DOCUMENT RESUME

ED 129 913 95 TM 005 786

AUTHOR Pelavin, Sol H.; Barker, Pierce
TITLE A Study of the Generalizability of the Results of a Standardized Achievement Test.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 76
CONTRACT Rand-B2c-5326
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Education; Examiners; Scores; *Standardized Tests; Statistical Analysis; Student Testing; *Testing; Testing Problems; Testing Programs; *Test Interpretation; *Test Results
IDENTIFIERS *Generalizability Theory; Metropolitan Achievement Test

ABSTRACT

A standardized achievement testing program was begun in Alum Rock, California in the fall of 1972 as part of an evaluation of an Educational Voucher Demonstration. During each of the first three years of the demonstration both the form of test administration and the particular level of the standardized achievement test that a student is assigned have varied. This study assesses what, if any, were the effects of different modes of test administration and what, if any, were the effects of students being assigned out-of-level tests. (Author)

A Study of the Generalizability of the Results of a Standardized Achievement Test

Sol H. Pelavin
Stanford Research Institute

Pierce Barker
Harvard University

Presented at the Annual Meeting
of the American Educational Research Association, 1976
Session # 27.15

This paper is based on work done at the Rand Corporation for the National Institute of Education under Contract B2C-5326.

Since the inception of the Education Voucher Demonstration in Alum Rock, California, standardized achievement tests have been a major component of the demonstration's evaluation. However, for a variety of reasons, the results of the achievement tests have been difficult to interpret. Two of the reasons are that, over the years, different levels of the test were used within the same grade and different types of people administered the tests. To determine the effects of different test levels and administrators, we designed two studies. The studies were simultaneously conducted in Alum Rock during the last week in November and first week in December 1974. This paper presents first the study of the effects of different test levels followed by the study of different test administrators. For each study, we describe the problem that was investigated, the design and results of the investigation, and, finally, the conclusions that may be drawn from the results.

Study 1: Levels of the Metropolitan Achievement Test

Since the fall of 1972, when the standardized testing program began in Alum Rock, the only achievement test has been the Metropolitan Achievement Test (MAT) (Durst, et al.). Different levels of the MAT have been used, not only across grades, as is usual, but also within the same grade. For those readers not familiar with the MAT, we first describe it briefly and then discuss its use in Alum Rock.

The MAT is not one test, but rather a series of six different achievement tests, each of which tests a student on more advanced content as the series progresses from the first test to the sixth. These six tests are referred to as the levels of the MAT. The lowest level is the Primer. The next five levels, in order of increasing difficulty, are Primary I, Primary II, Elementary, Intermediate, and Advanced. On any level of the MAT, a student's raw score represents the number of items that the student has answered correctly. Comparing students' raw scores on different levels of the MAT is not informative since the tests are not of equal difficulty

or length. To facilitate comparisons of the results of different levels, the MAT publishers created the standard score scale which they defined to be equal interval. From tables furnished by the MAT publishers, a student's raw score on any level of the MAT can be converted into a standard score. Theoretically, these tables permit standard scores on different levels to be compared. Thus, the standard score scale supposedly allows all levels of the MAT to be viewed as alternate forms of the same test. To the extent that this is true, the MAT becomes suitable for measuring growth across years and across test levels. In practice, comparisons are usually limited to test levels which are adjacent, or one level apart; for example, Primary II and Elementary are used as alternate forms of the same test.

The six levels of the MAT are standardized only for students in particular grades; for example, the Intermediate level is standardized only for fifth and sixth grade students. Any student who is given a level other than that standardized for the student's present grade is considered to be tested "out-of-grade-level" or "out-of-level." In Alum Rock, being tested out-of-level usually meant that a student was given a level of the MAT one or two levels lower than that recommended by the test publisher. Out-of-level testing occurred frequently because many Alum Rock students were more than one year behind their grade-level. Teachers feared that if the students were tested at grade-level, they would do so poorly that their morale and their academic work would suffer.

