thus, we focus on teacher effects and do not deal honestly with the role of student motivation for learning and outside-the-school factors that also affect school learning.

A further note about responsibility for learning is that, over the years, when the sources of variance in student achievement were examined, teachers were never found to be as powerful a force as the press, public, or politicians believed them to be. Analysis of many studies reveals that out-of-school factors (e.g., SES, ability level, and attendance rate) account for about 60% of the variance in student achievement; inside-the-school factors account for about 20% of that variance; and error (unknown factors)

account for another 20% of the variance in student achievement scores (Goldhaber, 2002). Thus, outside-the-school factors are about three times more powerful than inside-the-school factors. But, teachers are only one of many inside-the-school factors accounting for that relatively weak effect. Besides teachers, many other inside-the-school factors influence student achievement, such as school leadership, teacher turnover, provision of school counselors and nurses, etc. Estimates are that teachers are the most powerful of the school effects, but still only account for 10-20% of the variance in mean student achievement for their classes.

A dissent from this view comes from a recent study by Kesting, Chen, and Stigler (2013) on the sources of variance in achievement scores. They claimed that a larger percentage of student test score variance is attributable to teachers, net of other considerations. They suggested that teacher effects on the variance in student test results may be as high 38%, but some of that is confounded with school effects, which are nontrivial and represent the context issues I am concerned about.

No one I know wants to deny a teacher's power to affect the achievement, motivation, or life course of an individual student. But, there is also consensus that teachers are not the only agent affecting mean student achievement in their classrooms. The press, public, and politicians much too often overestimate the power of the teacher effect on classroom achievement and underestimate the effects of other variables on the test scores that we use to measure achievement. Notable among these other variables are the effects of school, family, and neighborhood on test scores.

To greater or lesser effect, school and student achievement is moderated by the interpersonal relationships among students in a classroom and their teacher, among students and their peers, and the relationships of the school community to the school. Classroom teacher main effects on achievement are not easily separated out from these other effects and that is why value-added models for evaluating teacher quality have proven to be so unreliable from year to year and classroom to classroom in the same year. As their classrooms vary, so will the value-added scores obtained by teachers, as described below. The commonly heard cry by educational reformers is "If we only had good teachers, all would be well in our schools." But, it is more likely that "if we only had well-behaved, highly-committed learners, who came from caring, middle class, two-parent families as our students, all would be well in our schools!" Figuring out the many variables, and the power of those variables to influence student achievement, is not nearly as easy as it appears to politicians and the press.

Currently, American policymakers are blind to the teacher x student or teacher x peer group two-way interactions, and they cannot even conceive of the three-way interactions among teachers, characteristics of classroom peers, and characteristics of neighborhoods from which students come. This unrealistic view occurs, in part, because we Americans hold on to a myth of the individual as hero in so many of the stories we read or see portrayed in movies. We go to theaters to see Clint Eastwood as "Dirty Harry" or as a lone and moral cowboy who always cleans up the city or guns down the bad guys all by himself. In *High Noon*, winner of four Academy Awards and nominated for seven, Gary Cooper had to go it alone. Even the Soviet Union was officially moved to comment on this movie's glorification of individualism. We also watched approvingly as Sylvester Stallone in his role of either Rocky or Rambo, always overcoming the bad guys, mostly by himself.

America loves individuals as heroes. No complexities, no interactions. Only one good teacher is needed to fix everything and run the bad students or the bad administrator out of town. *Blackboard Jungle*, *Lean on Me*, and *Dangerous Minds*—all are films about schooling with strong individual heroes. The most dramatic example of this genre may be *Stand and Deliver*, which portrayed Jaime Escalante's heroic work in Los Angeles teaching calculus to poor Hispanics. He was a lone hero and, thus, was held up as a model for what all teachers ought to be able to do. But as I looked into this hero, I found that his Los Angeles students chose to be in his class. They were not a random draw of general mathematics students in a school that served the poor, although that was not featured in the movie (see Wainer, 2011). While this should not take away from Escalante's achievements, it does put a different light on them. Success as a teacher is always a lot easier to attain with handpicked and volunteer students. Furthermore, when Jaime Escalante moved to Sacramento, he was not as successful in developing his program. What do we make of that? Was Escalante the loner who took on the education system and won because he was a great teacher with great skills, or was Escalante's success and failure influenced by unknown social and contextual factors that made him both more and less effective in different contexts?

Escalante's success and his failure to replicate his success suggest that it is not the teacher, but the teacher in interaction with the students, the school culture, the neighborhood, and the social context of the school that determines a lot of what we call teacher effectiveness. Despite the continuing American belief in the power of teachers to transform lives and the many anecdotes that confirm this belief, and despite the occasional demonstration that a school or teacher actually has these kind of main effects, overall, the power of the individual teacher, independent of the group with which that teacher works, may be much less than most people think.

SEARCHING FOR STABILITY IN TEACHER CLASSROOM BEHAVIOR AND TEACHER EFFECTS ON ACHIEVEMENT

We all can agree that individual teachers frequently influence particular students in their classrooms, just as we hope they might. We can also all agree that certain teachers seem, year after year, to maintain their reputations as teachers, even as students change from year to year. But, the research community now knows that a teacher's behavioral stability and stability in effects on achievement are more illusory than real.

Our misunderstanding about teacher's stability in behavior is based on what we call high-inference (global) characterizations of the teacher's behavior. A warm teacher is generally warm, class after class and year after year. An organized teacher is generally organized year after year and from class to class. A collegial teacher or an enthusiastic teacher or a teacher who is a strong disciplinarian is likely to be seen as consistent in the eyes of school administrators, parents, and colleagues, and in the eyes of the

teachers themselves. At these high levels of inference, there appears to be consistency in teacher behavior. And, we also happen to believe that warm is good, while cold is bad; organized is good, while disorganized is bad; maintaining discipline is good, but chaos in the classroom is bad. So, this is a perfectly adequate method for making choices about who the better and worse teachers are in a school. A reputation for being warm, organized, collegial, fair, able to maintain discipline, and the like are quite reasonable indicators of being a “good” teacher, perhaps also an effective teacher.

Two relatively new observational systems help to evaluate teachers in this way. The observation system developed by Pianta, called CLASS, looks at the teacher’s emotional support of students, instructional support, and classroom organization (Curry School of Education, n.d.). The observation system of Danielson (2012) looks at such things as whether the teacher has developed an environment of respect, is a good manager of classroom procedures, and can keep pupils engaged in learning.

Both of these systems use variables that have, through research, actually been related to measures of teaching effectiveness. But, the variables examined are more global than those that are called low-inference variables, variables that are more molecular, less global. Among the low-inference variables, which are quite reliably measured, are such things as questions asked per hour, the percentage of higher order questions per period, the number of times teachers waited over 3 seconds for a student answer, the positive or negative statements uttered per hour, percentage of time the less able children are called on, and so forth. At the more global level, where it is harder to get observer reliability, teachers show reasonable stability in behavior over observations made in their classrooms. But at the more molecular level, where variables are more reliably measured, teachers are remarkably unstable from one observation to another.

Because of this instability, it is much harder to categorize and rate teachers using these small-grained, more precisely defined behavioral measures. Measuring the frequency and magnitude of variables that seem to be important, such as percentage of high-order questions asked, or number of times that students were asked to extend their answers to questions, or the time spent with the least able students in the class, or the percentage of questions initiated by students, indicates that teachers are highly reactive to particular situations in their class and, thus, unstable in their classroom behavior over time, settings, and circumstances. To obtain barely reliable measures of teacher classroom behavior at this more molecular level requires many different well-trained observers, observing in classrooms over many different occasions, at costs that are usually prohibitive (Bill & Melinda Gates Foundation, 2012).

