Journal of Personnel Evaluation in Education 11: 57–67, 1997

Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation

S. PAUL WRIGHT, SANDRA P. HORN AND WILLIAM L. SANDERS
University of Tennessee, Value-Added Research and Assessment Center, 225 Morgan Hall, P.O. Box 1071, Knoxville, Tennessee 37901-1071

Abstract

The Tennessee Value-Added Assessment System (TVAAS) has been designed to use statistical mixed-model methodologies to conduct multivariate, longitudinal analyses of student achievement to make estimates of school, class size, teacher, and other effects. This study examined the relative magnitude of teacher effects on student achievement while simultaneously considering the influences of intraclassroom heterogeneity, student achievement level, and class size on academic growth. The results show that teacher effects are dominant factors affecting student academic gain and that the classroom context variables of heterogeneity among students and class sizes have relatively little influence on academic gain. Thus, a major conclusion is that teachers make a difference. Implications of the findings for teacher evaluation and future research are discussed.

Overview

Over the years, educational researchers have investigated many factors considered to affect student learning. At the heart of this line of inquiry is the core belief that *teachers make a difference*. There are continuing debates about how much the extant teacher-effectiveness literature (e.g., Brophy, 1986; Porter & Brophy, 1988) can be trusted to identify characteristics of effective teachers, and additional debates as well about how such research findings should frame the subsequent development of teacher evaluation systems (e.g., Ellett, 1990; Scriven, 1990; Peterson, Kromrey & Smith, 1990). In addition, there is considerable argument over the logic behind and the extent to which student achievement data should be used as a basis for teacher evaluation (Berk, 1988; Schalock & Schalock, 1993). These debates aside, few attempts have been made to directly measure the influence of individual teachers on the academic progress of large *populations of students* using measurements available from traditional standardized testing programs. Partial confounding of educational (teacher) effects with factors exogenous to schooling influences (see Wang, Haertel & Walberg, 1993 for an explication of these issues) and the nonrandom assignment of students to teachers are two of the reasons most often assumed to be insurmountable obstacles to this type of inquiry.

In criticizing and arguing equity issues in the fair application of teacher evaluation instruments and procedures, teachers have often directed their comments to classroom context characteristics. Key among these has been the issue of the ability level of students and the range in individual differences among students in ability levels. As the argument typically proceeds, teachers who have classes more heterogeneous than homogeneous in ability levels are at a distinct disadvantage in producing effects on student learning and subsequent achievement, particularly as inferred from standardized test scores.

Recently, new processes for estimating the effects of teachers and schools on student academic outcomes free of these traditional objections have been developed. One of these—the Tennessee Value-Added Assessment System (TVAAS), which uses statistical mixed-model methodology to enable a multivariate, longitudinal analysis of student achievement data— has been demonstrated to produce estimates of school and teacher effects that are free of socioeconomic confoundings and do not require direct measures of these concomitant variables (see Sanders & Horn, 1995b, and Sanders, Saxton & Horn, in press, for greater detail). To support TVAAS, a massive database of longitudinally merged student, teacher, school, and school system information has been compiled for the primary purpose of determining system, school, and teacher effects on the academic gains of students. Utilizing this database, the present study attempts to measure the relative magnitude of teacher effects while simultaneously considering the influences of intraclassroom heterogeneity, student achievement level, and class size on academic growth. Among these influences, intraclassroom heterogeneity is of special interest. The magnitude of this variability may be a natural occurrence or can result from intentional grouping of students. Regardless of cause, the evaluation of the influence of intraclassroom, variability on the academic growth of student populations and its interaction with teacher effects is another important research objective of this study.

Methodology

For the purpose of this investigation, results are derived from analyses of a subset of data from the 1994 and 1995 TCAP scores for five subjects (math total, reading total, language total, social studies, and science) and three grades (third, fourth, and fifth). TCAP tests are given each spring to all students in Tennessee in grades two through eight. An important property of these tests is that the scale scores form a single, continuous, equal-interval scale across all grades (CTB/McGraw-Hill, 1990, pp. 4–5), allowing for measurement of student academic progress from year to year. The analyses reported here are based on student academic gain—that is, the student's scale score this year minus that student's scale score last year. Thirty separate analyses were done. Each of the fifteen subject–grade combinations was analyzed separately, and each of these fifteen analyses was carried out on two different sets of school systems in Tennessee. One set consisted of thirty East Tennessee school systems, and the other consisted of twenty-four Middle Tennessee systems. A mixed-model analysis of variance was obtained by fitting the following model[1] to the data:

$$\begin{aligned} Y =& M + S + H + C + H^*C + T(S^*H^*C) + A + A^*S \\ &+ A^*H + A^*C + A^*H^*C + A^*T(S^*H) + E, \end{aligned}$$

where

Y is the student's gain score,
M is an overall mean gain,
S is the school system,
H is heterogeneity-in-achievement (three groups were used),
C is the class size (two groups were used),
*H**C* is the heterogeneity-by-class-size interaction,
T(*S**H**C*) is the teacher, each one nested within a particular combination of system, heterogeneity groups, and class-size group,
A is achievement level (four groups were used),
*A**S* is the achievement-by-system interaction,
*A**H* is the achievement-by-heterogeneity interaction,
*A**C* is the achievement-by-class-size interaction,
*A**H**C* is the achievement-by-heterogeneity-by-class-size interaction,
*A**T*(*S**H**C*) is the achievement-by-teacher interaction,
E is the random ''error'' term.

The *T*(*S**H**C*), *A**T*(*S**H**C*), and *E* terms represent random effects. All the other effects are fixed. The analyses were done with the MIXED procedure in SAS/STAT version 6.09 running on an IBM RS/6000 Model 590 work station at the Value-Added Research and Assessment Center at the University of Tennessee, Knoxville.

The response variable—the educational outcome of the student—was the student's gain score from 1994 to 1995—that is, the student's 1995 scale score on the TCAP minus the student's 1994 scale score. The student's achievement level was defined operationally as the average of the student's 1994 and 1995 scale scores. Classroom heterogeneity in achievement was defined operationally as the standard deviation of the achievement level scores of the students in the class, as defined above. The larger the standard deviation, the more heterogeneous in achievement were the students in the class. For the analysis, classrooms were classified into three groups—low, moderate, and high heterogeneity—using their standard deviation of achievement level. The moderate group contained about half of the classrooms, and the two extreme groups each contained about one-fourth of the classrooms. Students were classified into four achievement level groups of roughly equal size using the achievement level scores described above. Inclusion of an achievement level variable was thought to be particularly important in view of the results of earlier studies indicating that the value of tracking or not tracking depended on the achievement level of the student (Kulik, 1992).

Two class-size groups were used: small (ten to nineteen students) and large (twenty to thirty-two students). Classes of fewer than ten or more than thirty-two students were omitted. There were several reasons for omitting the larger classes. The first was that the database currently does not actually identify the classroom of each student. It does identify the teacher for each student and subject. The reason that only third, fourth, and fifth grades were analyzed is because, in these grades, it is more commonly the case that each student is in a single classroom with a single teacher. Nevertheless, some teachers in the database were shown to have a large number of students, too many to represent a single classroom. Omitting teachers with more than thirty-two students provided a way to avoid treating as one classroom what was in fact several classes taught by the same teacher.

Results

Table 1 through 3 summarize the results for grades three through five, respectively. As an aid for assessing both the statistical significance and the effect sizes of the various effects in the model, z-scores are reported for each effect. For random effects, z-scores were obtained by dividing the estimated variance component for the effect by its estimated standard error. For large samples (such as those in this study), this z-score is approximately distributed as a standard normal variate. For fixed effects, first p-values were obtained

Table 1. z-Values for Analyses of Third-Grade Gains.