Some students in Alum Rock were tested at grade-level and some were tested out-of-level. If the standard score transformations are accurate, the results of adjacent levels of the MAT should be comparable, and out-of-level testing should not affect a student's performance. However, the publishers of the MAT never verified the accuracy of the standard score transformation; that is, they performed no studies of the reliability or validity of the standard scores. Prior to this study, a small experiment was carried out

in the fall of 1973 to determine the effects of out-of-level testing in third grade (Barker and Pelavin). All third grade students were given both the Primary I and Primary II levels of the MAT. The results of this study showed that if the two levels of the MAT were viewed as alternate forms of the same tests, the reliability of the tests was quite low, at most 0.5. There was also a slight indication of bias; that is, for some students, one level of the test would produce a higher score than would the other level. The results of the third grade study raised sufficient doubts about the accuracy of the standard score transformations to cause us to want to investigate the effects of out-of-level testing for other grades and other levels of the MAT. Therefore, a study was designed to be carried out in grades 4, 5 and 6 to determine the effects that out-of-level testing have on students in these grades.

Design

The main objective of this study was to determine whether or not a student's standard score could be generalized across different levels of the MAT; that is, would a student's standard score be the same regardless of the level of the MAT taken. The subjects of this study were four classes from each of grades 4, 5 and 6. Two of the fourth and fifth grade classes came from one elementary school and two came from another. Two sixth grade classes came from each of two middle schools. In terms of reading achievement, these classes were representative of Alum Rock. The district's mean scores for the composite Total Reading, expressed in standard score units, were 56.9, 61.4, and 68.5, respectively, for grades 4, 5 and 6. The mean scores of the students in the study were 56.3, 61.9, and 69.1, respectively, for grades 4, 5 and 6. There seemed to be no meaningful differences between the average reading achievement of the students in the Alum Rock School District and the students included in the study.

During October 1974, the fourth grade students were given the Primary II level of the MAT, the fifth grade students were given the Elementary level,

and the sixth grade the Intermediate level. These tests were administered as part of the normal fall achievement testing program in the Alum Rock School District. When the study was initially conceived in the summer of 1974, it was our intention that this study should begin within two weeks of the October achievement testing. The two-week time interval would have allowed the October testing to be viewed as the first of three repeated measures (or as part of a test-retest). However, because of internal concerns within the Alum Rock School District, this proved to be infeasible. The shortest period between the fall testing and the beginning of the study that was acceptable to the school district was six to seven weeks. It was, however, then feasible to test the students for a third time within a week after the second testing.

Beginning late in November 1974, the students in the study were tested twice more with the reading portion of the MAT. The fourth and fifth grade students were given both the level below and the level above the one they had received in October. The sixth grade students were given the level below that given in October and an alternate form of the same level as the October test. The specific levels for each grade and their time of administration are presented in Table 1. Half of the classes in each grade were initially given the lower of the two levels while the other half were initially given the higher level.

Table 1

MAT LEVELS BY GRADE

Grade	October	November/December
Fourth Grade	Primary II	Primary I, Elementary
Fifth Grade	Elementary	Primary II, Intermediate
Sixth Grade	Intermediate (Form G)	Elementary, Intermediate (Form F)

In order to standardize test administration conditions and thereby attempt to minimize the error variance for the later two testings, each class was tested at the same time of day and on the same day of the week as during the October testing. Moreover, for each class, the same person conducted all three test administrations.

Results: Part I

To determine whether or not the level of the MAT affected a student's performance, we compared mean scores for the different test levels given within each grade. The Total Reading* mean scores are presented in Table 2

Table 2

TOTAL READING MEAN SCORES

	Type of Score and Date of Administration				
	Raw Score	Standard Score		Equivalent Grade	
		Oct.	Nov/Dec	Oct.	Nov/Dec
Fourth Grade (N = 116)					
Primary I	66.4		53.3		2.9
Primary II	59.2	56.3		3.1	
Elementary	46.1		57.6		3.3
Fifth Grade (N = 97)					
Primary II	70.1		65.1		4.1
Elementary	53.5	61.9		3.8	
Intermediate	37.8		67.4		4.3
Sixth Grade (N = 121)					
Elementary	64.1		68.9		4.6
Intermediate (Form G)	41.3	69.1		4.6	
Intermediate (Form F)	43.7		70.6		4.8

*For economy, all of the analyses were done on the composite Total Reading Score.

in three metrics. Mean raw scores are included merely for informational purposes and should not be used to compare the results of different levels.