This now accepted general finding of instability in the more fine-grained classroom behavior of teachers over occasions and (as discussed below) the instability of teachers’ effects on student achievement over years or over similar classes in the same year, suggests three things: First, that the observation instruments and the tests used to judge teachers are unreliable. This is often, but not always true. Second, the observed instability in teacher behavior and teacher effects is due to the fact that practicing teachers are themselves unstable in their professional behavior. That is not likely. The third hypothesis, and the most likely of the explanations, is that the behavior of teachers and their effects on students are highly conditioned by classroom life, along with characteristics of the classroom, school, and community. It is likely that the behavior of students and the characteristics of the classroom, school, and community, as well as the nature of the curriculum being worked on, and even the local weather, are the sources of these instabilities in teacher behavior and teacher effects. This means that we may be overestimating the main effects that teachers have on students and underestimating the interaction effects, and especially the reciprocal nature of the influence patterns in a classroom.

This instability in teacher behavior in the classroom and teacher effects on student achievement results in great complexity for those charged with making policy. A teacher may be fired for being ineffectual in classroom behavior or for effects on student achievement, or the teacher may be rewarded for being excellent in classroom behavior or effects on achievement. But if, in fact, as argued here, both teacher behavior in the classroom and teacher effectiveness with regard to achievement is actually a response to the myriad interactions of the teacher with the class, school, and neighborhood, we might be as well off if we fired the class or rewarded the students instead of firing or rewarding the teacher! Thinking about simple main effects (Teacher → Student) makes life easy for policymakers, but the real world may call for much more interactionist thinking, for complexifying rather than simplifying the way we perceive the world.

In sum, if the power of teachers to influence student achievement is moderated by many other factors, then what we now call the teacher effect is an overstatement about what teachers accomplish on their own, as independent actors. This poses a serious problem for value-added systems of teacher evaluation, all of which depend upon a clear relationship between what teachers do in classrooms and what students achieve on tests. Value-added assessments require an assumption of teacher causality of student growth in achievement. But, if the teacher effect from class to class and year to year is moderated by many different variables, often in interaction, then the reliability of value-added models will always remain low. Always. Simple main effect thinking, coupled with a desire for lone heroes and heroines, ignores the complexity of classroom life while shutting down the search for more sensitive and probably more expensive measures for assessing teacher quality.

TEACHER INSTABILITY IN BEHAVIOR AND IN EFFECTS

We know that the correlations between teacher behaviors assessed by observation instruments and student gain scores in achievement of subject matter from year to year are low and often negligible. Brophy (1979), among others, understood why this was so 30 years ago. After reviewing the extant research on teaching he said, “The influence of context is being recognized as more and more important. [Thus] there do not appear to be any universal teaching competencies . . . that are appropriate in any and all circumstances” (p. 5). Teacher classroom behavior, then, is necessarily unstable, though the better term might well be strategically flexible, adapted to circumstances. But, the outcome of this kind of strategic flexibility is that the reliability of the

instruments we use to measure teaching behavior is usually quite low. This is one reason for the generally low correlations we find between teacher behavior and measure of teacher effects.

The newest version of these old truths was released recently by a stellar group of researchers in a well-done and immensely expensive study funded by the Gates foundation (Bill & Melinda Gates Foundation, 2012). The researchers promoted stringent training of raters, use of multiple raters for coding each teacher, multiple occasions for the ratings, and pooling results across lessons. Even then, the reliabilities for the observational data were not as high as desired and were barely adequate for the study. The authors quite properly noted that the vast majority of the total variance in observational scores using the CLASS and the Danielson instruments described above, as well as other observational instruments, is due to factors *other than* stable differences among teachers. Teacher classroom behavior is just not as stable over classroom sections and time as the research and political communities would like it to be. Reality does not conform to political wishes.

As important as the issue of teacher stability of behavior is, the issue of greater importance is that the correlations of these observations with measures of student gain scores, a value-added approach, were quite low.

In mathematics, across years, these correlations of teacher behavior with value-added growth scores ranged from .12 to .25. Across different sections taught in the same year by the same teacher the correlations ran from .16 to .26. In predicting English language arts scores from year to year, correlations ran between .09 and .12. When predicting value-added scores across sections in the same year, scores ranged from .10 through .24.

These data were just barely adequate enough for the instruments to predict some of the top teachers and the bottom teachers in a distribution of value-added student achievement scores. The predictions were only good at the very extremes, not throughout the midrange of teacher effectiveness. That may actually be prediction enough for policy. So, this approach clearly has some limited promise. But, what is troubling is the fact that even if the correlation between some pool of carefully measured observation instruments and some measures of value added were as high as .3, then those teachers identified as the bottom 25% of effective teachers would be misclassified 62% of the time.

In fact, about a third of the teachers believed to be in the bottom 25% of all teachers would actually be in the top half of teachers who produced value-added scores that were above the average, and 14% would actually be in the top quartile. Furthermore, as the correlation of observation instruments with measures of student value added dips below .3, as was usually the case, then the chances of misclassification of teachers as effective or not effective is much higher. Because of this, you simply cannot ethically and legally use these data for high-stakes decisions about teacher competence, though these data might be used for purposes of formative evaluation of teachers. But even then, the data will often be misleading (see also Rothstein & Mathis, 2013), and teachers are likely to learn to mistrust the data quite rapidly.

This study is a landmark in using classroom observations and value-added achievement test scores in an attempt to identify more and less effective teachers. Two important points may be derived. One is that teacher behavior is not stable across lessons or years, and the other is that teacher effects on achievement are not stable across lessons or years. Something in different sections of a class in the same year, or in different classes over years, is affecting, or conditioning, or moderating, or regulating, or controlling, or subduing, or intensifying teacher classroom behavior and teacher effectiveness as measured by student test performance. The primary candidate for this mysterious factor is a classroom peer effect and its interactions with a host of other exogenous variables. This is so difficult for legislators to understand. They talk of value-added assessments as if they were cornerstones of stability in the sea of subjectivity that characterizes most contemporary teacher evaluations (Gabriel & Lester, 2013). But, the stability they seek is not just missing, it is almost impossible to obtain.

Another recent study hailing teacher effects without regard to magnitude made headlines in *The New York Times* and its lead author, Raj Chetty, recently was awarded a MacArthur Fellowship grant for his genius (Chetty, Friedman, & Rockoff, 2012). But, it may have overstated its case. The authors studied 2.5 million children's records over 20 years and drew the following conclusion:

Teachers have large impacts in all grades from 4 to 8. On average, a one standard deviation improvement in teacher [value added scores based on student achievement] in a single grade, raises earnings by about 1% at age 28. Replacing a teacher whose [value added scores based on student achievement] is in the bottom 5% with an average teacher would increase the present value of students' lifetime income by more than \$250,000 for the average classroom. . . . We conclude that good teachers create substantial economic value.

These sound, of course, like big teacher effects, and I would never want to be quoted as saying that teachers do not have the power to affect the lives of their students. They do. Instead, I want to put those teacher effects in perspective so we don't go astray with our national policies. This particular study received enormous attention, even from President Obama, who has also bought the argument about getting rid of the lowest performing teachers based on their students' test scores. But, a little thought about this particular study reveals that a raise in earnings of 1% at age 28, given the current earnings-per-year of a 28-year-old, will likely be about \$300 per year. At the time of the study, it may have resulted in earnings of only \$182 per year at age 28, but something even more important was not emphasized at the time the study. That was the fact that the raise in income for 28-year-olds was essentially a raise in income of zero by age 30 (Adler, 2013).