Source	*Set*	*Math*	*Reading*	*Language*	*Social Studies*	*Science*
System (*S*)	1	6.12	2.26	4.34	4.03	3.13
	2	4.86	3.55	5.39	5.55	3.92
Heterogeneity (*H*)	1	1.39	0.25	0.61	0.81	0.05
	2	1.54	0.09	1.64	0.61	0.30
Class size (*C*)	1	0.57	0.02	1.45	0.14	1.92
	2	1.03	0.64	0.16	0.97	0.38
H*C	1	0.58	0.49	0.29	0.45	1.83
	2	0.20	0.47	2.21	0.20	0.83
Teacher (*S*H*C*) (*T*)	1	12.48	7.85	11.04	6.09	7.76
	2	13.14	8.69	12.06	8.33	8.88
Achievement level (*A*)	1	17.00	12.65	8.49	10.04	6.76
	2	28.04	20.14	8.96	14.53	8.41
*A*S*	1	2.19	1.88	2.70	2.49	2.19
	2	1.25	5.31	1.46	3.34	3.26
*A*H*	1	2.05	4.64	1.15	4.36	0.53
	2	1.41	0.76	1.29	3.78	4.27
*A*C*	1	1.37	0.53	0.40	0.18	1.53
	2	0.12	0.67	1.14	2.33	1.19
*A*H*C*	1	0.07	0.22	0.32	0.10	0.70
	2	2.05	0.94	0.37	2.12	2.18
*A*T*	1	2.35	4.88	2.02	0.61	1.05
	2	0.73	0.68	1.27	1.69	2.39
N	1	10751	10564	10916	10005	9939
	2	13632	13506	14079	13651	13624

Set: 1 = 30 East Tennessee school systems.
2 = 24 Middle Tennessee school systems.
N = total number of students.

from *F* statistics, then corresponding *z*-scores were calculated from the *p*-values by treating the *p*-values as if they were two-tailed and from a standard normal distribution. This technique of converting *p*-values to *z*-scores is commonly used in meta-analysis to convert results from a variety of tests to a common metric (see, for example, Rosenthal, 1984, p. 65). For reference, the *z*-values correspond to the two-tailed *p*-values of 0.10, 0.05, 0.01, 0.001, and 0.0001 are 1.64, 1.96, 2.58, 3.29, and 3.89, respectively.

It is clear from Tables 1 to 3 that the two most important factors impacting student gain are the teacher and the achievement level for the student. The teacher effect is highly significant in every analysis and has a larger effect size than any other factor in twenty of the thirty analyses. The achievement-level effect is significant in twenty-six of the thirty analyses and has the largest effect size in ten of the thirty analyses. These results are discussed in more detail in the Discussion section below.

The third most important factor overall was the school system. There were significant

Table 2. *z*-Values for Analyses of Fourth-Grade Gains.

Source	*Set*	*Math*	*Reading*	*Language*	*Social Studies*	*Science*
System (*S*)	1	5.63	3.66	5.68	4.23	2.55
	2	5.56	5.07	4.62	4.02	3.00
Heterogeneity (*H*)	1	0.20	0.03	0.13	2.53	0.62
	2	1.84	1.32	0.94	1.47	1.00
Class size (*C*)	1	1.65	1.00	1.30	2.83	1.47
	2	0.39	1.14	1.14	0.81	0.49
H*C	1	2.29	0.80	0.98	2.30	0.75
	2	1.31	0.69	0.62	2.40	1.11
Teacher (*S*H*C*) (*T*)	1	11.17	6.04	9.24	7.17	7.93
	2	12.49	5.72	10.48	6.69	7.62
Achievement level (*A*)	1	2.45	13.04	8.61	3.37	10.99
	2	6.70	11.92	8.36	4.59	10.91
*A*S*	1	2.63	3.01	1.86	2.14	1.55
	2	3.50	4.50	1.43	5.27	3.74
*A*H*	1	0.28	1.32	2.53	2.01	0.12
	2	0.59	0.89	1.02	0.55	2.06
*A*C*	1	2.96	0.84	1.18	1.53	0.34
	2	1.09	1.99	0.99	0.42	1.68
*A*H*C*	1	1.13	1.33	0.02	0.73	1.25
	2	1.50	0.18	0.05	1.09	0.78
*A*T*	1	1.75	0.56	1.40	2.45	1.24
	2	2.14	2.61	1.10	1.06	0.47
N	1	10344	10477	10497	9438	9329
	2	13102	13102	13498	12320	12406

Set: 1 = 30 East Tennessee school systems.
2 = 24 Middle Tennessee school systems.
N = total number of students.

differences among school systems in twenty-seven of the thirty analyses, and the effect sizes are in most cases impressively large, though not nearly as large as for the teacher and achievement-level factors. A notably nonsignificant factor was class size. The main effect for class size was significant in only three of the thirty analyses. In two of these three instances, the smaller-size class had the higher gains; in the other case, the larger-size class had higher gains. Class size also appeared in a number of statistically significant interactions, though most of these had relatively small effect sizes. The interpretations of these interactions are as varied as those for the class-size main effect. Since the objective was not to investigate the class size effect per se but merely to control for that effect where it occurs, no further discussion of this point is offered.

