DOCUMENT RESUME

ED 334 267 TM 016 899

AUTHOR Lanese, James F.
TITLE Test Familiarity: Evidence of "Practice Effects"?
PUB DATE Apr 91
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cohort Analysis; Comparative Testing; *Elementary School Students; *Equated Scores; Grade 4; Grade 6; Grade 8; Intermediate Grades; Junior High Schools; *Reading Tests; School Districts; *Standardized Tests; Testing Problems; Testing Programs; *Test Wiseness
IDENTIFIERS *California Achievement Tests; Cleveland Public Schools OH; *Parallel Test Forms

ABSTRACT

A study involving all students taking the California Achievement Test (CAT) in the Cleveland City School District (Ohio) during the fall of 1989 was conducted to assess the effects of the use of parallel test forms. Each spring, all students in grades 4, 6, and 8 are selected to take the appropriate level of the CAT reading test, Form E. As a result, the question of test familiarity was raised. To address this question, Form F of the CAT battery was selected for administration in the fall of 1989. The sample population included 4,367 fourth, 3,768 sixth, and 2,770 eighth graders. Data from each annual testing from the spring of 1988 through the spring of 1990 were compared with the fall 1989 data. Normal curve equivalent reading scores were used to identify differences between CAT-E and CAT-F scores. To determine whether scores differed significantly for the CAT-E versus the CAT-F, a regression discontinuity design was used to investigate the trends in reading achievement evident for each cohort. Results refute the notion that repeated use of the same form of the test improves scores through "practice effects." The alternative form of the test used during the same academic year provided comparable, or higher, reading achievement levels when used for the first time in the school district. Six figures and one table are included. (TJH)

Test Familiarity: Evidence of "Practice Effects"?

James F. Lanese

Cleveland Public Schools

Paper presented at the annual meeting of the Amertican Educational Research Association in Chicago, Illinois, April, 1991.

Test Familiarity: Evidence of "Practice Effects"?

Introduction

The recent attention turned towards standardized achievement testing on a national level has included the issues of test design adequacy, extent of testing, the use of test results and their validity over the period of time the test is in use. Additionally, the escalation of test related accountability issues has initiated close scrutiny of testing parctices among schools and teachers including charges of unethical behavior in the preparation of students for the testing experience.

One element of the test design involves the utilization of parallel forms of the test to provide an optional form for use in a testing program. Another element of the parallel form issue involves the use of a test's alternate form as a method of preparation for a testing experience. This study investigates the measured achievement performance of students participating in a testing program which includes the use of both forms of a standardized test.

Lake Woebegon and Ethical Testing Practices

In 1987 and again in 1989, John J. Cannell published

reports which asked serious questions about the use of, validity of, and reliance upon standardized achievement test results by school districts and states throughout the nation. Considerable reaction to these reports has fostered numerous studies and reviews which have criticized the attacks while admitting that test results could be used to mislead the public. Highlights of the issues and practices related to the use of parallel forms of published achievement tests are noted below.

Parallel Form Testing

Parallel (or equivalent) forms of tests are made available by publishers in order to "allow teachers to retest students at different times to evaluate progress without having them retake the same test or to test students in different sections of a class without being concerned about test security" (Sax, 1974). Typically, alternate form reliability studies are cited by the publisher in order to support the notion of equivalency of measure despite the form being employed. Additionally, Sax (1974) notes that the use of alternate forms makes it difficult for a student to recognize items from a previously administered test or to help another person planning to take the same test. Although parallel forms serve these purposes, in practice they are seldom used. School districts who face significant cost increases to purchase two sets of a test in order to

implement a parallel form testing program typically employ alternate forms only for specialized or specific situations.

Recent criticisms of local and state reliance upon standardized achievement test measures have included allegations of various test preparation practices utilized in schools. Koretz (1988) indicated that "the vast gray area of teaching to the test stretches from frank cheating at one extreme to appropriate remediation and instruction at the other. Both educators and educational researchers disagree strenuously about where the line between appropriate and inappropriate teaching should be drawn." In a response to Dr. Cannell's allegations (1987), Phillips and Finn (1988) cite the

> practice (among schools) of continually using the same test for many years within the school system. Teachers become familiar over time with the test objectives, and students get accustomed to the item formats. In some cases, teachers may actually teach the specific items on the test. These "practice effects" greatly inflate the students' scores and give a misleading impression of achievement gains.