A rise in a classrooms' earnings of \$250,000 at the end of their lifetime of work, because we fired the bottom 5% of teachers in our sample, results in about \$10,000 extra income *per lifetime* for each student in a classroom of 25 students who worked for approximately 40 years. A median income earner will earn about \$2,200,000 for a lifetime of work, but with a "better" rather than a "worse" teacher, as identified from an unreliable value-added measure, that worker will earn \$10,000 more over their working

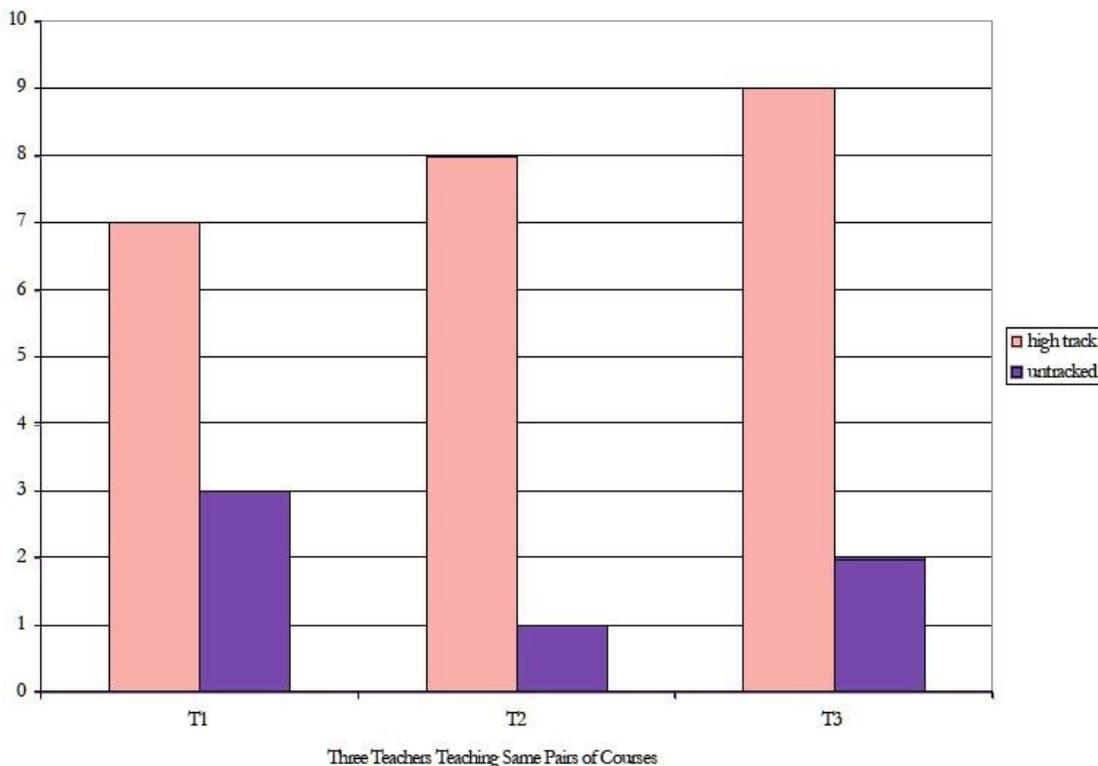
lifespan. That is a total of \$250 more per year worked! This does not sound like a big teacher effect to me. Surely that was a mistaken inference by those who cited the study. Studies like this one, apparently so supportive of value-added assessments and receiving such widespread publicity are likely to be examined carefully. Thus, recent reanalysis of the data suggested that it was not just a small effect that was found, but actually none at all (Adler, 2013).

Another study looked at different versions of value-added models now in use (Newton et al., 2010). Their main finding was this: Even though the different value-added models of teacher effectiveness controlled for student characteristics, it was still found that English teachers' rankings on effectiveness were negatively correlated with the proportion of students in their classes who were English Language Learners, free lunch recipients, or Hispanic. And, the measures of teacher effectiveness were positively correlated with the proportion of students who were Asian or had highly educated parents. Remember, these significant correlations were found *after* student demographics were accounted for by traditional statistical methods. The statistical controls didn't control enough! The researchers also found that English teachers were ranked higher in effectiveness when they had more girls in their classes, and math teachers were ranked higher in effectiveness when they had a higher proportion of students who had better math preparation in previous years. Again, this was *after* controlling for these variables.

Those authors said what I also think to be the case, namely, that the presence of significant correlations between teacher effectiveness rankings and the demographic composition of classrooms signals the compositional or contextual effects that many other scholars have identified as powerful influences on school and classroom achievement (eg., Raudenbush & Bryk, 2002; Willms, 2006). Individual student achievement is affected not only by their own individual background characteristics, a well-established finding, but also by the characteristics of other students in the class. In the development of all current value-added systems, we note a failure to adequately account for many of the possible classroom composition effects. But even if a researcher could specify them all, controlling for them seems difficult to do. And even if that were done, the researcher probably wouldn't have any power to detect the interactions of teacher characteristics and student characteristics. All of these problems, blithely ignored by value-added advocates, lead to misunderstandings about the nature and magnitude of the teacher effect on student learning.

Apparently, teaching a class in which most students come from highly educated families yields higher value-added scores because each individual student's learning is boosted by the presence of other well-supported and highly motivated students. We confuse that with a pure teacher effect. Similarly, a teacher assigned a class in which most students come from poorly educated families, or teaching a class with a larger percentage of boys, or teaching one with more English Language Learners might produce lower value-added scores because each individual student's learning is affected by the presence of other students with difficulties adjusting to school. This is not a simple teacher effect either. These authors cautioned against making any high-stakes decisions about teachers' ability because of the variability they found in classifying teachers as "better" or "worse" teachers in producing student achievement. Figure 1 illustrates this. It provides the decile ratings for the amount of value added by three teachers who taught the same course to two different sections in the same year.

Figure 1. Value-added scores, in deciles, for three teachers teaching the same course, in the same year, to two different classes



Notes. From Newton et al. (2010)

Clearly, *who* these teachers taught and the kinds of classes they were assigned affected their ratings of effectiveness. Also studied by these researchers were the misclassifications that would occur if teachers were ranked by decile of their value added across two sections of the same subject matter, in the same year, or across two years of teaching. The misclassification rates, as seen in Table 1 are so high as to make the present value-added systems immoral to use for any high-stakes decisions about teaching.

