DOCUMENT RESUME

ED 109 064 SP 009 310

AUTHOR Acland, Henry
TITLE A Study of Teacher Effects Based on Students' Achievement Scores.
PUB DATE 75
NOTE 37p.

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS *Academic Achievement; *Effective Teaching; Elementary Education; Grade 5; Student Evaluation; Teacher Behavior; *Teacher Influence; Teaching Skills; Tests
IDENTIFIERS *Metropolitan Achievement Test (MAT)

ABSTRACT

This report tests the assumption that teachers have an impact on how much students learn. The results of this study indicate that teachers have an effect on average class achievement scores, and that this effect can be broken down into a stable component attributed to the teachers' consistency, and an unstable effect which varies from year to year. The stable component can be obtained by measuring (a) consistency teachers have in teaching different skills to the same students, and (b) consistency in teaching the same skill to different students. The data were collected from 89 fifth-grade teachers. Student achievement was tested in October and April in two consecutive years on the Intermediate Battery of the Metropolitan Achievement Test (MAT:1959). Adjusted gain scores were computed, based on class means, and the gain was used as an index of relative teacher effectiveness. The following three assumptions are implicit in the use of these gains: (a) the MAT is a relevant index of student performance, (b) gain scores measure teachers' deliberate behavior and variables beyond control of the teacher, and (c) students in average or below-average classes may learn considerably during the year, although in comparison to other classes they may have learned less. Results of the study also indicate that teachers are not found to have a consistent effect on the spread of achievement scores in their classes. (Author/JS)

A STUDY OF TEACHER EFFECTS

BASED ON STUDENTS' ACHIEVEMENT SCORES

HENRY ACLAND

1975

Introduction

There is a widely shared view that some teachers yield better results than others. It is not just that some are more skilled at teaching but that their abilities have a degree of constancy. This is reflected in the implicit assumptions of hiring and promotion policies. One teacher is chosen, and another turned down, because the past teaching records of both are thought to predict their future performances. The same beliefs underlie training programs that are designed to teach competencies that make graduates of the program consistently more effective than the recruits. Finally, the belief in variation in teacher effectiveness underlies research on teachers which has sought to identify the behavioural correlates of effective teaching as if the prior problem of establishing the existence of variation in teacher productivity had been dispensed with. The research reported here tests the evidence on which such beliefs in teacher effectiveness might rest, by examining the importance of their effects.

Several different definitions of teacher impact are used so it is essential to separate and clarify them at the outset. The definitions rest on a set of distinctions, the most important of which is the distinction between the stable

and unstable components of teacher impact. The stable component of the consistent teacher effect can be measured in two ways: the consistency teachers have in teaching different skills to the same group of students and their consistency in teaching the same skill to different groups of students. The first way of looking at consistent effects examines the degree to which a teacher who is good at teaching one kind of competence, say in reading, is also effective at teaching another, say arithmetic computation. The second way of looking at consistency examines the degree to which a teacher who is effective in one year, with one class of students, is effective in the following year with a different group of students. It is this second way of defining teacher effects that gets most attention here. The reason for this is that the second definition corresponds to the implicit belief, referred to above, in consistency over time in the way teachers contribute to student learning.

Stable teacher effects are distinguished from unstable effects. The unstable effect is defined in temporal terms as the component which varies from one year to the next. It is the teacher effect which is specific to a given period, here the school year, and which cannot be attributed to that teacher's stable effectiveness. If this effect is found to be considerable it would imply that teacher effects vary from one context to another, and this opens up the question of the nature and determinants of this context.

For example, it might be that the student composition of the classroom affects the teacher's style, and hence the teacher's effectiveness. Alternatively, teachers may alter their techniques independently of the students they teach. Such possibilities suggest sets of hypotheses about the way teachers operate, but these questions lie outside the scope of this paper which is concerned with the simpler problem of the importance of the unstable effect.

Cutting across the distinction between stable and unstable teacher effects is the distinction between effects on the class average score and effects on the spread of scores within the class. Analaysis of teacher effects on the class average implies a concern with the average shifts in the performance of the students in the teacher's class. By contrast, analysis of the spread of scores, measured by the standard deviation or the variance, implies an interest in the degree to which teacher increase or decrease the disparities in the performance of students within their class. More attention here is given to teacher effects on class mean scores than their effects on the variance. Once again, this is because the prevalent conception of the difference between an effective and less effective teacher is that the former is successful in improving the overall performance of the class. Perhaps the most obvious way of measuring the overall performance of the class is with the class' average score. A less common definition of an effective teacher is one

who has a differential effect on students such that the spread of scores in a class increases or decreases over the course of the year. This definition leads to examination of teacher effects on the dispersion of students' scores.

Litte research has adopted the framework used in this paper. Several short-term studies of teacher consistency have been carried out, but only four long-term studies were located, where the teacher's contact with the students was longer than four weeks. All have concentrated on teachers' impact on class average scores. Three of these studies (Morsch et al 1955, Harris et al 1968, Soar 1966) have been reviewed by Rosenshine (1970) and his summary requires little embellishment:

> "...on the basis of these studies, evidence on the consistency of teacher effects is weak because correlations as high as .5 were obtained in only one study and all other correlations were about .35 or much lower."

