DOCUMENT RESUME

ED 053 202 TM 000 734

AUTHOR Lewis, Ernest L.; Beggs, Donald L.
TITLE The Effects of Repeated Testing on Verbal and Nonverbal Ability Assessment.
PUB DATE Apr 71
NOTE 24p.; Paper presented at the Annual Meeting of the American Personnel and Guidance Association, Atlantic City, New Jersey, April 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Conditioning, Educational Programs, *Grade 6, *Intelligence Tests, *Memory, Nonverbal Tests, *Recall (Psychological), Response Mode, Scores, Test Reliability, Test Validity, *Test Wiseness, Verbal Tests
IDENTIFIERS *Lorge Thorndike Intelligence Test

ABSTRACT

The purpose of this study was to attempt to determine if score gains obtained upon repeated testing with an intelligence test result from a practice effect, from students remembering specific items, or from a combination of both. The verbal and nonverbal batteries of an I.Q. test were administered to 860 sixth graders on three occasions with two-month and four-month intervals between testing sessions. Some students received the same form of the test each time they were tested while others received alternate forms of the test. The results indicated that the subjects did experience an increase in verbal mean I.Q. In the nonverbal results, only groups retested with the same form of the test experienced significant mean gains. The verbal mean gains appeared to result from a practice effect while the nonverbal results appeared to result from students remembering specific items from one testing session to the next. (Author)

Paper Presented to a meeting of
The American Personnel and
Guidance Association

April 1971

The Effects of Repeated Testing on Verbal and Nonverbal Ability Assessment

Ernest L. Lewis
Southern Illinois University

and

Donald L. Beggs
Southern Illinois University

Intelligence test scores have been used by school systems in the United States to group students for instruction, to determine which students should be considered as gifted, for assigning students to instructional tracks, and for advising students to choose a vocational or college prepatory program of study. Since intelligence scores have traditionally been assigned such an important role by our educational system, one would expect them to be valid, reliable, and usable. Although validity certainly is of utmost importance with respect to intelligence tests, this study was primarily concerned with the reliability and usability of intelligence test scores.

For a test to be reliable, each individual has to obtain approximately the same standard score on two separate administrations of the test. In other words, each individual has to have approximately the same position with the respect to the means of the distributions of scores on two administrations of the test. This does not require that each individual receive the same score on both administrations of the test. As a result, the conditions of reliability would be met if each individual score increased or decreased by some constant amount from the first testing session to the second.

A number of researchers (Eichelberger, 1970; Kreit, 1968; Vernon, 1954; Watts et. al., 1952; Peel, 1951; Odell, 1925), who have investigated the effects of repeated testing, have reported that intelligence test scores increase upon repeated testing. Kreit (1968) reported that when an intelligence test was given to the same group of students on four different occasions

there was a statistically significant increase in mean scores from the first to second administration. Eichelberger (1970), Watts et. al., (1952), and Peel (1951) reported findings which were essentially the same as those reported by Kreit (1968). Although these results do not indicate that the tests are unreliable, the relative large increases from test session one to test session two do have a marked adverse effect on the usability to the test scores. Consider the difficulty in deciding the score to use as a cut-off value in comprising groups for instruction, for determining who is to be considered gifted, or who should be directed to a vocational program of study if one knows that students would receive higher scores if they were given the same intelligence test a month later. As these examples illustrate, a general increase can affect the usability of a test score even though it does not affect the test's reliability. Certainly intelligence scores have to be reliable, but they also need to be usable if they are to serve as the basis for instructional decisions.

A number of researchers (Mann, et. al., 1970; Kreit, 1968; Vernon, 1954; Heim and Wallace, 1949) have attributed the increases which occur in intelligence test scores upon repeated testing to a practice effect. This concept indicates that scores on intelligence tests rise not because the instrument is unreliable, but rather because the students develop skill in taking the test through practice with it. In other words, the student learns how to take the test as a result of taking it and as a result, scores higher on a subsequent administration.

In his doctoral dissertation, Eichelberger (1970) identified a second possible explanation for the increases which occur in intelligence scores upon repeated testing. Eichelberger( )1970) concluded that large initial increases observed in his data resulted from students remembering specific test items rather than from a practice effect.