Table 1. Changes in Teachers' Decile Rank From a Value-Added Measure of Effectiveness Either for the Same Course Taught to Two Sections During the Same Year or for the Same Course Taught in Two Consecutive Years

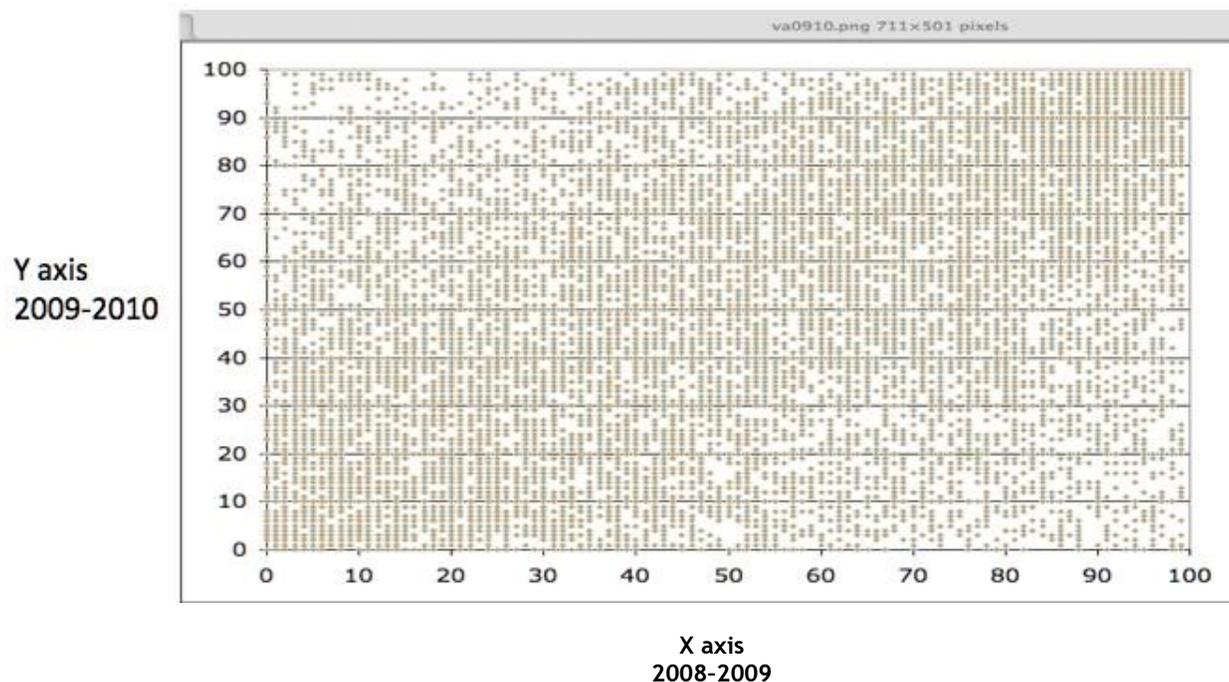
	The change in teacher ranking in effectiveness was one or more Deciles	The change in teacher ranking in effectiveness was two or more Deciles	The change in teacher ranking in effectiveness was three or more Deciles
Changes in teachers rankings across two courses taught in the same year	85%- 100%	54%- 92%	39%- 54%
Changes in teachers rankings across two consecutive years teaching the same type of course	74%- 93%	45%- 63%	19%- 41%

Notes. From Newton et al. (2010)

Although these data are not supportive of value-added approaches to evaluation, the field is still in flux. More recent data (Kersting et al., 2013) suggested that value-added scores over a few years are more stable than others have found them to be. These same researchers also assessed different value-added models and found two expected and two unexpected findings. The expected findings were, first, that class size matters a lot. You really don't get stable value-added scores from small classes, and 25 or fewer is likely to be regarded as small. Second, the value-added scores of inexperienced teachers grew over time, suggesting that this method of evaluating new teachers is unfair. The most interesting of their conclusions, however, was first that the type of value-added model chosen had big effects on the estimates of a teacher's value-added score. This led to quite different conclusions about who is and who is not effective under different models of value-added assessment, as also discovered in the study by Newton et al. (2010). We seem not to know enough yet to make wise choices about which models to use to bolster the stability of the scores we intend to use for judging teacher competency. Finally, and most important for policy, also in line with other research (Goldhaber & Hansen, 2010; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004), the teachers who appeared to be high in value-added scores were considerably more stable in their year-to-year effects than the teachers who appeared to be low in their value-added effects. Thus, the reliable identification of "poor" teachers is considerably harder than the reliable identification of "good" teachers. This is a problem for advocates of value-added assessments who are often more desirous of finding the nation's poor-performing teachers and firing them, than they are of finding the nation's best performing teachers and rewarding them.

Another study makes the point on stability (or instability) quite vividly (Rubenstein, 2012). In New York City, they attempted to create a sophisticated system for determining the value added by a teacher. Students' scores from the previous year are put into an equation with the following variables used to moderate those scores: a poverty index, the number of limited-English-speaking immigrants, the number of students with disabilities, the number of students with student suspensions, the number of students who repeated a grade, the number of students who attended summer school, class size, newcomers to the school, and demographic information such as race and gender. Such a lengthy list of variables should catch most of the variables that are ordinarily exogenous, yet affect value-added scores. Presented in Figure 2 is the scatter plot for value-added scores for 13,000 New York teachers from the 2008-2009 and 2009-2010 assessments. (Rubenstein, 2012). The stability of value-added scores from year to year is still poor.

Figure 2. Scatter plot of value-added scores for New York City teachers for two consecutive years



Note: Zero indicates no gains compared to classes similar to a teachers' own class, and 99 indicates maximum gains compared to classes similar to a teachers' own class. The correlation between scores over one year is .35 (Rubenstein, 2012).

The reason for this instability seems simple enough: There are many exogenous variables out there, and the psychometricians simply could not analyze all the interactions associated with the set of variables that they did measure. Rubenstein (2012) raised an additional issue about the validity of these assessments. He reran the data for just the 707 teachers who were first-year teachers in 2008-2009, expecting that their value-added scores would go up in their second year of teaching. Most teachers believe that they learn a lot during their first year of teaching and, therefore, believe that they did much better in their second year of teaching. The data from Kersting et al. (2013) confirmed that. But, the data from New York didn't show that effect. About half the teachers went up in value-added scores and half went down, raising an issue of whether the tests are valid, given the expectations that a different outcome would be found (see also Rothstein, 2010, for perplexing patterns associated with value-added scores).

On the other hand, because the correlation of these value-added scores was not out of line with many validity coefficients, these data suggested that the chances of being a good or poor teacher three, four, or five years in a row are quite low. Thus, it is argued that if a teacher were in the bottom or top decile a few years in a row, more validity could be claimed for the argument that "good" and "bad" teachers had been identified, at least when student test score growth is the criterion for those designations. This was the argument made by Kersting et al. (2013) from even higher year-to-year correlations than were present in the New York data. But it is not clear that this will actually work out in practice, even when data is pooled across a few years. From the thousands of teachers in the New York value-added data set, Fuller (2012) estimated that for those teachers who were in the top 20% in 2005-2006, a group called "irreplaceable" at the time, only 14 teachers of mathematics and only 5 teachers of English language arts were in the top 20% every year through to 2009-2010. So, empirically, not theoretically, out of the thousands of teachers for whom value-added scores were computed, only a total of 19 were found to be "wonderful" teachers every year for four years. This is not reassuring.

Moreover, relying on the accumulation of value-added data over years does not free teacher evaluators from trying to determine the circumstances in which the teaching took place, particularly if high-stakes decisions are to be made. As demonstrated by Newton et al. (2010), even after accounting for many of the likely variables affecting value-added scores, those scores were still affected by many of those same variables. Those inadequately controlled for variables, along with the many exogenous, unaccounted for variables, make high-stakes decisions about a teacher's skills problematic, even were they to show some stability in scores over a few years.