The fourth study, by Brophy (1973), looked at the stability of teacher effects across three years and found correlations "generally higher than those obtained in the long-term studies reviewed by Rosenshine." Median correlations for the four subgroups of teachers lie between .25 and .42. These studies are congruent and imply a modest stability of teacher effects across years. However, it will be argued in this paper that the correlation by itself does not really indicate the importance of the stable teacher effect. This

study is also justified because it seems likely that conclusions about teacher effects will be pieced together from small studies rather than deriving from large-scale surveys. The main reason for this is the complicated student testing program that is required for the research design. Given this practical problem, research on teacher consistency will probably be based on several small, and possibly unrepresentative samples of teachers.

Method

The data were collected from 89 fifth grade teachers, who form a systematic selection of all fifth grade teachers in a large school system. Teachers were included only if they taught at the same grade level in the same school in two consecutive years and if the school's fifth grade classes were self-contained. Where students were taught by more than one teacher the school was dropped. It is assumed in this analysis that the variation among these 89 teachers bears some resemblance to the variation that would be found in samples of larger populations.

Student achievement was tested in this school system in October and April in two consecutive years on the Intermediate Battery of the Metropolitan Achievement Test (1959). Different forms of this test were given in fall and spring. The results of the tests were obtained for each of nine MAT subtests for each student. The main analyses of this paper are based on class average achievement scores which are calculated for each subtest. For each subtest there are (at least) ten possible ways of deriving an average score.

In the first place, average scores can either be calculated on the basis of "matched" students or on the basis of "unmatched" students. Matched students are those who took a given subtest in both fall and spring of a given year; the unmatched group is that which was tested on one or the other of these occasions, either fall or spring, but not necessarily both. The second

source of complication is the variety of metrics which can be derived from the raw scores. Only two are considered here: publisher's standardized and grade equivalent scores. Following the procedures laid down in the publisher's manual, raw scores can be converted into nationally normed standardized scores and these standard scores can be transformed into grade equivalents. The third source of complication concerns whether scores are transformed before or after aggregation. For instance, pupil-level raw scores can be transformed into standardized and grade equivalent scores and subsequently aggregated to give a class average on all three metrics. Alternatively, the raw scores can be aggregated, to give a class average raw scores, and this average score can be transformed into the standardized and grade equivalent scores. These ten possible routes to an aggregate score are summarized in the diagram:

	Matched	Unmatched
Raw scores	1	2
Standardized scores	3	4
Grade equivalent scores	5	6

1,2 = no transformation possible, one class average score computed in each cell;

3-6 = transformation from raw score into metric can occur before or after aggregation of data. Two class average scores computed in each cell.

It was important to see if different aggregation routines altered the analysis. To test the proposition that it would make a difference, the class average fall scores set out above were correlated to examine differences in the relative ordering of classes. In the first place, the effect of using different metrics was examined by correlating raw score averages, with standardized score and grade equivalent averages both the latter being transformed prior to aggregation. Correlations were calculated for each subtest between each type of metric. They were all extremely high and sufficiently close that it is unnecessary to report each correlation. The mean correlation between raw and standardized averages is 0.997, between raw and grade equivalent averages, 0.984 and between standardized score and grade equivalent averages, 0.980. Since the cla averages are so closely associated any one metric could be substituted for the other; in fact the standardized metric was chosen because it is easier to make comparisons across subtests.

The second concern was whether the point of aggregating scores, either before or after transformation, would make a difference to the ranking of classes. Again, correlations were computed between scores that had been transformed and subsequently aggregated and between aggregated scores which had subsequently been transformed. Here too the correlations were high, averaging 0.994 for standardized scores and 0.982 for grade equivalent scores.

The final question was whether the class averages should be based on matched or unmatched groups of students. There are obvious reasons for preferring the matched groups, but an argument

could be made for using unmatched groups if the increase in the number of students improved the reliability of class average scores. In fact, the correlations between matched group and unmatched group means were sufficiently high for this decision to seem unimportant. For raw scores the average correlation between matched and unmatched averages across subtests is 0.990, for standardized scores, 0.990 and for grade equivalent scores 0.987. As a result of these analyses, only standardized average scores derived from matched groups, where the standardization preceded the aggregation of data, will be used in the rest of the paper.

For the analyses concerned with teacher effects on class mean scores, adjusted gain scores are computed, based on the class means, for each teacher in Year 1 and again in Year 2. This gain is used as an index of relative teacher effectiveness. Effectiveness is here defined as the amount by which the class average scores in the spring exceeds the level that would be predicted on the basis of the class average score in the fall. It is a relative measure in the sense that it compares one teacher's effectiveness to the average effectiveness of this group of teachers. A class adjusted gain is the difference $(\hat{Y} - Y)$, where $\hat{Y}$ is equal to $a + b.X_{fall}$. In this expression, a is the intercept and b the regression coefficient from the regression of spring on fall scores of a given subtest. Only one independent variable was used since it was found that use of more independent variables did not seriously affect the main analyses (see Tables 2 and 3).

There are three assumptions implicit in the use of these gains. First it is assumed that the MAT is a relevant index of student performance. The study can be criticized on the grounds that these teachers were not trying to improve skills which the MAT measures, but this issue cannot be resolved without investigating these teachers' goals directly. The validity of the MAT for the purposes of this study is a matter of judgement.