We expect the mean scores for the test given during November/December to be higher than the October mean score. From an analysis of data from previous years, we estimate that students' scores increase at least 1.5 standard score units during any two-month interval (Barker). Therefore, if the standard score transformations are accurate, we expect both November/December means to be about 1.5 standard score units greater than the mean for October. We would also expect that the two levels of the MAT administered one week apart in November and December would have the same mean scores. We will now discuss the differences in mean scores for each grade.

In fourth grade, the mean score for the Elementary level of the MAT is 1.3 standard score units higher than the mean score for Primary II, which is about what we expect given a two-month interval between test administrations. However, the mean score for Primary II is 3.0 standard score units lower than that for Primary I, which is quite surprising since Primary I was administered almost two months later than Primary II. The difference of 4.3 between Primary I and Elementary is also quite large since they were administered only a week apart. This difference in standard score units is equivalent to a difference of four months on the grade equivalent scale, a difference which might be quite important. Tests more than one level apart do not ask questions on the same content and, probably, their results should not be compared.

We conclude that for students in fourth grade the Primary I level of the MAT is not interchangeable with either the Primary II or the Elementary level of the MAT. Most students score lower on Primary I than on Primary II; we attribute this difference mainly to problems in the standard score scale. A complete discussion of the possible causes of the differences in scores on Primary I and Primary II is contained in our earlier study (Barker and Pelavin). The Primary II and Elementary levels of the MAT do seem to

be interchangeable. The difference in mean scores of 1.3 standard score units might well be caused by growth. We realize that growth is confounded with test level, and therefore the above explanation should be viewed with care.

In the fifth grade sample, the difference of 3.2 standard score units between the Primary II and Elementary levels of the MAT is larger than expected. The difference between Primary II and Elementary scores is greatest for students scoring above the norm for their grade level. One possible cause of this difference is that as students approach the ceiling of the Primary II level (that is, are answering almost all of the questions correctly), their standard scores become inflated.

This is only one possible explanation of the higher Primary II mean score. There is an even larger difference (5.5 standard score units) between students' scores on the Elementary and Intermediate levels of the MAT than between Primary II and Elementary. Part of this difference can be explained by the transformations of chance scores.* On the Intermediate level of the MAT, a chance score is transformed into a substantially higher standard score than is a chance score on the Elementary level. Data from the Educational Testing Service's Anchor Test Study allows the Elementary and the Intermediate levels of the MAT to be anchored by the California Achievement Test (CAT). Scores on both levels of the MAT can be translated into CAT scores which can then be compared directly. Conversely, any CAT score can be translated into a score on both levels of the MAT. When the same CAT score is translated into a standard score for both the Intermediate and the Elementary levels of the MAT, the Intermediate level's standard

---

*Students are said to have scored at the chance level if their raw scores are no higher than the scores that they would have received had each multiple choice question been answered at random.

score is consistently higher (Linn). This suggests that the standard score transformation has a bias beyond that caused by the chance score transformation.

The difference of 2.3 standard score units between the Primary II and Intermediate levels of the MAT is larger than we would expect since both tests were administered in November/December. As previously stated, the results of tests more than one level apart probably should not be compared.

We conclude that for students in fifth grade, the Elementary level of MAT is not interchangeable with either the Primary II or the Intermediate. Moreover, the Primary II and the Intermediate levels of the MAT are not interchangeable.