Test preparation activities were addressed by Mehrens and Kaminiski in 1989 whereupon they described a continuum of practices ranging from always ethical practices on one hand

to never ethical practices on the other. Seven categories included; 1)general instruction, 2) teaching test taking skills, 3) instruction by objectives not matched to the test, 4) instruction by objectives matched to the test, 5)instruction on matched objectives using the test format, 6) practice on a published parallel form of the test, and 7) practice on the same test. The authors concluded that the point where one crosses over from legitimate to illegitimate practice "must be somewhere between (3) and (5)." While the authors would indicate that the use of parallel forms of tests for practice is at the fraudulent end of the continuum, the use of the parallel form of a test as an alternate measuring instrument remains consistent with the parallel testing design.

Beyond the issues of test preparation activities and their appropriateness, lies the issue of test familiarity. In a second inquiry into the issues raised by Dr. Cannell, Gary Phillips (1990) restated the potential impact of test familiarity. "As the test is repeatedly used, the school system becomes more and more familiar with the test content and format. Again, this gives the test user an advantage not shared by the norming sample." In a related discussion, Shepard (1990) addresses the test familiarity question. Shepard speculates that "test familiarity might allow teachers to improve the performance of their students innocently, without consciously deciding to cheat, by

xeroxing a copy of the test." The author indicates that a teacher who simply remembers the nature of a test item and adjusts his or her instruction towards that item could serve to influence a child's percentile score by two to seven points. This phenomenon could easily, even innocently, occur when a particular test is used for several years in a school system.

Background for the Study

The alternate form scheduling approach has not been utilized in the Cleveland City School District since the adoption of the California Achievement Test in 1987. Annually, the reading subtest (form E) of the CAT has been used at all grade levels within the district. Each spring, all test eligible pupils in the district are selected to take the appropriate level of the CAT reading test. In light of this factor, a question of test familiarity was raised concerning its potential impact upon achievement measures.

During the 1989-90 academic year, the District initiated a separate testing program at grades four, six, and eight in order to conform to a State achievement/ability testing program. Consequently, form F of the CAT battery was selected for this administration in the fall of 1989. The availability of these scores enabled the district to address

the questions stated above.

Purpose of the Study

The purpose of this study was to explore the impact of evidence of test familiarity upon obtained achievement scores in reading.

Specifically, the following questions were addressed.

1. Did significant differences exist between CAT-E reading achievement scores and CAT-F reading achievement scores for students at various grade levels?

2. Did student reading achievement trends differ significantly when measured by CAT-E or CAT-F tests?

Methodology

Sample

The sample for this study included all students who were tested (grades four, six, and eight) in the program in fall, 1989. This group formed the basis for the derivation of a cohort at each grade level. Reading test scores for these students were then selected from spring, 1988, spring, 1989

and spring, 1990. Those students who had available scores for all four data points were included in the sample. The first cohort contained 4,362 members who were fourth graders in 1989-90. The second cohort contained 3,768 members who were sixth graders in 1989-90 and the third cohort contained 2770 students who were in the eighth grade in 1989-90. Between three and five percent of each cohort membership repeated a grade during the period included in the study; these students took the same level and form of the reading test during their retention year.

Table 1.

Grade Levels of Cohort Members

Cohort	SPring88 GRD	SPring89 GRD	Fall89 GRD	SPring90 GRD	n
1	2	3	4	4	4362
2	4	5	6	6	3768
3	6	7	8	8	2770

Data Analysis and Results

Normal curve equivalent reading scores were included for data analyses to address the first question. Specifically, each cohort group's mean reading comprehension NCE was computed for the four testing sessions included in the study. A set of correlated t-tests between test sessions one and two, two and three, three and four, and two and four were conducted. Normal Curve Equivalent scores were chosen since the nationally norm referenced test was administered on level and scored during the empirical norming period for all testing sessions. Due to these factors, the mean NCE score for each cohort would be expected to remain consistent from one session to the next. Correlation coefficients would also be expected to remain high for all pairs.

The results illustrated in the following tables are discussed below.

The first cohort (grade four,1989-90) evidenced an inconsistent achievement pattern from spring, 1988 to spring, 1989 to spring, 1990 on the CAT form E. Mean NCE's increased by three points then dropped four points over the three spring measures. All means differed significantly ($p < .01$). However, the fall, 1989 measure obtained by administering form F of the CAT was consistent with the spring, 1990 form E measure and indicated no significant difference with that measure.