Second there is uncertainty about designating the gain score as a measure of teacher effectiveness. The adjusted gain scores may measure the consequences of teachers' deliberate behavior; however, other influences may be at work which are associated with the teachers, but which are not under the teacher's control. For example, gains may reflect uncontrolled school level variables such as the presence of a particular curricular model or they could be due to aggregate differences in home background such as the level of parental encouragement. If factors such as these play a part it would lead one to a conclusion, directly opposed to Brophy's, that the adjusted gains are best regarded as maximal, not minimal estimates of teachers' effects. However, this conclusion should be offset against the observation that the MAT, like other standardized tests, is constructed in a way which makes it insensitive to the unique effects of different teachers. Standardized tests are designed to be fair in the sense that they test skills which all students could be expected to have had the chance to learn. This lessens the chance that students who have learned particular things in uncommon situations, a special program perhaps or an

unconventional teacher, will stand out above students who were exposed to more conventional situations. A different kind of achievement test could well give different and possibly higher estimates of teacher effectiveness.

Third, students in average or below average classes may still learn considerable amounts during the year although in comparison to other classes they have learned less. Teachers presumably have absolute effects in addition to having relative effects to one another. Unfortunately, these absolute effects could not be measured with these data.

Analysis

1) Consistency across subtests within years

The first four sections of the analysis are concerned with th ect teachers have on the class average score. This part looks at the degree to which class adjusted gains based on class mean scores of a given subtest of the MAT are consistent with gains measured on other subtests during the same school year. If high levels of consistency are found it would suggest that teachers who are effective in increasing a class' average score on one skill are also effective at teaching other skills measured by the nine MAT subtests or, alternatively, that the nine subtests are measuring essentially the same kind of competence and this single competence is being influenced by the teacher. The credibility of the second of these alternatives can be tested by factor analyzing the nine MAT subtests derived from each of four different testing occasions: fall and spring of Year 1 and fall and spring of Year 2. Class mean scores are used in each analysis. If the first principal component accounts for a large proportion of the total variation in the nine subtests it would be reasonable to conclude that the subtests measure a common skill. In the four factor analyses the first principal component accounts for between 82.0% and 84.5% of the variance--a relatively high percentage. This means that if the within-year correlations among adjusted gains are high, the consistency could be attributed to the commonality of the nine MAT subtests. However, the inter-correlations of the adjusted gain scores of the subtests, presented separately for Year 1 and Year 2 in Table 1, are only moderate.

TABLE 1. Zero order correlations among adjusted gain scores, within years. Based on standardized scores for matched groups, transformed before aggregation.

YEAR 1

	WK	RD	LG	LS	AC	AP	SS	SK
RD	0.454***							
LG	0.497***	0.325**						
LS	0.356**	0.284**	0.655***					
AC	0.399***	0.267*	0.458***	0.440***				
AP	0.350**	0.387***	0.419***	0.462***	0.561***			
SS	0.485***	0.265*	0.345**	0.293**	0.470***	0.556***		
SK	0.506***	0.273*	0.311**	0.340**	0.296*	0.308**	0.354**	
SC	0.612***	0.287**	0.361***	0.327**	0.412***	0.435***	0.645***	0.365***

Correlations based on minimum of 70 cases

YEAR 2

	WK	RD	LG	LS	AC	AP	SS	SK
RD	0.365***							
LG	0.376***	0.487***						
LS	0.459***	0.568***	0.521***					
AC	0.292**	0.487***	0.482***	0.506***				
AP	0.229*	0.511***	0.521***	0.446***	0.689***			
SS	0.301**	0.770***	0.435***	0.609***	0.468***	0.581***		
SK	0.273*	0.469***	0.473***	0.415***	0.328**	0.526***	0.603***	
SC	0.295**	0.328**	0.292**	0.512***	0.371***	0.511***	0.411***	0.417***

Correlations based on minimum of 84 cases

WK = Word Knowledge, RD = Reading, LG = Language, LS = Language Study Skills, AC = Arithmetic Computation, AP= Arithmetic Problem Solving, SS = Social Studies Information, SK = Social Studies Study Skills, SC = Science

The average correlation among the adjusted gains in Year 1 is 0.40 and in Year 2, 0.46.*

The size of these correlations is related to the reliability of the gain scores. If the gains can be shown to have an appreciable error component the correlations will be under-estimated. To exaggerate the possible size of this error, the reliability of the class mean scores was estimated by substituting the most conservative values for the MAT subtests in Shaycroft's formula (Shaycroft, 1962). The estimated reliability, 0.96, is consequently the lowest estimate that could be obtained for these subtests and class sizes. About 60% of the variance in the spring class means can be accounted for by the fall mean scores, which leaves 40% of the original variance containing all the error of those scores. The proportion of this error to the residual variance is the estimated reliability of the adjusted gain scores. In this instance, the error variance in class mean scores is 4% of the total and the residual variance is 40%, indicating a reliability of 0.90 for the gains. Bearing in mind that this estimate is conservative, it seems unlikely that correcting the correlations reported in Table 1 will change them substantially. In view of this and the fact that the subtests load heavily on a single principal component, it is reasonable to conclude that teachers could have differentiated effects on student learning.

*These correlations appear slightly lower than Brophy's within-year correlations. For Brophy's 12 teacher subgroups the average within-year correlations range between 0.29 and 0.71.

A good mathematics teacher is not necessarily good at teaching language skills. However, the fact that all the correlations are positively correlated indicates that there is some correspondence in the degree to which teachers' classes change above or below the average rate in different tested skills.

