DOCUMENT RESUME

ED 076 657 TM 002 664

AUTHOR	Eichelberger, R. Tony
TITLE	Effects of Repeated Standardized Testing on Different Types of Students.
PUB DATE	73
NOTE	13p.; Paper presented at American Educational Research Association Meeting (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Grade 6; *Intelligence Tests; Measurement Instruments; Speeches; *Standardized Tests; *Student Characteristics; *Student Testing; Testing; *Test Results
IDENTIFIERS	*Otis Lennon Mental Abilities Test

ABSTRACT

The effects of repeated I.Q. testing were investigated to ascertain the necessity of constructing and using alternate test forms. There were also attempts made to describe selected individual characteristics of subjects who improved the most over the repeated testing. One hundred and forty-five students were tested at one month intervals for three months. Two forms of the Otis-Lennon Mental Abilities Test were used in a counter-balanced design. The total group improved only from the first to second testing session. Persons repeating the same form did significantly better than persons taking alternate forms over the same testing sessions. It appeared that the students did tend to remember items from testing session one to testing session two, but this trend did not hold into testing session three. In general, the mean scores tended to decrease from testing session two to testing session three. Persons who appeared to improve most were from the upper class, or girls, or had relatively high grade point averages. (Author)

EFFECTS OF REPEATED STANDARDIZED TESTING ON DIFFERENT TYPES OF STUDENTS

by

R. Tony Eichelberger

Learning Research and Development Center

University of Pittsburgh

ABSTRACT

A standardized intelligence test was administered three times to 145 sixth-grade students at monthly intervals. A number of variables, such as social status, sex, grade-point average (GPA), initial IQ test score, and test-wiseness, were used to predict the resulting changes in IQ test scores. Overall, the scores increased from the first to second testing session, but decreased from the second to third. The types of students whose scores increased were girls either from the middle or upper class or with a relatively high fifth grade GPA. The conditions under which most studies of repeated testing occur usually include some type of motivational technique. In order to obtain more reliable estimates of the effects of variables related to standardized testing, it seems imperative that some specific testing conditions need to be adopted.

A review of previous research suggests that test scores on the average increase when a standardized instrument is repeatedly administered (Kreit, 1968; PMA, 1968; Dearborn and Rothney, 1941; Peel, 1952). Present measurement theory (Thorndike in Lindquist, 1951) assumes that much of this increase is due to remembering specific test items. Many studies on repeated testing (PMA, 1958; Heim and Wallace, 1949; Kreit, 1968; Droege, 1966) report results that are similar when either the same form or alternate forms of a test are used, which would cast doubt on this assumption.

A number of studies have been reported which investigate the effects of repeated testing with standardized instruments. The majority of these studies indicate that scores for individuals fluctuate (Thoulass, 1936), but the mean score for the group increases significantly from the first to second administration of the test (PMA, 1968; Dearborn and Rothney, 1941). These group gains tend to decrease for each subsequent administration, and are usually found to be non-significant after the second testing session (Peel, 1952; Kreit, 1968). In previous experiments either all subjects were given the same form, or all subjects were given alternate forms of the testing instrument. The relative effect of remembering specific test items could not be investigated with these designs.

The purpose of the present study was to investigate the assumption that remembering specific test items is a major determinant of increases in test scores resulting from repeated testing. In addition, an attempt was made to identify the types of students who improve most over repeated testing.

METHOD

Sample

All sixth grade students of a rural school district in southeastern Missouri were selected for the study. The school qualified for Title I funds with over 50% of the students' families receiving some form of welfare aid.

Procedure

All students were given identifying numbers which were selected randomly from a bowl to make up six groups, each composed of approximately 28 students. The six groups were administered two forms of the Otis-Lennon Mental Ability Test as indicated in Table 1. Since members of each group came from each of six classrooms tested, the effects of testing conditions and test administration differences were minimized.

As the design indicates, three of the groups took Form J first and three groups took Form K first. The design was balanced so the effects of Forms J and K were taken into account. By administering the same form to some students each time, and alternate forms to other students, the effect of remembering specific test items could be investigated.