F89GRD: 4 COHORT 1

- - - T-TESTS FOR PAIRED SAMPLES - - -

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S88NCE	SPRING 88 OBTAINED SCORE			
	4362	46.6862	17.809	.270
	4362	49.6130	16.355	.248
S89NCE	SPRING 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	I I	CORR.	2-TAIL PROB.	I I	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-2.9269	15.701	.238	I	.580	.000	I	-12.31	4361	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S89NCE	SPRING 89 OBTAINED SCORE			
	4362	49.6130	16.355	.248
	4362	44.7334	15.893	.241
F89NCE	FALL 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	I I	CORR.	2-TAIL PROB.	I I	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
4.8796	13.682	.207	I	.640	.000	I	23.56	4361	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S89NCE	SPRING 89 OBTAINED SCORE			
	4362	49.6130	16.355	.248
	4362	45.1729	16.309	.247
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	I I	CORR.	2-TAIL PROB.	I I	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
4.4402	13.876	.210	I	.639	.000	I	21.13	4361	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
F89NCE	FALL 89 OBTAINED SCORE			
	4362	44.7334	15.893	.241
	4362	45.1729	16.309	.247
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	I I	CORR.	2-TAIL PROB.	I I	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-.4395	12.080	.183	I	.719	.000	I	-2.40	4361	.016

Figure 1. Achievement score comparisons for Cohort 1.

- - - T-TESTS FOR PAIRED SAMPLES - - -

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S88NCE	SPRING 88 OBTAINED SCORE			
	3768	47.3992	14.786	.241
	3768	47.6746	14.683	.239
S89NCE	SPRING 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-.2755	10.724	.175	.735	.000	-1.58	3767	.115

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S89NCE	SPRING 89 OBTAINED SCORE			
	3768	47.6746	14.683	.239
	3768	53.5387	15.840	.258
F89NCE	FALL 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-5.8641	11.116	.181	.737	.000	-32.38	3767	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S89NCE	SPRING 89 OBTAINED SCORE			
	3768	47.6746	14.683	.239
	3768	48.3129	16.604	.270
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-.6383	11.772	.192	.725	.000	-3.33	3767	.001

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
F89NCE	FALL 89 OBTAINED SCORE			
	3768	53.5387	15.840	.258
	3768	48.3129	16.604	.270
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
5.2258	11.554	.188	.747	.000	27.76	3767	.000

Figure 2. Achievement score comparisons for Cohort 2.

F89GRD: 8 COHORT 3

- - - T-TESTS FOR PAIRED SAMPLES - - -

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S88NCE	SPRING 88 OBTAINED SCORE			
	2770	50.7614	14.410	.293
	2770	46.8354	13.865	.263
S89NCE	SPRING 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
3.9260	10.405	.198	.752	.000	19.86	2769	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S89NCE	SPRING 89 OBTAINED SCORE			
	2770	46.8354	13.865	.263
	2770	44.6383	17.515	.333
F89NCE	FALL 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
2.1971	11.610	.221	.750	.000	9.96	2769	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
S89NCE	SPRING 89 OBTAINED SCORE			
	2770	46.8354	13.865	.263
	2770	45.0430	17.250	.328
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
1.7924	12.738	.242	.685	.000	7.41	2769	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
F89NCE	FALL 89 OBTAINED SCORE			
	2770	44.6383	17.515	.333
	2770	45.0430	17.250	.328
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-.4047	13.503	.257	.698	.000	-1.58	2769	.115

Figure 3. Achievement score comparisons for Cohort 3.

The second cohort (grade six, 1989-90) evidenced a consistent pattern of achievement (means within one NCE point) over the three spring test sessions. Fall testing of this cohort evidenced a six NCE point (significant) rise from the previous form E session.

Finally, the third cohort (grade eight, 1989-90) realized a four and two NCE point decline respectively (both significant) on three spring Form E measures. Fall (form F measures also showed the two point significant decline and a non-significant change to the subsequent spring measure.

Although the assumption concerning consistency of the mean score over time was not demonstrated by these data, the consistency of fall (form F) to spring (form E) means among the first and third cohort members illustrates little apparent impact of the use of the different form of the test.

To address the second question, a regression discontinuity design was utilized to investigate the trends (implicit in the above analysis) in reading achievement evident for each cohort. Spring, 1988 NCE scores were plotted with spring, 1989 NCE scores to build a linear prediction equation for each of the cohorts. The equations were then used to calculate a predicted NCE score for each cohort member for the 1989 academic year. The predicted scores were then compared to the obtained fall (form F) and

spring (form E) scores to determine the differences. The illustrated t-tests and discussion follow.

The first cohort evidenced similar significant differences in measure between both forms of the test and the predicted scores. In both comparisons, obtained scores were 1.2 and 1.6 NCE points lower than predicted levels.