2) Consistency across years within subtests

The central analyses of this paper are concerned with the teachers' consistency across time. This consistency is measured by the inter-year correlations for the gain scores of the different subtests (see Table 2). A majority of these correlations are statistically significant,* and the median correlation, 0.398, compares well with Brophy's median correlation between successive** annual gain scores of 0.39.

These correlations vary considerably across subtests, although the variation is not as large as found in Brophy's study where the range for successive annual gain scores is -0.12 to 0.78. There are two explanations of the subtest variation. First, as a consequence of different psychometric properties, some of the MAT subtests may be better at measuring the stable component of teacher effects than others. Straightforward examination of the MAT did not reveal any obvious differences between the subtests, but a

*The statistical significance levels are reported even though they are not strictly meaningful when, as in this case, teachers have not been randomly selected.

**This implies Year 1-Year 2 correlations and Year 2-Year 3 correlations, but excludes Year 1-Year 3 correlations.

TABLE 2. Zero order correlations between Year 1 and Year 2 adjusted gain scores. Based on standardized scores for matched groups, scores transformed before aggregation.

Subtest	r	N
Word Knowledge	.488***	81
Reading	.198	82
Language	.398***	80
Language Study Skills	.132	83
Arithmetic Computation	.405***	82
Arithmetic Problem Solving	.457***	80
Social Studies Information	.433***	83
Social Studies Study Skills	.310**	73
Science	.228*	83

simple analysis of this kind is not definitive. Second, the impact teachers have may be more stable in some areas of achievement skills than it is in others. This speculation compounds the earlier finding (Table 1) which showed that gains were only moderately correlated within years, and suggests that teachers' effects are related to the particular achievement test that is used to measure student learning.

A caution is in order. The correlations are circumstantial evidence of a stable teacher effect. They imply the existence of teacher behaviors which are stable and which have consistent effects in successive years, but it must be remembered that these behaviors have not been identified nor have they been observed directly. As mentioned above, effectiveness could be related to an effective curricular model so that teachers who use it are found to be consistently more effective in comparison with other teachers who do not use that curriculum. The same applies to other class-level factors. Therefore, it is important to regard these findings as a tentative indication of a stable teacher effect rather than a proof that some teachers are superior to others as a result of their classroom practices. It is important to bear this proviso in mind in the following analysis, which treats the data as if the consistency measured by the year-to-year correlations could be attributed to the teachers' influence.

3) The Size of Teacher Effects

The analyses in this section are concerned with the practical importance of teacher effects. This will be assessed in two ways; in terms of achievement test units and in relation to the pupil-level variation in test scores.

The first way of expressing the size of the stable component of teacher effects is in terms of the achievement test score units. This requires consideration of both the correlations reported in Table 2 and the variances of the adjusted gain scores, since the correlations alone do not indicate the practical impact of consistent teacher effects. If there is little variation among teachers in terms of their relative effectiveness such that the best teacher is not so different from the worst, then the evidence of consistency will assume less importance. Conversely, the larger the variation in teacher effectiveness, and the more consistent teachers are, the larger their overall impact on student learning.

The method of estimating the size of teacher effects depends on the assumption that Year 1 and Year 2 adjusted gains are imperfect measures of the true differences between teachers. These differences are defined as their ability to consistently change the average level of achievement in their classes above or below the predicted level. Seen this way, the square root of the correlation between Year 1 and Year 2 gains is an estimate of the correlation between the true, unmeasured teacher consistency variable and the observed, adjusted gain scores. Thus it is possible to estimate the proportion of the variance in the adjusted

gains that can be attributed to true differences in the consistent component of teachers' influence. The analysis is summarized in Table 3. For the Word Knowledge subtest, the inter-year correlation is 0.488, and the average standard deviation of the adjusted gain scores is 3.09 test points. The product of the square root of the correlation and the standard deviation (3.09 x 0.698) gives the number of test points associated with one standard deviation difference on the underlying teacher consistency measure. The estimated effects, reported in Column 3 of Table 3, are concrete in the sense that they suggest how much student achievement can be attributed to the stable element of teachers' impact. For instance, a contrast between the average teacher and the teacher at the 84th percentile on the distribution of the unmeasured teacher effect variable is associated with 2.16 test score points on the Word Knowledge test; a more extreme contrast, say between the average teacher in the top and bottom fifths of the effectiveness distribution (2.8 standard deviations) is associated with a difference of 6.05 test score points (Column 4). The average effect associated with the top and bottom fifth contrast, 5.34 achievement test points, implies that teachers can have important consequences for the amount students learn. Some teachers are not only consistently better than others but their practical effects make an appreciable difference to the average student in their classes.

The second way of expressing the importance of teacher effects is based on decomposition of pupil level variance in spring test scores. There are two components of this variance

TABLE 3. Estimates of the size of teachers' impact. Based on standardized scores for matched groups, scores transformed before aggregation.

Column 1: Correlation between Year 1 and Year 2 adjusted gain scores (see Table 2)

Column 2: Average standard deviations for Year 1 and Year 2 adjusted gain scores

Column 3: First estimate of teacher effect. The test points associated with one standard deviation difference on the underlying measure of teacher effectiveness.

Column 4: Second estimate of teacher effect. The test points associated with the contrast between top and bottom fifths of teachers on the underlying measure of teacher effectiveness.