Table 1

Procedure for the Administration of the Different Forms of the Otis-Lennon Mental Ability Test

Group	Form Taken by Testing Session		
	TS 1	TS 2	TS 3
1	J	J	J
2	J	J	K
3	J	K	J
4	K	K	K
5	K	K	J
6	K	J	K

The Index of Social Status (ISS), developed by McGuire and White (1955), was used to compute the social status of each student. The ISS provides a score which is a weighted sum of ratings on the occupation, source of income, and education of the students' parents. The scores can range from 12 (high) to 84 (low).

The Otis-Lennon Mental Ability Test is composed of 80 multiple choice items to be completed in 40 minutes. General purpose answer sheets were used with the test booklets. The test was administered at monthly intervals (March 18, April 15, and May 13).

RESULTS

To investigate the primary question concerning the effect of remembering specific test items, both the general, or overall, practice effect and the differential effect of the same form and alternate forms of the test were analyzed. Raw scores were used as the criterion scores.

General Practice Effect

There was a significant increase from the first to second testing session, but not from the second to third nor from the first to third testing session as indicated in Table 2. This result agrees with previous results, suggesting that a practice effect occurs quickly and then seems to disappear.

Table 2

Results of Three IQ Testing Sessions with all Subjects

		Mean Difference Scores		
Testing Session	Mean Score	TS2-TS1	TS3-TS2	TS3-TS1
1	43.76	2.83*		
2	46.59			.74
3	44.50		-2.10	

*Significant at .01 level using correlated t-tests (one-tailed tests, df = 144)

Practice Effect of Different vs. Same Form of Test

As indicated in Table 3, students who took the same form during the first two administrations increased significantly more than students who took different forms. Multiple linear regression procedures (Kelly, Beggs, McNeil, Eichelberger, and Lyon, 1969) were used in doing the statistical analyses. To test the effect of differences in test form the following models were used.

$Y = a_0U + a_1 X_1 + a_2 X_2 + E_1$, where

Y = Gain scores on Otis-Lennon Mental Ability Tests.

X_1 = 1 if person took the same form of the test on both testing sessions in question, 0 otherwise.

X_2 = 1 if person took different forms of the test on the two testing sessions in question, 0 otherwise.

U = Unit vector

a_0, a_1 and a_2 = Least-square weights

E_1 = Error vector $(Y - \hat{Y})$

The hypothesis that the two groups were from the same population was tested by assuming $a_1 = a_2$. The resulting equation in this case was:

$$Y = a_0U + E_2$$

The proportion of criterion variance accounted for by the predictor variables in each case were compared. The results of these tests are reported in Table 3.

Table 3

Mean Gain Scores for Students Taking the Same Form vs. Students Taking Alternate Forms of IQ Test During the Two Testing Periods

Testing Period	Same Form (N=99)	Different Forms (N=46)	F	df_1	df_2
TS2-TS1	4.05	.15	13.2*	1	143
TS3-TS2	-3.35	-1.42	1.8	1	143
TS3-TS1	.75	.71	.01	1	143

*Significant at .01 level (one-tailed).

Analysis of Response Patterns to the Items

The response patterns of all students who took the same test form during the first two testing sessions (TS1 and TS2), which was the only time a significant change occurred, were analyzed in the following manner. Since nearly all students answered all 80 items during both testing sessions, only the following four combinations were studied: (1) answered correctly both times (++), (2) answered incorrectly both times (--), (3) answered correctly during TS1 and incorrectly during TS2 (+-), and (4) answered incorrectly during TS1 and correctly during TS2 (-+). The results are broken down in Table 4 to high social status (HSS) and low social status (LSS) for both students whose scores increased on TS2 and those whose scores did not. Because the students tended to have low social status, the dividing point for high and low social status was arbitrarily set between scores of 62 and 63 on the ISS. McGuire and White (1955) report that these scores are indicative of lower middle class status.