For students in sixth grade, as well as those in fifth, the Elementary and Intermediate levels of the MAT do not produce equivalent results. As we have previously noted, a test administered during the November/December study is expected to have a higher mean score than a test administered in October. In sixth grade, the mean score for the Elementary level of the MAT is not higher than that of the October administration of Form G of the Intermediate level. Our discussion above of the biases in the standard score transformations for these two levels is a possible explanation of this result. These biases are also the probable cause of the difference of 1.7 standard score units between the mean scores for the Elementary level and Form F of the Intermediate level. These two tests were both administered in November/December, and their mean scores are not expected to be very different.

The difference of 1.5 standard score units between the two parallel forms, F and G, of the Intermediate level is about what we expect. We attribute the difference to growth.

We conclude that for students in sixth grade, the Elementary level is not interchangeable with either Form F or G of Intermediate level. As

in fifth grade, the results of the Intermediate level are consistently higher than the results of the Elementary level. However, the two forms, F and G, of the Intermediate level do seem interchangeable.

Results: Part II

Looking at the means of the different levels of the MAT test does not give a complete answer to the question of whether or not a student's score on one level of the MAT can be generalized to other levels. It is possible for the means of different levels of the achievement test to be quite different and yet for coefficients of generalizability or coefficients of reliability to be high. This would imply that even though the means differ from level to level, the relative order or rankings of the students would remain the same. One measure of the stability of the ranking of the student is the coefficient of generalizability, ρ. (Cronbach, et al., 1972)

The coefficients of generalizability and standard errors of measures are presented in Table 3 (coefficients for raw scores are listed for information purposes only and shall not be used to judge the tests' generalizability). The coefficients of generalizability for fifth and sixth grade indicate that the relative order of the students is preserved in fifth and sixth grade. The lower coefficient in fourth grade could be caused by changes in relative order of students either between the Primary I and the other two levels, Primary II and Elementary, or among all three levels. If the coefficient for fourth grade is calculated only for the two higher levels--Primary II and Elementary--a different picture emerges. Table 4 lists the coefficients of generalizability and the standard errors of measure for only the higher two levels administered in each of the three grades. In fourth grade, the coefficients of generalizability have increased which indicates that the Primary II and Elementary levels of the MAT preserve the relative order of the students. The coefficients for fifth and sixth grade are about the same in Table 4 as they are in Table 3.

Table 3

COEEFICIENTS OF GENERALIZABILITY (ρ) AND STANDARD ERRORS OF MEASURE (S.E.M.) FOR THE THREE LEVELS OF THE MAT GIVEN IN EACH GRADE

	Raw Scores	Standard Scores	Grade Equivalents
Fourth Grade			
ρ	.58	.58	.56
S.E.M.	14.4	7.0	.70
Fifth Grade			
ρ	.74	.79	.76
S.E.M.	18.0	6.1	.73
Sixth Grade			
ρ	.81	.77	.81
S.E.M.	14.5	5.9	.64

Table 4

COEFFICIENTS OF GENERALIZABILITY (ρ) AND STANDARD ERRORS OF MEASURE (S.E.M.) FOR THE TWO HIGHEST LEVELS OF MAT GIVEN IN EACH GRADE

	Raw Scores	Standard Scores	Grade Equivalents
Fourth Grade			
ρ	.84	.80	.78
S.E.M.	11.6	4.7	.50
Fifth Grade			
ρ	.76	.78	.80
S.E.M.	13.6	6.4	.69
Sixth Grade			
ρ	.82	.80	.84
S.E.M.	16.2	5.7	.61

Study 2: Test Administrators

In each of the academic years during which the achievement testing has occurred, there have been different test administrators. During the first year of the Voucher Demonstration (1972-73), all students were tested by the district's classroom teachers, who were sometimes the student's own teacher and sometimes not. In the fall of the second year (1973-74), the MAT was administered to students by either their own classroom teacher, another teacher from within the same school, a member of the district's evaluation staff, or by a person registered with the Alum Rock School District as a substitute teacher. In the spring of the second year, the situation changed and only classroom teachers were used as test administrators. Spring 1974 was also the first time that teachers were given in-service training in how to administer standardized achievement tests.