Subtest	1.	2.	3.	4.
Word Knowledge	.488	3.09	2.16	6.05
Reading	.198	3.44	1.53	4.28
Language	.398	3.81	2.40	6.73
Language Study Skills	.132	4.09	1.49	4.16
Arithmetic Computation	.405	3.54	2.25	6.31
Arithmetic Problem Solving	.457	2.76	1.87	5.24
Social Studies Information	.433	3.32	2.18	6.10
Social Studies Study Skills	.310	3.11	1.73	4.85
Science	.228	3.23	1.54	4.32

which are crucial to the analysis. The first is attributable to the stable teacher effect, the second to the unstable teacher effect. This can be explained by reference to the ANCOVA design. Teachers and Years are defined as two factors in a crossed design and students are nested within each Teacher-Year cell. The dependent variable is the spring score for the student and the covariate his fall score for the same subtest being used for the dependent measure. The percentage of student variance that can be assigned to the main teacher effect is called the stable teacher effect; it is that part which is consistent from one year to the next. The second component of variance, the unstable effect, is that which can be attributed to year-specific effects. It is the part of the variance assigned to the interaction term (Teachers x Years). This is also a teacher effect, being the part of their effect which is variable from one year to another. There are several reasons to expect teachers to have such an unstable effect. For example, they may adjust their instructional technique to meet different needs of different groups of students and in doing so alter the amount they teach. Alternatively, the students in the class may create an informal social ambience that makes instruction more or less difficult in a given year. As the composition of the class changes so may the teacher's effectiveness change. This part of the analysis seeks to identify the unstable teacher effect and compare its size to the stable teacher effect.

The results of the ANCOVA are summarized in Table 4, which shows the percentage of variance attributable to the main teacher effect (Teacher) and the interaction term (Teacher x Years). The consistent teacher effect accounts for an average of 4.76% of the student-level variance in spring scores; unstable teacher effects account for slightly more: 5.85% of the variance. Both kinds of teacher effects together account for an appreciable proportion of the overall student-level variance in achievement scores.

The results add to those presented earlier by showing the relative importance of stable and unstable teacher effects. By establishing the provisional evidence for both stable and unstable teacher effects, the findings suggest that teacher are predictable, to some degree, in the effect they have on students. Of course, the decision about whether this effect is large enough to be educationally significant will depend on the immediate context of a policy decision and the goals of the decision-maker. However, it may be added that since the unstable teacher effect is about as large as the stable component, there is little reason to select or allocate teachers on the basis of a belief that teachers are mainly consistent.

TABLE 4. Percentage of student level variance in achievement scores that can be attributed to two sources; the teacher main effect (stable component) and the teacher x years interaction effect (unstable component). Based on standardized scores for matched groups, scores transformed before aggregation.

Subtest	Teacher	Teacher x Years
Word Knowledge	5.95	4.75
Reading	3.01	8.21
Language	6.63	7.17
Language Study Skills	3.18	9.57
Arithmetic Computation	6.43	3.56
Arithmetic Problem Solving	4.53	.68
Social Studies Information	6.74	6.83
Social Studies Study Skills	3.38	2.95
Science	2.95	4.95

4) Specially Effective Teachers

Inspection of the frequency distribution of the adjusted gain scores showed that they were positively skewed with a small number of teachers scoring well over two standard deviations from the mean. The question here is whether this small group of specially effective teachers was consistently effective between years within the same subtests. If specially effective teachers also perform consistently, it is conceivable that the stable teacher effect reported in the previous section can be partly accounted for in terms of a small numbers of teachers.

The most direct way of looking at the part that exceptional teachers play is to inspect the bivariate plot of adjusted gain scores of one subtest for Year 1 and Year 2 (Table 5). This plot shows that there are certainly three, and possibly five teachers who stand out from the rest in the upper right hand portion of the plot. Plots for other subtests revealed similar outlying points. The outlying teachers tend to be consistent as well as specially effective. Results of other teacher consistency studies have not explored the question of outlying data points so the finding cannot be corroborated. This is unfortunate since the finding suggests an important qualification of the results reported above. The specially effective teachers make a disproportionate contribution both to the variance of the adjusted gain scores and to the size of the between-year correlation. Therefore, the teacher effect that has been reported here can be attributed to some degree to the existence of small numbers of special teachers.