Table 4

Mean Item Responses for Increasing and Non-Increasing Subjects of High and Low Social Status.

	Testing Period		Increasing		Non-Increasing	
	TS1	TS2	HSS (N=27)	LSS (N=49)	HSS (N=8)	LSS (N=15)
Response Combinations	+	+	49.5	31.6	27.5	25.6
	+	-	5.7	7.2	13.5	12.8
	-	+	10.8	11.7	10.5	10.4
	-	-	11.9	23.3	27.6	29.2

As indicated in Table 4, the HSS increasing students (12-62) averaged 49.5 ++ items while LSS (63-84) students in this group averaged 31.6. On -- items HSS subjects averaged 11.92, while LSS subjects averaged 23.3. In both cases the increasing students with LSS did not score much differently from the non-increasing students. The overall differences between the increasing and non-increasing groups appear to be due primarily to increasing students with HSS.

Summary of Results Related to Remembering Specific Test Items

These results suggest that a practice effect occurs rapidly and then dissipates. The practice effect does appear to be due to students remembering specific test items, as students repeating the same form improved 4.05 items while students taking different forms improved on the average of only .15 items. Therefore, the assumption does appear to be supported.

At least two observations require tempering of this conclusion. First, there did not appear to be an adequate ceiling on the test used. One student answered all 80 items correctly during one testing session and 79 items during a second session. A number of students answered over 70 items correctly during all testing sessions. Thus, there was little opportunity for other test-taking skills, such as use of time or use of answer sheet, to be indicated as related to improvement. Second, nearly all previous studies indicate continuous improvement--even over 9 or 10 administrations (Heim and Wallace, 1949), while a decrease occurred from TS2 to TS3 in this study.

A possible explanation for these results, especially as they differ in many ways from previous results, is a motivational one based on the situation in which the tests were given. Almost no previous study

reported a decrease in average score from the second to third testing session. In nearly all previous studies some attempt was made to motivate the subjects to improve (e.g., volunteer subjects were used, rewards were given, directions were read which indicated improvement was expected, etc.). In the present study no such attempts were made. The students were told that the test scores would be given to the school administrators, but no indication was given either that the scores would count toward their grades, or that they would be retested. The person administering the tests simply indicated that he was interested in finding out what would happen to the scores, if he said anything at all. Only after the final administration were the children told that the testing was over. The teachers and administrators of the School System gave tremendous cooperation, and there appeared to be no hostility toward either the experimenter or the time taken from class. The students also left many more blank spaces on the third test than on the second one. During TS3 a number of students completed the first row on the answer sheet and then stopped. The same format had been used twice previously, so they knew how to carry out the task. Therefore, it appears that many students simply became bored with the task on the third testing (three IQ tests in two months). Other studies that follow must be concerned with this possibility.

Identifying Students Improving Most on the Tests

Six predictor variables were used to describe the participating sixth grade students. These variables were: (1) social status (ISS), (2) initial IQ test score, (3) fifth-grade grade point average (GPA), (4) sex, (5) test-wiseness, and (6) a moderator variable, (GPA/IQ)* ISS. The measurement of each variable has previously been described except for test-wiseness. This skill was measured by a 16-item instrument originally

devised by Slakter and Koehler (1969). The rationale for including these particular variables is indicated elsewhere (Eichelberger, 1970).

The independent contribution of each variable to the change in test score variance was investigated from TS1 to TS2, TS2 to TS3, and TS1 to TS3. The Multiple Linear Regression approach was used in the following manner to analyze the data.

$Y = a_0U + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + E_3$, where

Y = Change in test score

X_1 = ISS score

X_2 = Initial IQ test score

X_3 = Fifth-grade GPA

X_4 = Sex

X_5 = Test-wiseness score

X_6 = Moderator variable $(X_3/X_2)^* X_1)$

a_0 through a_6 = Least-square weights

E_3 = Error vector $(Y - \hat{Y})$

Each predictor variable was dropped in turn from the equation to test its independent contribution to the change in test scores during the different testing sessions.