The second cohort showed a wider (significant) difference between the two measures and the predicted scores. Form F measured reading 6.3 NCE points higher than predicted while form E measured reading achievement 1.2 points higher than predicted.

The third cohort, like the first, indicated lower obtained scores on both forms than predicted (6.5 and 6.1 respectively. Again, the differences were both significant.

With the exception of the second cohort, the alternate forms of the test indicate similar patterns of achievement (as might have been anticipated from the first analysis).

Conclusions

The review of the achievement patterns over time of the three cohorts in this study indicated that in two of the three cases, the alternate form of the test yielded no

F89GRD: 4 COHORT 1

- - - T-TESTS FOR PAIRED SAMPLES - - -

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
ERCNCE	PREDICTED SCORE			
	4362	46.3554	10.336	.157
	4362	44.7334	15.893	.241
F89NCE	FALL 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
1.6221	12.208	.185	.640	.000	8.78	4361	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
ERCNCE	PREDICTED SCORE			
	4362	46.3554	10.336	.157
	4362	45.1729	16.309	.247
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
1.1826	12.544	.190	.639	.000	6.23	4361	.000

Figure 4. Predicted vewrsus Obtained Achievement scores for Cohort 1.

F89GRD: 6 COHORT 2

- - - T-TESTS FOR PAIRED SAMPLES - - -

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
ERCNCE	PREDICTED SCORE			
	3768	47.2792	10.865	.177
	3768	53.5387	15.840	.258
F89NCE	FALL 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-6.2595	10.733	.175	.737	.000	-35.80	3767	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
ERCNCE	PREDICTED SCORE			
	3768	47.2792	10.865	.177
	3768	48.3129	16.604	.270
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR.	2-TAIL PROB.	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
-1.0337	11.522	.188	.723	.000	-5.51	3767	.000

Figure 5. Predicted vewrsus Obtained Achievement scores for Cohort 2.

F89GRD: 8 COHORT 3

- - - T-TESTS FOR PAIRED SAMPLES - - -

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
ERCNCE	PREDICTED SCORE			
	2770	51.1544	11.591	.220
	2770	44.6383	17.515	.333
F89NCE	FALL 89 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	I I	CORR.	2-TAIL PROB.	I I	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
6.5161	11.690	.222	I	.750	.000	I	29.34	2769	.000

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
ERCNCE	PREDICTED SCORE			
	2770	51.1544	11.591	.220
	2770	45.0430	17.250	.328
S90NCE	SPRING 90 OBTAINED SCORE			

(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	I I	CORR.	2-TAIL PROB.	I I	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
6.1114	12.573	.239	I	.685	.000	I	25.58	2769	.000

Figure 6. Predicted vewrsus Obtained Achievement scores for Cohort 3.

significant differences than the subsequent administration of the test's other form. This fact was not evident among the second (sixth grade) cohort members in the study. These results among the first and third cohorts (as well as the higher achievement measure obtained by the second cohort) refute the notion that repeated use of the same form of the test improves scores through "practice effects."

Concerning reading achievement trends, again, two of the three cohorts (the first and third) demonstrated significant but comparable differences between the predicted and measured level of reading achievement. The middle cohort evidenced a considerably higher obtained than predicted reading level using this analysis. As above, the evidence does not support the contention that repeated use of a test contributes to improved scores.

It appeared that alternative form test utilized during the same academic year provided comparable (or higher) reading achievement levels when utilized for the first time in the school district.

References

Cannell, John J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are testing above the national average. Educational Measurement: Issues and Practices. 7 (2), 5-9.

Cannell, John J. (1988). How public educators cheat on standardized achievement tests. An invited presentation to the annual meeting of the National Council for Measurement in Education. Boston: April, 1990.

Koretz, Daniel. (1988, Summer). Ariving in lake wobegon. Are standardized tests exaggerating achievement and distorting instruction? American Educator. 8-14, 46-52.

Mehrens, William A. and Kaminski, John. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent. Educational Measurement: Issues and Practices. 8 (2), 14-22.

Phillips, Gary W. and Finn, Chester E. (1988). The lake wobegon effect: A skeleton in the testing closet? Educational Measurement: Issues and Practices. 7 (2), 10-12.

Phillips, Gary W. (1990). The lake wobeogon effect. Educational Measurement: Issues and Practices. 9 (3), 3.

Sax, Gilbert. (1974). Principals of educational measurement and evaluation. Belmont, California: Wadsworth Publishing Co.