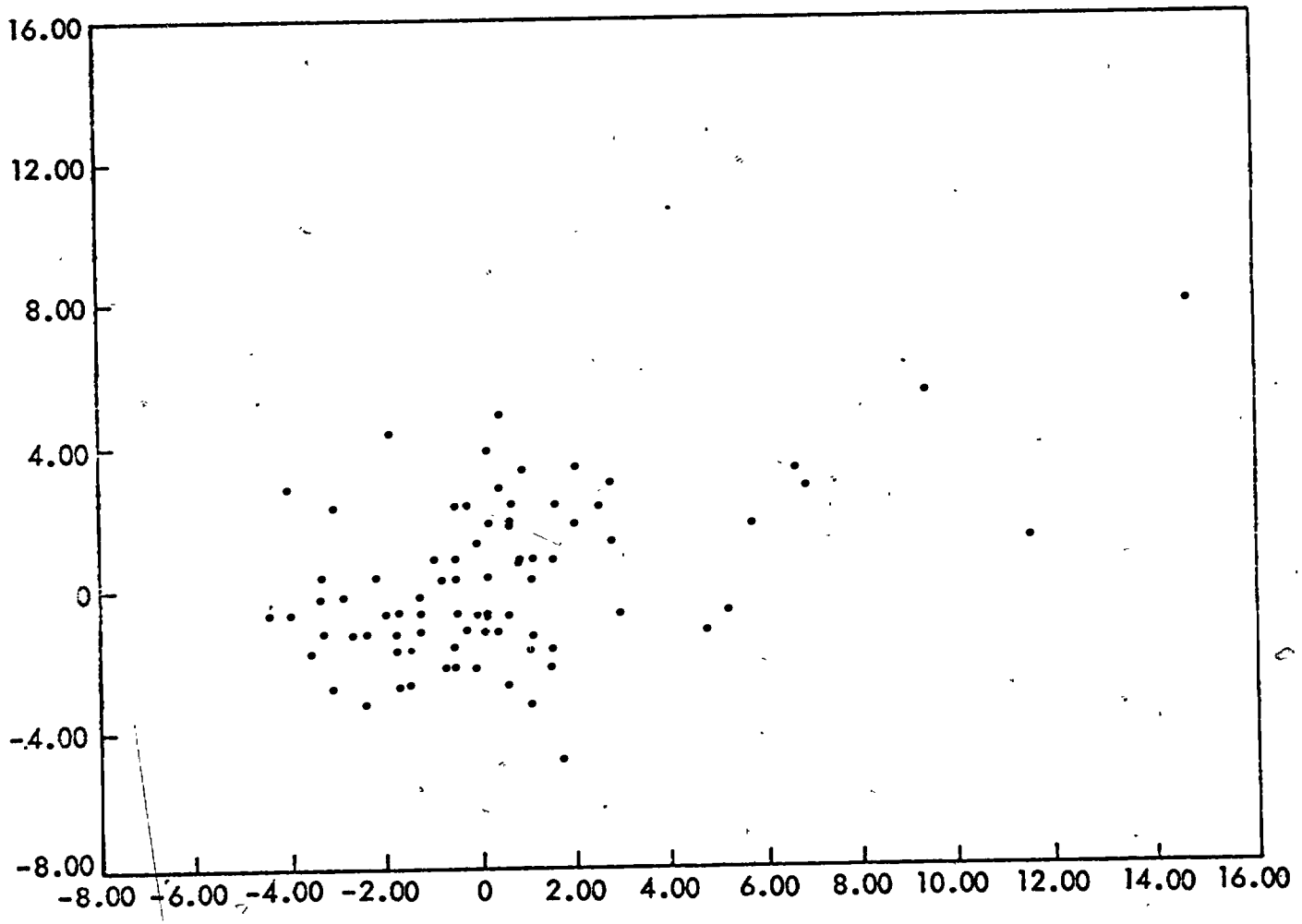

Table 5. Bivariate plot of Year 1 (Y-axis) and Year 2 (X-axis) adjusted gain scores for the Arithmetic Problem Solving subtest, N = 80

The implication for future research is twofold. First, it is important to know if this finding is duplicated in similar studies. Second, in the event that it is, the case could be made for special studies of these teachers on the argument that effective teacher behaviors would be especially evident in this group of teachers, and therefore easier to observe.

5) Teacher Effects on the Spread of Achievement Scores

The first four sections of this analysis have been concerned with the effect teachers have on the average level of performance in the class. The average score, and changes in the average, can and may be unrelated to the dispersion of achievement. So the average scores of two classes may change in the same way while the dispersion of scores changes in very different ways. For instance, the dispersion might shrink in one class relative to the other if the teacher is effective in bringing students within a narrower range of scores than they began with. This might happen as a consequence of differential attention being paid either to the slow or the clever students. Alternatively, the dispersion in one class might increase if the teacher's effects are proportional to a student's initial achievement level. The question raised in this part of the analysis is whether, and to what extent, teachers alter the dispersion of achievement scores.

Within-class variances are computed for each class on each subtest for both Year 1 and Year 2. The central tendencies of these variances are summarized by their means in Table 6. There are three observations to be made about the results. The average

TABLE 6. Average within-class variances, by year, by subtest for standardized scores based on matched groups, scores transformed before aggregation.

	YEAR 1			YEAR 2		
	Fall	Spring	Difference	Fall	Spring	Difference
WK	44.30	49.60	5.30	45.34	49.32	3.98
RD	50.59	57.99	7.40	49.95	60.36	10.41
LG	58.61	64.83	6.22	55.33	65.03	9.70
LS	59.44	72.51	13.07	57.03	72.81	15.78
AC	27.87	53.86	25.99	26.42	54.97	28.55
AP	35.21	43.00	7.79	33.03	44.56	11.53
SS	40.72	43.72	3.00	38.92	39.87	0.95
SK	54.24	59.93	5.69	51.96	57.09	5.13
SC	48.88	59.66	10.78	48.20	58.04	9.84

WK = Word Knowledge, RD = Reading, LG = Language, LS = Language Study Skills, AC = Arithmetic Computation, AP = Arithmetic Problem Solving, SS = Social Studies Information, SK = Social Studies Study Skills, SC = Science.

within-class variance always increases from fall to spring; the increase for a given subtest in Year 1 is very similar to the increase in Year 2 and, most strikingly, there are substantial disparities in the results across subtests.