In the third year of achievement testing (1974-75), an entirely new form of administration was adopted. A group of approximately 25 people, all of whom were registered as substitute teachers within the district, were selected for special training in administering standardized achievement tests. Following a four-day, intensive training program, the substitute teachers administered all achievement tests that were given in the district during 1974-75.

Since students in Alum Rock had been tested under so many different modes of test administration during the three years 1972-73, 1973-74, and 1974-75, we thought it was important to determine whether or not the mode of administration had any effect upon the students' achievement scores.

Design

To determine the effects of different types of test administrators, students from the second and third grades were given the reading portion of the MAT by the three types of administrators that had been most frequently used in the first three years. Substitute teachers registered

in Alum Rock, but not familiar to the students, were one type of test administrator. Under the supervision of the school district, the substitute teachers had been given intensive training in how to administer the MAT. Hence, their method of test administration was quite uniform. A second type of test administrator included in the study was the student's own classroom teacher. Teachers in the same school who were not the student's regular classroom teacher were the third type of administrator. Both groups of classroom teachers had received a minimal amount of instruction in how to administer the MAT.

The subjects for this study were four second grade and four third grade classes. Two classes from each grade were in one elementary school and two classes from each grade were in another. Based on the district's mean score, the four second and four third grade classes seemed representative of the district's reading achievement. The district's mean scores for the MAT composite Total Reading, in standard score units, were 37.8 and 47.6 for second and third grade, respectively. The second and third grade classes in the experiment had mean Total Reading scores of 38.4 and 49.4, respectively.

In October 1974, all students in the Alum Rock district were given the MAT administered by the trained substitute teachers. As with Study 1, we had initially planned for the study of the effects of test administrators to begin within two weeks of the October testing. A two week interval would have allowed the October testing to be viewed as the first of three repeated measures. However, as with Study 1, internal concerns in the district delayed the study until the last week in November. The third test was administered a week after the second MAT administration.

The eight classes were tested a total of three times as shown in Table 5. Only the reading portion of the MAT was used during the second and third administrations. Of the 176 students who were in our study, half (88) of the students chosen at random were tested the second time by

Table 5

DESIGN OF TEST ADMINISTRATION EXPERIMENT

Group	Pattern of Test Form and Administration						n
	October 1974		First Retest		Second Retest		
	Form	Administration	Form	Administration	Form	Administration	
I	F	Subst. Teacher	F	Own Teacher	G	Other Teacher	45
II	F	Subst. Teacher	F	Other Teacher	G	Own Teacher	45
III	F	Subst. Teacher	G	Own Teacher	F	Other Teacher	43
IV	F	Subst. Teacher	G	Other Teacher	F	Own Teacher	43
							Total: 176

their own classroom teacher and the third time by a teacher other than their own. The other 88 students were tested the second time by a teacher other than their own classroom teacher, and the third time by their own classroom teacher.

During the second and third test administration, approximately half of the students within each classroom, chosen at random, were given Form F of Primary I, and half of the students were given Form G. These alternate forms have been prepared by the MAT publishers to avoid "learning effects" which might occur when students are repeatedly tested at the same level. An odd number of children in some of the classrooms caused 90 of the students to be given Form F during the second test administration and then given Form G during the third administration. The remaining 86 students were first given Form G and then given Form F.

In order to minimize the error variance by standardizing testing conditions, all classes were tested the same day of the week at the same time of the day by all three types of administrators.

Results: Part I

To determine whether or not the type of test administrator affected a student's performance, we compared the mean scores for the different types of administrators. The Total Reading* mean scores and their standard deviations are presented in Table 6. For the same reasons discussed in Study 1, if the type of administrator were unchanged, we would expect that growth would cause the mean scores for the November/December test to be about 1.5 standard score units higher than the October results. We would also expect that the two forms of the MAT administered one week apart in November and December would have approximately the same mean scores.

*For economy, the analyses done in Study 2 will be done only for the composite Total Reading Score.