Based upon an effect size (z-value) of 2.0 (corresponding to a significance level of approximately 0.05), the main effect for heterogeneity was statistically significant in only two of the thirty analyses, approximately the number that would be expected to occur by

Table 3. z-Values for Analyses of Fifth-Grade Gains.

Source	*Set*	*Math*	*Reading*	*Language*	*Social Studies*	*Science*
System (*S*)	1	1.30	3.52	3.18	1.04	1.30
	2	5.69	3.50	2.49	4.20	3.02
Heterogeneity (*H*)	1	0.55	0.57	1.44	0.37	2.56
	2	0.66	0.33	1.41	0.12	0.59
Class size (*C*)	1	2.19	0.72	0.59	1.58	2.35
	2	1.13	1.40	0.71	0.14	0.01
H*C	1	0.29	0.82	0.23	1.13	1.77
	2	0.66	0.79	1.37	0.10	0.11
Teacher (*S*H*C*) (*T*)	1	9.70	5.80	6.29	5.65	6.24
	2	9.13	6.33	9.68	6.62	6.27
Achievement level (*A*)	1	1.94	4.42	1.51	0.14	5.20
	2	3.88	5.12	2.26	1.29	2.24
*A*S*	1	2.60	2.03	2.64	0.91	2.15
	2	3.36	2.15	0.98	4.24	0.59
*A*H*	1	2.81	1.07	1.10	0.78	1.18
	2	0.70	2.40	0.91	1.22	0.97
*A*C*	1	2.07	1.09	1.70	0.94	0.93
	2	2.35	1.18	0.13	0.86	0.88
*A*H*C*	1	1.49	0.06	1.31	0.24	1.63
	2	1.46	0.39	1.43	0.45	3.04
*A*T*	1	1.79	2.52	1.52	0.05	0.63
	2	3.48	0.64	0.00	0.00	1.87
N	1	8259	8874	8615	6527	6662
	2	9939	9629	10141	9136	8569

Set: 1 = 30 East Tennessee school systems.
2 = 24 Middle Tennessee school systems.
N = total number of students.

chance. The statistically significant effects for heterogeneity were found in fourth-grade social studies and fifth-grade science in East Tennessee. In the first instance, the estimated mean gains for the three groups (low, moderate, and high heterogeneity) were 26.9, 26.4, and 21.6. In the second instance, the estimated mean gains were 10.8, 10.7, and 15.9. So in one case, higher gains occurred under lower heterogeneity, and in the other case higher gains occurred under higher heterogeneity. (Note that the scales for social studies and science are not comparable, so the larger point gains in social studies do not indicate greater academic progress than the smaller ones indicated for science.)

In addition to significant main effects, there were a number of statistically significant interactions, including a significant three-way interaction of achievement level, heterogeneity, and class size in four of the thirty analyses. Specifically, in the thirty analyses there were a total of 180 interaction effects of which fifty-one were statistically significant. However, the effect sizes were relatively small: only seventeen exceeded 3.0 (in absolute value) and only eight exceeded 4.0. The largest interaction effect had a z-value of 5.31. For comparison, the smallest teacher effect size was 5.65. While some of the interaction effects appear to be different from zero, their interpretation tends to vary from subject to subject and grade to grade so that no general conclusions can be drawn. For example, there were seventeen significant interactions involving the heterogeneity factor (out of a total of ninety interactions involving heterogeneity in the thirty analyses), mostly with relatively small effect sizes. From these analyses, we conclude that the effect of intraclassroom heterogeneity neither as a main effect nor interacting with other factors is important in the academic growth of students.