The increase in spread indicates one of three possibilities: students with high scores move further from the mean, students with low scores move further from the mean, or students near the mean move away from the mean. Since there is no evidence of bimodality in the spring distributions the third alternative seems unlikely. But the question remains of what part teachers play in this shift. The results only hint at the likely direction of teachers' influence; they do not demonstrate to what degree teachers are responsible for changes in variance. In addition, the wide variation in results for different subtests raises the possibility that the psychometric properties of these subtests might account for some of the increase in variance. This deserves consideration.

If these tests are generally too difficult for students in the fall, but become more appropriate for their range of achievement in the spring, an increase of variance would be anticipated such as that reported in Table 6. If this happens the tests which have the most marked floor effect in the fall should also show the largest increase in variance. To test the possibility, an analysis was carried out in which the floor of each subtest is defined, the difference between the average class mean fall scores and the floor for each subtest calculated and this difference score related to the change in variance over the school year for

the subtest in question. If floor effects explain the increases in variance, then there will be a negative relationship between the two difference scores: (average of class means - floor for that subtest), (spring variance - fall variance).

For the purpose of this analysis the floor of the subtest is defined as the chance score, that is, the average score that would be obtained if students checked answers at random. This score could not be calculated for the two arithmetic subtests which have open-ended items. The difference between the chance score and the average of the fall mean scores is defined as the extent to which the subtest has a floor effect. This difference score forms the X-axis of Table 7; the Y-axis is the difference between spring and fall variances. Each subtest contributes two points on the plot, one for each year. The two variables are positively correlated ($r = 0.26$). Thus, the hypothesis that the floor effects of the subtests might explain the increases in variance is rejected, and this leaves open the possibility that some of this increase might be accounted for by the teachers.

Like earlier parts of the analysis, the focus here is on teachers' consistent effects, but the present analysis differs in looking at teachers' impact on the spread of achievement scores rather than changes in the class average scores. The purpose is to establish the existence of a stable teacher effect on within-class variance in the spring while controlling for the initial differences among classes in their fall variances. To this end a two-way ANCOVA is used in which teachers and years are the two factors. The dependent variable is the spring within-class

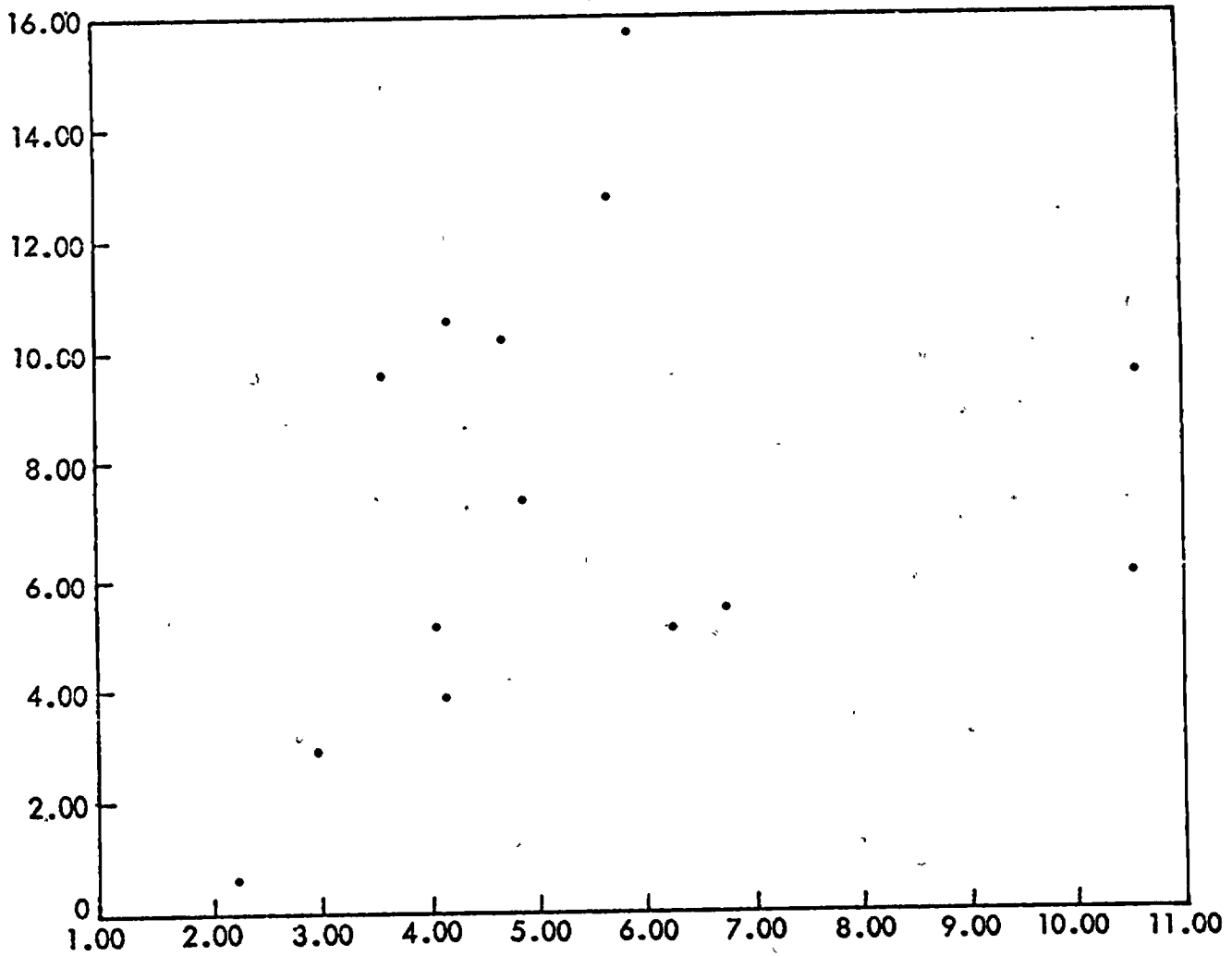

Table 7. Y-axis: change in variance between fall and spring for a given subtest. X-axis: difference between average fall mean score on the subtest and the chance score for the test (see text)