Table 6

MEANS FOR DIFFERENT TEST ADMINISTRATORS

(N=176)

	October	November/December Retests	
	Substitute Teacher	Regular Teacher	Other Teacher
Raw Score (Standard Deviation)	54.4 (18.8)	57.2 (19.1)	55.7 (18.3)
Standard Score (Standard Deviation)	44.5 (11.9)	47.2 (13.0)	46.3 (12.5)
Grade Equivalents (Standard Deviation)	2.39 (0.82)	2.59 (0.96)	2.42 (0.91)

Both of the retests have higher mean scores than tests administered in October. The test administered by the other teacher has a mean score 1.8 standard score units higher than the same test administered in October by the substitute teacher. This difference is about what we could attribute to growth. The mean score for the test administered by the students' regular classroom teacher is 2.7 standard score units higher than the October mean score. This difference is somewhat higher than we expected, but not enough to be considered educationally significant. The difference of 0.9 between the mean scores for tests administered by students' own teachers and other teachers is small, especially in light of the standard deviations. The differences in mean scores do not allow us to conclude that the type of administrator has an effect upon the students' performance. Although the students' growth is confounded with type of administrator, (substitute teachers only administered tests in October), in our opinion, the differences in mean scores are caused by either growth or random variation and not by type of administrator.

Results: Part II

Looking at the means of the various types of test administration does not give a complete answer to the question of whether or not a studcnt's score under one type of administrator can be generalized across the other types of administrators. It is possible that the means of the three types of test administration would not be equal and yet the coefficient of generalizability (ρ) might be very high (see Study 1).

A high coefficient of generalizability would mean that the relative standings of the students remained unchanged. To investigate this question, we calculated estimates of the coefficients of generalizability and standard errors of measure. These are presented in Table 7. The coefficients show that the type of test administrator does little to effect the relative standings of the students. These results are quite similar to the results of the previous third grade study (Barker and Pelavin).

Table 7

COEFFICIENTS OF GENERALIZABILITY (ρ)
AND STANDARD ERRORS OF MEASUREMENT (S.E.M.)

	Raw Score	Standard Score	Grade Equivalent
ρ	.88	.84	.77
S.E.M.	6.6	5.1	.4

Conclusions

At the beginning of this paper, two general questions were posed. Those questions were: What are the effects of using different levels of the MAT within the same grade and what are the effects of the MAT being administered by different types of teachers. We believe both these questions have been answered. The level of the MAT administered to a student can have a substantial effect on the student's score. The standard score scale does not permit a student's score on one level of the MAT to be generalized to other levels, even adjacent levels. However, in most cases, the standard score scale does preserve the relative order of students from one level to another. These results suggest that the level of the MAT used in Alum Rock did affect the evaluation results. This is particularly true for fifth grade, where the Elementary level of the MAT was administered to most students. Our results show that had these students been given the Intermediate level (the level standardized for fifth grade), their mean score would probably have been higher.

In contrast to the findings of the first study, the type of test administrator seems to have had little effect on the students' scores. A student's score on a test administered by one type of teacher seems to generalize across the other two types. Hence, we believe that the use of different test administrators in Alum Rock did not affect the evaluation results.

References

Barker, P., Issues in Measuring Student Outcomes, The Rand Corporation, R-1497 (1975).

Barker, P., and S. Pelavin, Within-Subject Investigations of the Accuracy of Scaled Scores in Standardized Achievement Tests: The Case of MAT 70, The Rand Corporation, WN-9161-NIE (1975).

Cronbach, L. J., Essentials of Psychological Testing, Third Edition (Harper and Row, 1970).

Cronbach, L. J., G. Heser, H. Nanda, and N. Rajaratnan, The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles (John Wiley and Sons, Incorporated, 1972).

Durost, W. N., H. Bixler, W. Wrightstone, G. Prescott, and Balow, Metropolitan Achievement Test (Harcourt Brace Jovanovich, 1971).

Linn, Robert L., "Anchor Test Study Review," Journal of Educational Measurement, Volume 12, pp. 201-213 (1974).