Discussion

Despite ongoing debates about whether, and how much teachers make a difference in student learning relative to a host of other factors assumedly affecting student learning (Wang, Haertel & Walberg, 1993), and whether particular elements of teaching can be systematically and causally linked to student achievement (Scriven, 1990), the results of this study well document that the most important factor affecting student learning is the teacher. In addition, the results show wide variation in effectiveness among teachers. The immediate and clear implication of this finding is that seemingly more can be done to improve education by improving the effectiveness of teachers than by any other single factor. *Effective teachers appear to be effective with students of all achievement levels, regardless of the level of heterogeneity in their classrooms.* If the teacher is ineffective, students under that teacher's tutelage will achieve inadequate progress academically, regardless of how similar or different they are regarding their academic achievement. This finding is corroborated by recent research on the cumulative effects of teachers on the academic progress of students (Sanders & Rivers, 1996). These recent studies show that teacher effects on student learning as inferred from standardized test scores are additive and cumulative over grade levels with little evidence of compensatory effects. Thus, students in classrooms of very effective teachers, following relatively ineffective teachers, make excellent academic gains but not enough to offset previous evidence of less than expected gains.

The other dominant factor in the results of the analyses reported here was the achievement level of the student. Table 4 shows the estimated mean gains in each achievement

Table 4. Estimated Mean Gains by Four Achievement Levels with Standard Errors in Parentheses.

		Achievement Level				
	Set	*Lowest*			*Highest*	*z*
Third grade	1	64.2 (1.6)	56.0 (1.4)	45.2 (1.4)	35.9 (1.4)	17.0
	2	75.4 (1.2)	59.3 (1.2)	47.5 (1.1)	36.6 (1.1)	28.0
Fourth grade	1	20.8 (1.4)	19.3 (1.1)	19.9 (1.1)	16.1 (1.2)	2.5
	2	28.7 (1.1)	25.7 (1.1)	21.4 (1.0)	20.5 (1.0)	6.7
Fifth grade	1	23.6 (1.4)	26.1 (1.2)	27.0 (1.2)	24.0 (1.3)	1.9
	2	25.9 (1.1)	27.2 (1.0)	25.9 (1.1)	21.2 (1.2)	3.9
Reading:						
Third grade	1	42.5 (1.5)	34.0 (1.2)	27.7 (1.3)	19.4 (1.3)	12.7
	2	45.3 (1.2)	33.0 (1.0)	26.6 (1.0)	16.4 (1.0)	20.1
Fourth grade	1	10.5 (1.1)	16.8 (0.9)	20.4 (1.0)	28.5 (1.0)	13.0
	2	16.7 (1.0)	20.8 (0.9)	22.9 (0.9)	32.6 (1.0)	11.9
Fifth grade	1	9.7 (1.3)	9.7 (1.1)	16.0 (1.1)	13.6 (1.1)	4.4
	2	11.6 (1.1)	10.3 (1.1)	16.0 (1.0)	17.4 (1.1)	5.1
Language:						
Third grade	1	29.7 (1.1)	25.1 (1.0)	18.4 (1.0)	23.0 (1.0)	8.5
		30.7 (0.9)	26.6 (0.8)	21.3 (0.8)	23.4 (0.8)	9.0
Fourth grade	1	10.7 (1.1)	20.0 (1.0)	18.5 (1.0)	23.4 (1.1)	8.6
	2	16.2 (1.0)	21.7 (1.0)	21.1 (0.9)	27.3 (1.0)	8.4
Fifth grade	1	14.8 (1.1)	16.9 (1.1)	15.8 (1.0)	17.9 (1.1)	1.5
	2	13.5 (1.0)	14.6 (1.1)	15.8 (1.0)	17.5 (1.1)	2.3
Social studies:						
Third grade	1	40.8 (2.0)	46.9 (1.7)	37.1 (1.6)	24.4 (1.6)	10.0
	2	46.2 (1.7)	49.0 (1.4)	39.8 (1.3)	23.6 (1.4)	14.5
Fourth grade	1	26.7 (1.9)	27.5 (1.6)	26.3 (1.6)	19.5 (1.7)	3.4
	2	28.5 (1.6)	31.4 (1.4)	29.4 (1.4)	22.3 (1.4)	4.6
Fifth grade	1	30.2 (1.8)	30.1 (1.6)	29.1 (1.6)	30.8 (1.8)	0.1
	2	28.9 (1.6)	28.3 (1.5)	25.6 (1.5)	25.7 (1.3)	1.3
Science:						
Third grade	1	18.1 (1.9)	28.5 (1.5)	24.5 (1.5)	15.9 (1.5)	6.8
	2	23.3 (1.5)	30.1 (1.3)	25.2 (1.2)	15.8 (1.3)	8.4
Fourth grade	1	24.9 (1.7)	22.6 (1.4)	17.6 (1.4)	5.6 (1.4)	11.0
	2	25.0 (1.5)	24.4 (1.2)	20.0 (1.2)	8.3 (1.3)	10.9
Fifth grade	1	19.6 (1.7)	10.2 (1.5)	8.2 (1.4)	11.8 (1.6)	5.2
	2	13.7 (1.6)	9.4 (1.4)	9.3 (1.3)	12.9 (1.3)	2.2