An instructive example of the instability of scores in value-added systems and the apparent worship of the numbers that emerge from this system without regard for circumstance, came from Houston, Texas (Amrein-Beardsley & Collins, 2012). The value-added record of Ms. J, with 10 years of experience in the Houston Independent School District, is provided in Table 2. Houston used the Education Value-Added Assessment System (EVAAS®), a proprietary system designed by William Sanders and owned by SAS®, the large statistical software company. Proprietary means that independent scholars have no way of checking how their system adjusts for the kinds of variables discussed in this paper. SAS will not let anyone check their statistical models, though states and districts are contracting for their services as if their statistical models are approved by the social science community. Users of EVAAS have to take the word of SAS that it controls for all the important factors, and that the scores obtained are reliable and valid for use in high-stakes decisions such as firing teachers or giving them bonuses. This is not sensible. Nevertheless, Houston and other school districts use this program to determine a teachers' value-added score. In the table provided, scores between -1 and +1 indicate that the teachers' value-added score was within plus or minus one standard error of the scores obtained by teachers like them,

teachers with the same kinds of students that Ms. J had taught. Ms. J's pattern of success and failure, as determined from value-added assessments, is obviously variable. Very variable!

Table 2. Four Years of Value-Added Scores for Ms. J of Houston, Texas

	2006-2007	2007-2008	2008-2009	2009-2010
	Grade 5	Grade 4	Grade 3	Grade 3
Math	-2.03	+.68	+.16	+3.46
Reading	-1.15	-.96	+2.03	+1.81
Language Arts	+1.12	-.49	-1.77	-.20
Science	+2.37	-3.45	-----	-----
Social Studies	+.91	-2.39	-----	-----

Notes. Negative scores indicate less value added than teachers of similar classes; positive scores indicate more value added than teachers of similar classes. Scores between -1.00 and +1.00 are close to the mean value-added scores of teachers with similar classes. Scores beyond those values are considered significant (Amrein-Beardsley & Collins, 2012).

Note also the lack of any correlation of language arts and reading value-added scores. This too suggests a validity problem. But the bigger lesson to be learned from this wobbly pattern of effectiveness, as measured using the EVAAS, is that it yields unreliable estimates about a teacher's skills from one year to the next. Despite assurances by the EVAAS developers and the SAS company that their measures of teacher effectiveness are stable, we see that Ms. J's scores were above average 8 times, significantly so 5 of those times. Her scores were below average 8 times, significantly so 5 times. As Amrein-Beardsley and Collins (2012) noted: If you had flipped a coin and decided heads was a good teacher and tails was a bad one, you would have been right as often as this system revealed Ms. J to be. Other information obtained was that Ms. J's supervisor rated her extremely high in 4 of the 5 years for which there were data, and she received bonuses totaling \$7,800 for her excellence in 3 of those 5 years. Ms. J was also given a "Teacher of the Month" award in 2010, and a "Teacher of the Year" award in 2008, both awarded by her colleagues. Ms. J was, however, fired for incompetence after her fifth year.

Ms. K is another example of this apparently unsound Houston policy. The data on Ms. K covers 4 years. She was a new teacher and her value-added scores for reading and mathematics were positive in half those years, and negative in half those years, showing no particular pattern. Yet, in each of those years, she received bonuses that totaled about \$6500. Clearly the district was satisfied with her performance. And clearly, if they judged her performance satisfactory in years when her value-added score was poor, then there is a criterion validity problem. The value-added assessment of competence and the supervisors' assessment of competence were not measuring the same things, though they were intended to do that.

Ms. K's reading and language arts scores went down in 2009-2010, and her social studies score was also considered to be quite low. Nevertheless, the district apparently had reason to think she was quite competent since they awarded her \$1,250 as a bonus. But shortly after that, in light of her value-added scores, she was fired for incompetence. Not counted in the interpretation of her scores in year 4 of her career was the fact that an inordinate number of English Language Learners were transitioned into her classroom. These were children she was not trained to teach. But because of the trust we place in the numbers, rather than the opinions of people or the context within which the numbers were developed, Ms. K was considered a failure. As one team of researchers put it (Graue, Delaney, & Karch, 2013), the student outcomes we value in value-added instruction are made possible through classroom interaction and processes that are supported or undermined by contextual factors. Context really matters. Only multidimensional approaches to teacher evaluation can begin to capture the complexity of education in the classrooms of our nation's schools.

One of the teachers interviewed in the Amrein-Beardsley and Collins (2012) study put it this way: "I do what I do every year. I teach the way I teach every year. My first year got me a pat on the back. My second year got me a kick in the backside. And for three years [now] my scores were off the charts. I got a huge bonus, and now I'm in the top quartile of all the English teachers. What did I do differently? I have no clue."

But the teacher I described at the start of this paper and I do have a clue. Characteristics of the students we teach interact with our teaching skills and moderate our effectiveness as teachers. The instability in teacher behavior and teacher effectiveness that we so commonly observe across so many teachers suggests that teachers have different skills for different kinds of students and classes, resulting in different estimates of a teachers' effectiveness every year. These peer and compositional effects are pervasive and universal, as noted in the next section of this paper.

PEER OR COMPOSITIONAL EFFECTS

Peer or compositional effects condition, that is, interact with, measures of teaching effectiveness. In Chile, for example, McEwan (2003) stated unequivocally that the classroom mean of mothers' education is the most important peer determinant of achievement in a classroom. He said a one standard deviation increase in mothers' mean educational level led to over a one-quarter standard deviation rise in student performance in Spanish language arts. If inadequately controlled for or ignored, the composition of the class would mislead us about the magnitude of the teacher effect.

The mean of a fathers' education also had significant positive effects on achievement, but these were smaller in magnitude. In addition, an increase in percentage of indigenous students in a classroom tended to lower achievement, all else being equal, but again, the magnitude of the effect was rather small. Still, a one standard deviation increase in indigenous students in a classroom would be equivalent to a 9% increase in their population and this, in turn, leads to a decline in Spanish achievement of 3% of a

standard deviation. Though small, note how this might work. If you had 10 indigenous students in your class, and that number went up to 13, an increase of 30%, you can expect a loss in student achievement of nearly 10% of a standard deviation. That works out to about one month's worth of achievement in the ordinary test metrics that we use in the United States to determine student achievement and teacher effectiveness. If the teacher does not have the interpersonal skills to work with these children, on average, that teacher will be judged less effective than if the number of indigenous children in the classroom was less.

In another study by Lavy et al. (2009), this time in England, results showed that a large fraction of low-achieving peers at school, as identified by students in the bottom 5% of the national ability distribution, negatively and significantly affected the cognitive performance of all the other students in the school. These researchers found that it is the very bottom 5% of the students that negatively affect achievement, and not low-achieving students with scores that are higher, even though those scores might still be low in the national distribution. They also found that girls, particularly those in the lower half of the overall ability distribution, benefited significantly from interactions with very bright peers. Boys, however, showed an opposite effect. When there was a higher proportion of bright peers, the scores of the average boy and the below-average boy dropped.

Although the negative gender effect for boys is smaller than the positive gender effect shown with girls, both are statistically significant and in line with the extant literature. Lavy et al. (2009) said, "Classroom composition affects what and how much boys and girls learn *independent of teacher effects*." They estimated that if a student moves from a primary school where 20% of the children are in the lowest levels of the achievement distribution, to a school where no students are in the lowest levels of the achievement distribution, that student would experience an improvement in test scores of about 2.4 percentiles, which amounts to around 9% of a standard deviation in the national primary test score distribution. They further calculated that a 10% decrease in very low-scoring peers at a school is quite close to the effect of one extra hour of weekly instruction time for all students. Clearly, compositional effects are not trivial, though not always noticed. Thus, variables such as these had better be accounted for in assessments of teacher effectiveness. It appears extremely problematical to fire or reward teachers without fully accounting for compositional effects.