RESULTS

The results indicated that GPA was the only significant ($\alpha<.05$) predictor from TS1 to TS2, while sex and social status (ISS) were significant predictors from TS2 to TS3 and from TS1 to TS3. The proportion of criterion variance accounted for by the predictor set in each case is indicated in Table 5.

Table 5

Results of Predicting Test Score Change by 6 Predictor Variables

Testing Period	R^2	Significant Predictors	Obtained p	df
TS2-TS1	.0656	GPA	.013	1,138
		SEX	.003	1,138
TS3-TS2	.1187	ISS	.037	1,138
		SEX	.001	1,138
TS3-TS1	.1447	ISS	.008	1,138

Note: Significance level: $\alpha<.05$, N = 145

Although only a small proportion of the change in test score was predicted, some inferences concerning the types of students most likely to improve on repeated testing are possible. Further observation and manipulation of the data indicated that girls with middle or upper status, or with a relatively high GPA were most likely to improve on repeated IQ testing.

Again, these results may be peculiar to the situation in which this study was done. Perhaps upper class girls are more willing to persevere when given an apparently boring task. Also, these results might not replicate when periods between testing sessions are more like that which normally occurs within a school, i.e., a full year. But numerous theories would lead one to expect that sixth-grade girls from middle or upper class families, or with a history of high academic achievement, would be most likely to concentrate more and work harder to score well on standardized tests given within their schools. Therefore, the results from this study would tend to support these ideas, or theories.

SUMMARY

In summary, the data reported tend to support the assumption that students remember specific test items--at least for a period of one month. Other test-taking skills did not appear to be significant predictors of test score change. Persons who tended to improve most on repeated testing were girls, students from the upper class, or students with relatively high GPA's. Motivational effects on repeated testing appeared to be especially detrimental to an in-depth analysis of the two research concerns--remembering specific test items, and identifying students whose scores increased.

Further studies on repeated testing should attempt to standardize the conditions under which tests are given, while making sure that students are adequately motivated to do their best during each testing session. It would appear that only when a number of different researchers attempt to work under relatively standard conditions will the effects of variables related to changes in standardized test scores be adequately evaluated.

BIBLIOGRAPHY

Dearborn, W. F., and Rothney, J. W. Predicting the Child's Development. Cambridge, Massachusetts: Sci-Art Publishers, 1941.

Droege, R. C. Effects of practice on aptitude scores. Journal of Applied Psychology, 1966, 50, 306-310.

Eichelberger, R. T. Practice effects of repeated IQ testing and the relationship between IQ change scores and selected individual characteristics, unpublished doctoral dissertation, Southern Illinois University, Carbondale, Illinois, 1970.

Heim, A. W. and Wallace, J. G. The effects of repeatedly retesting the same group with the same intelligence test: Part I: Normal Adults. Quarterly Journal of Experimental Psychology, 1949, 1, 151-159.

Kelly, F. J., Beggs, D., McNeil, K., Eichelberger, T., and Lyon, J. T. Research Design in the Behavioral Sciences: Multiple Regression Approach. Carbondale, Illinois: Southern Illinois University Press, 1969.

Kreit, L. H. The effects of test-taking practice on pupil test performance. American Educational Research Journal, 1968, 5, 616-625.

Lindquist, E. F. (ed.) Educational Measurement. Washington, D. C.: American Council on Education, 1951.

McGuire, C., and White, G. D. The measurement of social status. Research paper in human development No. 3 (revised), Department of Educational Psychology, University of Texas, 1955.

Peel, E. A. Practice effects between three consecutive tests of intelligence. British Journal of Educational Psychology, 1952, 22, 196-199.

Primary Mental Abilities (PMA) Technical Handbook. Chicago: Science Research Associates, 1968.

Slakter, M. J., and Koehler, R. A. Test-Wiseness, Final Technical report. Teacher Education Research Center, State University College at Fredonia. Fredonia, New York, 1969.

Thoulass, R. H. Test unreliability and function fluctuation, British Journal of Psychology, 1936, 26, 325.