Set: 1 = 30 East Tennessee school systems.
2 = 24 Middle Tennessee school systems.

level group for all thirty analyses (including four in which the effect was not statistically significant). No universally applicable pattern emerges, but it is worth noting that out of the twenty-six analyses in which achievement level was significant, the largest gains occurred in the lowest achievement group twelve times, in one of the two middle groups eight times, and in the highest group six times. Similarly, the smallest gains occurred in the highest achievement group fifteen times, in one of the two middle groups six times, and in the lowest group five times. In other words, there is a disturbingly common but not universal pattern for the best students to make the lowest gains. Possible explanations include a lack of stretch in curriculum and instruction to accommodate the highest achievers and insufficient availability of higher level course offering in all schools.

Hundreds of studies on ability grouping have been conducted since the 1930s. Recent meta-analyses of these studies by Slavin (1987, 1990) and Kulik (1992) have synthesized the findings of the most rigorous studies. Slavin, in both of his studies, discovered that ‘‘study after study, including randomized experiments of a quality rarely seen in educational research, finds no positive effect of ability grouping in any subject or at any grade level, even for the high achievers most widely assumed to benefit from grouping’’ (Slavin, 1990, p. 491). Experts on ability grouping contend that the effects of grouping on achievement are minimal except in classrooms where there is significant curricular adjustment to meet the needs of students at different levels (Kulik, 1992; O’Neil, 1992; Rogers & Kimpston, 1992). Slavin (1990, p. 491) goes so far as to suggest that ‘‘the lesson to be drawn from research on ability grouping may be that unless teaching methods are systematically changed, school organization has little impact on student achievement,’’ This study supports Slavin’s conclusion.

Teachers seem to have far more to do with the academic progress of students than does the method used for assignment of children to teachers. The contention that high academic gains are more likely to be produced in highly homogeneous classrooms is not supported by our research, and, therefore, neither is the corollary that teachers with highly heterogeneous classrooms should not be expected to make those gains.

Perhaps the persistence of the phenomenon of ability grouping in American schools, despite the preponderance of research attesting to its ineffectiveness, can be attributed to the reluctance of the educational community to assign responsibility for student achievement to teachers. Travers (1981, p. 18) expresses this point of view thusly: ‘‘The extent to which a pupil learns in the school is a function of many different conditions, of which the teacher’s mode of operation is only one. . . . The teacher factor may well account for only a small amount of the differences in achievement.’’ Such statements as these, in turn, may derive from two widely held beliefs: that the interplay of the educational setting with factors outside the purview of formal education prevents the correct attribution of learning effects; and that most educational assessment tools and standardized tests, in particular, are poor indicators of academic progress (for a discussion of this latter point, see Sanders & Horn, 1995a). However, these beliefs do not seem supported and are contrary to the findings of this study. It is recognized here, however, that identifying a common set of factors and interpretation of their effects on student learning and achievement presents a highly complex set of methodological and theoretical issues (Wang, Haertel & Walberg, 1993).

Conclusions and Implications

Differences in teacher effectiveness were found to be the dominant factor affecting student academic gain. The importance of the effects of certain classroom contextual variables (class size and classroom heterogeneity) appears to be minor and should not be viewed as inhibitors to the appropriate use of student outcome data in teacher assessment. These results suggest that teacher evaluation processes should include, as a major component, a reliable and valid measure of a teacher's effect on student academic growth over time. The use of student achievement data from an appropriately drawn standardized testing program administered longitudinally and appropriately analyzed can fulfill these requirements. If the ultimate goal is to improve the academic growth of student populations, one must conclude that improvement of student learning begins with the improvement of relatively ineffective teachers regardless of the student placement strategies deployed within a school.