Lavy et al. (2009) also estimated the effect size of peer effects based on different achievement distributions of peers at a school. Their results were in line with those estimated in Ammermueller and Pischke (2006) across classes, within schools, in six different European countries, using the Progress in International Reading Literacy Study (PIRLS) data. In this study, "books in the home" was used as a proxy for social class, social capital, and parents' education for each fourth grade student. Their results, averaged across countries, predicted that a one standard deviation change in the number of books in a child's home leads to a change of about 17% of a standard deviation in reading test score. That is, when books in the home are used to estimate social class, social class effects on achievement at a school are quite large. If not controlled for, or controlled for inadequately, social class composition effects will often be confounded with teacher and school effects on achievement.

In Denmark, Schneeweis and Winter-Ebmer (2005) used Program for International Student Assessment (PISA) data for estimating student social class with a composite of home talk about political or social issues; home discussions of books, films, or TV programs; and listening to classical music at home. They reported that if you moved the average student to a new peer group one standard deviation higher on this index, you can expect the student's reading achievement to rise by about 5 points on the PISA scale. These researchers also estimated that if you are low in the index of cultural and social variables, and find yourself in a class where there is an increase of a standard deviation in the quality of the peers you go to school with, you will show an increase of about 10 PISA reading points. Thus, the peer effect is twice as high for low family background students. This seems to be a consensus opinion from a number of quite disparate researchers.

Additionally, an increase in peer quality will raise the bottom students' math and science scores by 7 points. So, if a nation wants to have its teachers show better effects on students, it might try integrating neighborhoods by social class. That seems to be a policy with more research behind it than one that advocates identifying and firing the least able teachers. Another way to think about this is to say that the U.S. PISA scores, which are often around the Organization for Economic Co-operation and Development (OECD) average, could be moved to above the OECD average, not by getting better teachers, as many in Congress demand, but by integrating schools by social class, a thought that terrifies many of our politicians and citizens, liberals and conservative alike. Despite the terror of our country's men and woman in facing this possibility, we have uncovered a very stable finding in the literature about composition effects and achievement.

In related research, Schindler-Rangvid (2003) in Denmark and Levin (2001) in the Netherlands have found that low-ability students can substantially benefit from a high-quality peer group, whereas high-ability students are not influenced much when mixed with low-ability students. This is the same finding reported by Oakes (2005) when studying tracking effects. In general, compositional effects of classrooms and schools are powerful in the life of the least academically talented, lowest social class students. And these are just the ones we worry most about and whose poor achievement in our schools started the American (and now worldwide) obsession with describing effective teachers. But, as pointed out in this paper, a persuasive argument can be made that it is not effective teachers who carry as much weight in the lives of our students as do scholastically and socially effective classmates.

In the United States, Hoxby (2000) found, like everyone else cited thus far, that students are affected by the achievement level of their peers. Using the TAKS (Texas Assessment of Knowledge and Skills), in grades 3-6, she found that a 1-point change in reading scores in the peer group score in the class, a tiny 1-point score increase on the TAKS, raises each student's score between .15 and .40 points, depending on which model she used to analyze the data.

Let us take a middle-level estimate of the peer effect and call it about .25. Any student who gets placed in a class that scores 10 points higher on the TAKS than the last class the student was in, is estimated to have his or her personal score go up 2.5 points. In

value-added models of teacher effectiveness, will the genuine growth in this student's score be attributed to teacher effectiveness or classroom compositional effects? The problem of inferring causality is always the problem to be faced.

It is precisely this underlying confusion about causality that motivates assertive parents to pursue districts, schools, and classrooms with higher mean scores. They know, without research informing them, that their own children's scores will go up because of compositional effects, which they typically confuse with teacher effects. They crave better teachers for their children, as any parent might, but they think those teachers are in better schools and neighborhoods with social and cultural indices that are higher than where they presently live or where they school their children now. These upwardly mobile parents, the tiger moms out there (Chua, 2011), have a correct understanding of the effects, but their understanding of the cause of those effects may not be accurate. Increased student achievement may occur not as a function of better teachers, as much it is caused by the composition of the classrooms and schools in which their children find themselves. Parents, like almost everyone else in the United States, may be overestimating the teacher effects and underestimating the compositional or peer effects in the search for effective classrooms and schools.

Hoxby (2000) also found that the more females in a class, the higher the scores of both males and females. Let us assume a class size in the United States of 30. Let us also assume 15 of the students are female. If that number went from 15 to 18, female scores in the reading tests would rise about .7 of a point and male scores would rise by almost 1 full point. An all-female class might well score about one-fifth of a standard deviation higher than a class with a normal division of males and females. In math, with the same change in the ratio of females to males, we can expect score increases for students of between .8 point and 1.5 points on the test, depending on the grade level examined. Though the compositional effects per student are small, these are not trivial, and may result in advantages in value-added scores for teachers in classes with more females. The possibility of receiving a bonus or being dismissed should not hinge on such things as the composition of your classroom! But, teachers already have figured this out. In Houston, a poster child for what can be done wrong in the search for effective teaching, teachers seek students for their classes who are midlevel in achievement, or well-behaved middle class low-achieving kids, because they are the "money kids" (Collins, 2012). They are the ones that teachers say will probably gain the most in a year, and that can help to get a big bonus. To be avoided like the plague, say these teachers, are English Language Learners and gifted students because they don't show growth on the tests and, thus, you could get fired or receive no bonus if you teach those kinds of students. Already, principals are helping teachers they like, and hurting teachers they do not like, by manipulating class rosters at the start of the school year. The principal's favorite teachers get easier to teach students and disliked teachers get the children who are expected to show the least growth on the test, allowing the principal to have those teachers removed. In Houston, we have reliable evidence of this kind of collusion to game the value-added system in order to receive bonuses and to avoid being fired (Collins, 2012).

Hoxby (2000) also found powerful and stable compositional effects when the race of students was examined. As the percentage of Asians grew, it raised mean achievement scores for all others in the class. As the American Indian population grew in number, classroom mean achievement went down. These effects were net of other effects, such as social class. Even more interesting was that intrarace scores were more affected by composition effects than scores between race. This has implications for countries like ours that have many schools that are highly segregated by race. Hoxby reported that for every 10% rise in the percentage of a class that is African American, African American students' reading scores fell a quarter point, Hispanic reading scores fell almost one point, and Anglo scores fell about 2/3 of a point. In mathematics, for every 10% rise in African American students in a class, African American mathematics scores dropped by almost 2 full points, Hispanic scores by almost one point, and Anglo student scores dropped almost a half point. Hoxby noted that when a class's mean achievement drops 1 point, *just 1 little point*, the effects on the various races in the two subject areas change about 1-7% on the achievement test. In the United States, about 40% of African American children and 40% of Hispanic children go to schools that have over 90% minority students (Orfield & Lee, 2005). These data suggest that in such schools a good part of the teachers' and schools' effects on achievement are compositional effects, often attributable to segregation by housing, though teachers are blamed for the poor performance of many of these children. But even when working with poor children, it is not just composition of the class that affects achievement: the social capital provided by the community from which a child comes influences the child's school achievement (Goddard, 2003). So, it is at least poverty x race x social capital, in complex interaction, that affects student achievement. Unless the interactions of these peer and compositional effects with teacher characteristics are fully controlled for in value-added models, teacher effects and compositional effects of classrooms, schools, and neighborhoods are hopelessly confounded.