variance of achievement scores and the covariate the fall within-class variance. Each cell in the design has a single observation on dependent variable and covariate.

The test of the existence of a consistent teacher effect on the variance of student achievement scores is the statistical significance of the teacher main effect. The results for each of nine subtests are presented in Table 8. The variability of the results makes it impossible to arrive at a clear conclusion. For some subtests (notably LS, AC, AP, SS), the teacher effect is statistically significant, suggesting that teachers make a consistent difference to the dispersion of achievement scores. But these results must be balanced by the non-significant findings for the Word Knowledge, Reading and Language subtests. Mixed results like these may simply reflect sampling fluctuations. Alternatively, they may be attributed to real effects, in this case a selective teacher effect on the dispersion of achievement scores which is dependent on the type of test that is used to measure student performance. However, the task of devising a hypothesis to account for a selective effect of this kind is formidable.

TABLE 8. Analysis of covariance on class-level variance scores. Dependent variable: spring within class variance. Covariate: fall within class variance. Factors: Teacher and Year. An asterisk in parentheses indicates the term used to test the effects.

SUBTEST		Sums of squares	d.f	Mean square	F-ratio
WK	Teacher	18910.38	80	236.38	0.744
	Covariates	7273.41	1	7273.41	22.90***
	(*)TeacherxYear			317.61	
RD	Teacher	41806.43	81	516.13	0.969
	Covariates	7677.12	1	7677.12	14.41 ***
	(*)TeacherxYear			532.89	
LG	Teacher	40852.10	79	517.12	1.211
	Covariates	7837.22	1	7837.22	18.22 **
	(*)TeacherxYear			427.17	
LS	Teacher	78884.69	82	962.01	1.795**
	Covariates	11446.11	1	11446.11	21.356***
	(*)TeacherxYear			535.98	
AC	Teacher	97831.69	81	1207.80	3.11 ***
	Covariates	10771.84	1	10771.84	27.76 ***
	(*)TeacherxYear			388.01	
AP	Teacher	40580.31	79	513.68	1.94 **
	Year	2584.91	1	2584.91	9.77 **
	(*) TeacherxYear			264.58	

SS	Teacher	36861.97	82	449.54	2.12***
	Covariates	53.15	1	53.15	0.25
	(*) TeacherxYear			211.90	
SK	Teacher	46379.47	72	644.16	1.581*
	Covariates	624.59	1	624.59	1.533
	(*) TeacherxYear			407.32	
SC	Teacher	57656.82	82	703.13	1.677*
	Covariates	1465.49	1	1465.49	3.496
	(*) TeacherxYear			419.17	

Discussion

The main finding of this study, which falls in line with those of four similar studies, is that teachers have a consistent effect on the average scores of the classes they teach in different years. They are also consistent in their effects measured on different subtests within the same school year. Finally, teachers are found to have a year-specific effect which the best estimates available show to be about as large as the stable teacher effect. Teachers do not appear to have a consistent effect on the spread of scores within their classes even though these tend to increase during the school year.

While the general finding of teacher consistency parallels earlier findings, there is one important departure: the discovery of specially effective teachers. This is a startling finding which cannot be confirmed or disconfirmed since data in previous studies have not been analysed in the appropriate manner. It is important to know if the finding is a quirk. If future research showed outliers were a general phenomena, detailed studies of these specially effective teachers would be justified. On the other hand, if replications support the finding of consistent teacher effects, but fail to identify a specially effective subgroup, a somewhat different direction can be envisaged for future work. The ultimate goal of the research should be the identification of the correlates of effective

teacher behaviour. This means defining and isolating the attributes of teachers and the nature of the teaching process which accounts for variations in the adjusted gains of classes. It also means identifying the correlates of the stable component of those gains in the way that stability has been defined here.

References

Brophy, J., 1973 Stability of teacher effectiveness. American Educational Research Journal, 10(3),245-252

Harris A.J.,Morrison C.,Serwer & Gold L. 1968 A continuation of the CRAFT project: comparing reading approaches with disadvantaged urban Negro children in primary grades. New York: Division of Teacher Education of the City University of New York (USOE Project # 5-0570-2-12-1). (ERIC ED 020 297)

Morsch J.E.,Burgess G.G., & Smith P.N., 1955 Student achievement as a measure of instructor effectiveness. Project # 7950, Task # 77243, Air Force Personnel and Training Research Center, San Antonio, Texas.

Rosenshine B. 1970 The stability of teacher effects upon student achievement. Review of Educational Research, 40(5), 647-662

Soar R.S., 1966 An integrative approach to classroom learning. Temple University, Philadelphia, Penna (ERIC ED 033 749)