In addition, student academic level was found to be significantly related to academic progress, although not nearly to the degree found for the teacher. Disproportionately, high-scoring students were found to make somewhat lower gains than average and lower-scoring students. Possible explanations include lack of opportunity for high-scoring students to proceed at their own pace, lack of challenging materials, lack of accelerated course offerings, and concentration of instruction on the average or below-average student. This finding indicates that it cannot be assumed that higher-achieving students will ''make it on their own.''

Though the debate about whether student achievement data should be used as part of an assessment, evaluation, and accountability system for teachers will assuredly continue, the results of this study suggest that *teachers do make a difference* in student achievement. It is recognized here, however, that there were no direct, systematic observations of the quality of teaching and learning at the classroom level in this study. Thus, identifying teachers that clearly get results over time, and comparing them to teachers over time who do not, seems a logical, worthwhile next step in addressing the issues raised here and in further developing general lines of inquiry about the important relationship between teacher effectiveness and teacher evaluation. If characteristics of teaching and learning environments that differentiate teachers who are demonstrably effective (as opposed to ineffective) in different contexts over time can be documented, subsequent teacher evaluation systems might be developed to accommodate these characteristics. Continuing debates aside, the results presented here suggest that *teachers indeed make a difference* and that homogeneity and heterogeneity of student ability levels within classes are not major concerns in assessing teacher effectiveness. Those developing future teacher evaluation systems might take comfort in the results reported here with the suggestion that variation in ability levels of students, despite teacher arguments and conventional wisdom, is not a major factor framing effectiveness in teaching.

Notes

1. This model would not be adequate and appropriate to provide the best possible estimate of an individual effect. Rather the full TVAAS model should be used (Sanders, Saxton & Horn, in press).

References

Berk, R. (1988). Fifty reasons why student achievement gains does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education*, 1(4) 345–364.

Brophy, J. (1968). Teacher influences on student achievement. *American Psychologist* (October), 1069–1077.

CTB/McGraw-Hill (1990). *Comprehensive test of basic skills* (4th ed.). Spring Norms Book. Monterey, CA: CTB/Macmillan/McGraw-Hill.

Ellett, C. D. (1990). *A new generation of classroom-based assessments of teaching and learning: Concepts, issues and controversies from pilots of the Louisiana STAR*. Baton Rouge: Teaching Internship and Statewide Teacher Evaluation Projects, College of Education, Louisiana State University.

Kulik, J. A. (1992). *An analysis of the research on ability grouping: Historical and contemporary perspectives*. National Research Center on the Gifted and Talented. Storrs, CT: University of Connecticut.

O'Neil, J. (1992). On tracking and individual differences: A conversation with Jeannie Oakes. *Educational Leadership*, 50(2), 18–21.

Peterson, D., Kromrey, J., & Smith, D. C. (1990). Research-based teacher evaluation: A response to Scriven. *Journal of Personnel Evaluation in Education*, 4(1), 7–18.

Porter, A. C., & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the work of the Institute for Research on Teaching. *Educational Leadership*, 45(8), (May), 74–85.

Rogers, K. B., & Kimpston, R. D. (1992). Acceleration: What we do vs. what we know. *Educational Leadership*, 50(2), 58–61.

Rosenthal, R. (2984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Sanders, W. L., & Horn, S. P. (1995a). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. *Educational Policy Analysis Archives*, 3(6).

Sanders, W. L., & Horn, S. P. (1995b). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. In A. J. Shinkfield & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp. 337–350). Boston: Kluwer.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Research Progress Report. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (In press). The Tennessee Value-Added Assessment Systems (TVAAS): A quantitative, outcome-based approach to educational assessment. In J. Millman (Ed.), *Assuring accountability? Using gains in student learning to evaluate teachers and schools*. Thousand Oaks, CA: Corwin Press.

Schalock, H. A., & Schalock, M. D. (1993). Student learning in teacher evaluation and school improvement: An introduction. *Journal of Personnel Evaluation in Education*, 7(2), 103–104.

Scriven, M. (1990). Can research-based teacher evaluation be saved? *Journal of Personnel Evaluation in Education*, 4(1), 19–39.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, 57(3), 293–336.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, 60(3), 471–499.

Travers, R. M. W. (1981). Criteria of good teaching. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 14–22). Beverly Hills, CA: Sage.

Wang, M. C., Haertel, G., & Walberg, H. J. (1993). Toward a knowledge base of school learning. *Review of Educational Research*, 73(3), 249–294.