In Australia, the work of Perry and McConney (2010) illuminated the magnitude of these compositional effects. They studied the PISA nations, and she and the Organization for Economic Co-operation and Development (OECD, 2004) reached the same conclusion, one that is at odds with the logic of value-added approaches for assessing teachers. She and the OECD believed that school social and cultural factors far outweigh an individual student's own social and cultural background, perhaps by a factor of 2. While in the United States, Rumberger and Palardy (2005) found that the effects on achievement of individual student social and cultural factors were about equal to the school's social and cultural effects on achievement. This set of findings demonstrates the power of compositional effects in shaping individual student achievement, independent of the teacher, and even independent of the skills and abilities brought to the class by the student. The data of Perry and McConney (2010) described the magnitude of this effect. If a poor Australian child is in a school that serves the poor, the child's PISA science score is 455, about half a standard deviation below international averages. But, that same poor child in a school that serves the wealthy would have a score of 512, a score at about the international average. Is that due to better teachers? The compositional effect? Interactions between the two? Or is it due to other unspecified factors? Similarly, a wealthy child in a school that serves the wealthy would have a score of 607 on the PISA science test, remarkably high, and about a full standard deviation above the international average. But, that same child in a school that serves the poor would have a score of 555, half a standard deviation lower than the international average. Is that due to teacher effects, compositional effects, interactions between the two, or other unspecified factors? These are important questions to answer. Currently, the evidence suggests the greater power of compositional effects, and the lesser power of teacher effects on classroom mean achievement.

Willms (2006) had equally convincing data. He asked what might be the average difference, across many countries, for a child whose family is in the middle of the social class distribution if the child attended one of the better performing schools in his or her country, rather than one of the poorer performing schools in the county. Using data from the Progress in International Reading Literacy Study (PIRLS), by grade 4 that child is estimated to be about one and a half grade levels ahead of where he or she would ordinarily be. This is growth, value added by composition effect rather than by teacher effect, and the two are quite easily confused.

If the data from the Program for International Student Assessment (PISA) were used in the same way, and this child from the middle social class attended one of the high-performing schools instead of one of the low-performing schools, Willms estimated that the difference in that child's test scores attributable primarily to composition effects would amount to an astonishing four grade levels!

One study at the preschool level described the magnitude of these compositional effects particularly well. Henry and Rickman (2007, p. 100) said, "The ability level of the peers in a child's classroom has direct effects on the child's cognitive skills, pre-reading skills and expressive language skills after controlling for preschool resources, family characteristics, and the child's skills at the beginning of preschool." These researchers concluded that studies that did not take into account peer effects on achievement probably overestimate the effects of school-level variables, such as class size, discipline policy, curriculum, leadership, and even teacher quality.

This preschool study provided a clear statement about this issue: Measures of teacher effects are probably inflated when we look at the outcomes of schooling without having really good measures of peer or compositional effects, or without being able to control for them, or without understanding the myriad exogenous variables that are interacting with life in classrooms.

CONCLUSION

Classroom and school compositional effects may be among the most powerful of the factors affecting student achievement, and the interactions of teacher, school, and family characteristics seem often to be unanalyzed; thus, they are often unknown and, perhaps, even unknowable. These separate and interactive effects are powerful moderators of student classroom achievement and are often erroneously confused with a teacher effect, leading at times to overestimation of those effects. As Lee J. Cronbach once so eloquently argued, when you begin to think seriously about interactions, you necessarily enter a hall of mirrors; everything is changed, everything is made smaller or bigger, fatter or thinner, and seems to appear closer or farther away. Interactions require modifications of the findings you thought were firmly established. The images you have about how things work are distorted, like the images in the mirrors. Complexity abounds. Added to this hall of mirrors that is associated with interactions are two additional problems that also make our interpretations of data problematic. One is the problem of unreliability or marginal reliability of the measures used both for observing in the classroom and for determining a value-added score. The other is the problem of choosing a defensible model for determining value-added scores. The problems associated with unreliability of our measures and model specification may ultimately be solvable, though I am not sanguine.

On the other hand, we may never be able to avoid entering that hall of mirrors when we study teachers' effects on students. That is because compositional effects in neighborhoods and in schools operate on, condition, and interact with peer effects in classrooms, and this has been demonstrated to occur over years and across countries, grade levels, and subject areas. Teachers, as might be expected of a profession that deals with such complexity, may not have the instructional or the interpersonal skills to optimize achievement given all the variations in students and student composition that they encounter from hour to hour, day to day, class to class, and year to year. On more careful consideration, variation in instructional behavior, given teachers' perceptions of shifting instructional contexts, is quite appropriate, and this leads naturally to variations in instructional effects. The miracle that is too often ignored is that in the face of all this complexity, so many of our teachers do a thoroughly adequate job of teaching, year in and year out.

Although hard to ferret out in their "pure" form as an independent main effect, teacher effects on student achievement exist, and they are likely to be strong enough for us all to worry about who teaches our children and what their training has been. There does seem to be a small percentage of teachers who show consistency no matter what classroom and school compositions they deal with. Those few teachers who have strong and consistent positive effects on student outcomes, we should learn from and reward. And, those few teachers who have strong negative effects on student outcomes need to be helped or removed from classrooms. But the fundamental message from the research is that the percentage of such year-to-year, class-to-class, and school-to-school effective and ineffective teachers appears to be much smaller than is thought to be the case. When the class is the unit of analysis, and student growth is the measure we use to judge teacher effectiveness, what we find is a great deal of adequacy, competency, and adeptness by teachers in response to the complexity of the classroom. And, we see much less of the extraordinarily great and horribly bad teachers of political and media myth. The thousands of welfare queens that Ronald Reagan railed against and the thousands of disability cheats that have contemporary Republicans in such a snit may be like the thousands of terrible teachers in our public schools—more hype than it is reality.

The continuous promotion by politicians and the press of the stand-alone, heroic, effective teacher who accomplishes miracles, or the opposite, the lazy, uncaring nine-to-fiver who cannot teach well at all, may also be another way to avoid paying for the social programs that could actually reduce the achievement gaps that we have in the United States. What we have learned about the powerful effects of peer and composition effects on school achievement suggests that if we had policies for placing public housing in sections of cities that are middle class; policies that push for the integration of school systems by social class; policies that eliminate tracking within schools; policies for making assignments to classrooms and schools that never allow for too many boys,

too many minorities, too many special education students, or too many English Language Learners; policies for matching teachers to the kinds of students they stand the best chance of succeeding with; policies to provide high-quality early childhood programs and high-quality culturally relevant summer school; and policies for increasing neighborhood social capital and a sense of communal efficacy, then we would stand a very good chance of having many of America's purportedly bad teachers suddenly become much better teachers!

References

- Adler, M. (2013). Findings vs. interpretation in "the long-term impacts of teachers" by Chetty et al. *Education Policy Analysis Archives*, 21(10). Retrieved from <http://epaa.asu.edu/ojs/article/view/1264>
- Ammermueller, A., & Pischke, J.-S. (2006, April). Peer effects in European primary schools: Evidence from PIRLS (Discussion Paper No. 2077). Bonn, Germany: Institute for the Study of Labor (IZA). Retrieved from <ftp://repec.iza.org/RePEc/Discussionpaper/dp2077.pdf>
- Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, 20(12). Retrieved from <http://epaa.asu.edu/ojs/article/view/1096>
- Attar, B. K., Guerra, N. G., & Tolan, P. H. (1994). Neighborhood disadvantage, stressful life events, and adjustment in urban elementary-school children. *Journal of Clinical Child Psychology*, 23, 391-400.
- Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, A., McCartney, K., Owen, M. T., The NICHD Early Child Care Research Network. (2007). Are there long-term effects of early child care? *Child Development*, 78(2), 681-701.
- Berliner, D. C. (2006). Our impoverished view of educational reform. *Teachers College Record*, 108(6), 949-995.
- Berliner, D. C. (2009). *Poverty and potential: Out-of-school factors and school success*. Boulder, CO: National Center for Educational Policy. Retrieved from <http://nepc.colorado.edu/publication/poverty-and-potential>
- Berliner, D. C. (2012). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record*, 115(12). Retrieved from <http://www.tcrecord.org/PrintContent.asp?ContentID=16889>
- Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Policy and practice brief). Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practitioner_Brief.pdf
- Brophy, J. E. (1979, September). *Teacher behavior and its effects* (Occasional Paper No. 35). East Lansing, MI: Institute for Research on Teaching, Michigan State University. Retrieved from <http://education.msu.edu/irt/PDFs/OccasionalPapers/op025.pdf>
- Chua, A. (2011). *Battle hymn of the tiger mother*. New York, NY: Penguin Books.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011, December). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (Working Paper 17699). Cambridge, MA: National Bureau Of Economic Research. Retrieved from <http://www.nber.org/papers/w17699>
- Collins, C. (2012). Houston, we have a problem: Studying the SAS Education Value-Added Assessment System (EVAAS) from teachers' perspectives in the Houston Independent School District (HISD). (Doctoral dissertation). Mary Lou Fulton Teachers College, Arizona State University.
- Cuban, L. (2012, March 20). The puzzle of student responsibility for learning [Web log post]. Retrieved from <http://larrycuban.wordpress.com/2012/03/20/the-puzzle-of-student-responsibility-for-learning/>
- Curry School of Education. (n.d.). *Measuring and improving teacher-student interactions in PK-12 settings to enhance students' learning*. Charlottesville, VA: University of Virginia, Center for Advanced Study of Teaching and Learning. Retrieved from http://curry.virginia.edu/uploads/resourceLibrary/CLASS-MTP_PK-12_brief.pdf
- Danielson, C. (2012). Teacher evaluation: What's fair? What's effective? Observing classroom practice. *Educational Leadership*, 70(3), 32-37.
- Fuller, B. (2012) On the stability (or not) of being irreplaceable [Web log post]. Retrieved from <http://schoolfinance101.wordpress.com/2012/11/17/on-the-stability-or-not-of-being-irreplaceable/>
- Gabriel, R., & Lester, J. N. (2013). Sentinels guarding the grail: Value-added measurement and the quest for education reform. *Education Policy Analysis Archives*, 21(9). Retrieved from <http://epaa.asu.edu/ojs/article/view/1165>
- Goddard, R. D. (2003). Relational networks, social trust, and norms: A social capital perspective on students' chances of academic success. *Educational Evaluation and Policy Analysis*, 25(1), 59-74.

- Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, 2(1). Retrieved from http://educationnext.org/files/ednext20021_50.pdf
- Goldhaber, D., & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance* (CEDR Working Paper 2010-3). Seattle, WA. Retrieved from <http://www.cedr.us/publications.html>
- Graue, M. E., Delaney, K. K., & Karch, A. S. (2013). Ecologies of education quality. *Education Policy Analysis Archives*, 21(8). Retrieved from <http://epaa.asu.edu/ojs/article/view/1163>
- Gysbers, N. C. (2008). Evaluating school guidance and counseling programs: Past, present, and future. In H. L. K. Coleman, & C. Yeh (Eds.), *Handbook of school counseling* (p. 739-749). New York, NY: Taylor & Francis Group.
- Henry, G. T., & Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review*, 26, 100-112.
- Hoxby, C. (2000). *Peer effects in the classroom: Learning from gender and race variation* (NBER Working Paper 7867). Cambridge, MA: National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w7867>
- International Association of School Librarianship. (2008, April). School libraries make a difference to student achievement [Online article]. Chicago, IL: Author. Retrieved from <http://www.iasl-online.org/advocacy/make-a-difference.html>
- Kersting, N. B., Chen, M., & Stigler, J. W. (2013). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives*, 21(7). Retrieved from <http://epaa.asu.edu/ojs/article/view/1167>
- Lavy, V., Silva, O., & Weinhardt, F. (2009, December). *The good, the bad and the average: Evidence on the scale and nature of ability peer effects in schools* (NBER Working Paper 15600). Cambridge, MA: National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w15600>
- Levin, J. (2001). For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. *Empirical Economics*, 26, 221-246.
- Little, P. M. D., Wimer, C., & Weiss, H. B. (2008, February). After school programs in the 21st century: Their potential and what it takes to achieve it (Brief No. 10). *Issues and Opportunities in Out-of-School Time Evaluation*. Cambridge, MA: Harvard Family Research Project. Retrieved from
- McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101.
- McEwan, P. (2003). Peer effects on student achievement: Evidence from Chile. *Economics of Education Review*, 22, 131-141.
- National Association of School Nurses. (2012). Five ways a school nurse benefits the school [Online flier]. Silver Spring, MD: Author. Retrieved from www.nasn.org/portals/0/about/snd2012_save.pdf
- Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, 18(23). Retrieved from <http://epaa.asu.edu/ojs/article/view/810>
- Oakes, J. (2005). *Keeping track: How schools structure inequality* (2nd ed.). New Haven, CN: Yale University Press.
- OECD (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: Author.
- Orfield, G., & Lee, C. (2005). *Why segregation matters: Poverty and educational inequality*. Cambridge MA: Harvard University, Harvard Civil Rights Project.
- Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, 112(4), 1132-1162. Retrieved from <http://www.tcrecord.org/content.asp?contentid=15662>
- Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Ronfeldt, M., Loeb, S., & Wyckoff, J. (2012, October). How teacher turnover harms student achievement. *American Educational Research Journal*. Advance online publication. doi 3102/0002831212463813
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay and student achievement. *Quarterly Journal of Economics*, 125(1), 175-214.

- Rothstein, J., & Mathis, W. J. (2013). Review of two culminating reports from the MET project [Review of the report "Have we identified effective teachers?" by T. J. Kane, D. F. McCaffrey, T. Miller, & D. O. Staiger and review of the report "A composite estimator of effective teaching" by K. Mihaly, D. F. McCaffrey, D. O. Staiger, & J. R. Lockwood]. *American Educational Journal*, 50(1), 4-36. Retrieved from <http://nepc.colorado.edu/thinktank/review-MET-final-2013>
- Rubenstein, G. (2012). Analyzing released NYC value-added data part 1 [Web log post]. Retrieved from <http://garyrubinstein.teachforus.org/2012/02/26/analyzing-released-nyc-value-added-data-part-1/>
- Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record*, 107(9), 1999-2045.
- Schindler-Rangvid, B. (2003) *Educational peer effects. Quantile regression evidence from Denmark with PISA 2000 data* (Doctoral thesis, Aarhus School of Business, Aarhus, Denmark). Retrieved from <http://www.oecd.org/denmark/33684822.pdf>
- Schneeweis, N., & Winter-Ebmer, R. (2005, April). *Peer effects in Austrian schools*. Vienna, Austria: Institut für Höhere Studien.
- Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies*. Princeton, NJ: Princeton University Press.
- Willms, J. D. (2006). Learning divides: Ten policy questions about the performance and equity of schools and schooling systems (UIS Working Paper No. 5). Montreal Canada: UNESCO Institute for Statistics. Retrieved from <http://unesdoc.unesco.org/images/0014/001470/147066e.pdf>

Cite This Article as: *Teachers College Record* Volume 116 Number 1, 2014, p. -
<http://www.tcrecord.org> ID Number: 17293, Date Accessed: 8/21/2015 1:15:57 PM

[Purchase Reprint Rights for this article or review](#)