The purpose of this study was to investigate the nature of gains which occur in intelligence test scores upon repeated testing. Specifically, this study was an attempt to determine if observed score gains result from a practice effect, from students learning specific items, or from a combination of both. Since several researchers (Vernon, 1954;, Derner, et. al., 1950; Hamister, 1949) have reported differential gains between verbal and nonverbal intelligence tests, this study examined the effects of repeated testing on both types of instruments.

METHOD

Sample

Approximately 860 students from 34 sixth-grade classrooms in three middle schools in Springfield, Illinois comprised the subjects for this study. This group represented a broad crossection of socio-economic backgrounds and scholastic abilities. All students who had attended Springfield schools as fifth graders had taken the Otis-Lennon Mental Ability Test and the Stanford Achievement Test during the 1969-70 school year. Although some variability existed in the testing backgrounds of the subjects, the majority of the students had taken at least six standardized tests prior to the 1970-71 school year.

Training Session

Since all subjects in this study had essentially the same past experience with standardized tests, it was decided that it would be necessary to develop a group of students who would be more sophisticated in test-taking principles than the remainder of the sample. In order to accomplish this task, a training program in test-wiseness was developed. This program was approximately 45 minutes in length and was presented in normal lecture fashion with the aid of prepared transparencies and an overhead projector. This training program was based, to a large extent, on the work of Slakter and Koehler (1969).

The program developed for this study was designed to serve two major purposes. First, the program was designed to teach students a general approach to taking tests. For example, students were informed that they should not spend too much time on any one item but should move ahead in order to have time to complete the entire test.

The second purpose of the training program was to teach students the types of clues to keyed responses which are often available in multiple choice items. It was pointed out in the training session that multiple choice items often can be answered even though an individual has no knowledge of the content of the item. The stem and the alternatives of the items often provide clues as to the keyed responses.

An example of one type of clue discussed in the training session follows:

1. The Flying Spider is known for its ability to:
    a. blend in with its surroundings
    b. glide through the air
    c. kill its prey
    d. make very large webs

Although there is no Flying Spider, this example taken from Slakter and Koehler (1969) demonstrates one test-taking skill. The test-wise individual would probably choose alternative "b" since it is the only alternative which refers to flying.

Several different clues were included in the program and each was presented through the use of several examples. Following the presentation of general principles and specific clues, the students were given a practice set of items consisting of ten examples of the various clues included in the presentation. After each student had completed the practice items, the students were provided with an opportunity to discuss their reasons for selecting various alternative choices and were informed as to the keyed response.

Testing Instrument

The criterion measures employed in this study were the verbal and nonverbal batteries of Level 3 of the Lorge-Thorndike Intelligence Tests. Level 3 was designed for 4-6 grade students and is available in two forms - Form A and Form B. Both forms were employed in this study. When the two batteries of the Lorge-Thorndike Intelligence Tests are given together, the testing session lasts

approximately 90 minutes. The reusable edition of Level 3 was used and the answer sheets used were those which Houghton-Mifflin Publishing Company had designed for use with the I.B.M. 1230 scoring machine.

Procedure

Since it was desired to train some students in test-wiseness and not to train others, to use varying time intervals between testing sessions, and to use alternate forms with some students and to use the same form with some students, the subjects were divided into eight groups. Random selection for inclusion in a group was made on the basis of classroom rather than by subject. Table 1 presents the design of this study in relation to group breakdowns, form of the test a group received and the dates on which each group was tested.

TABLE 1

GROUP DESIGNATIONS AND
ORDER OF TESTING

	Oct. 2, 1970	Dec. 1, 1970	Jan. 26, 1971
Trained Group	Group 1 Form A	Form A	Form A
	Group 2 Form A		Form A
	Group 3 Form B		Form A
	Group 4 Form B	Form A	
Nontrained Group	Group 5 Form A	Form A	Form A
	Group 6 Form A		Form A
	Group 7 Form B		Form A
	Group 8 Form B	Form A	

The training sessions were conducted by four doctoral students from Southern Illinois University. Each trainer had had at least one year of teaching experience in a public school. Each trainer conducted training sessions in four classrooms of about 30 students each on October 1, 1970. The classrooms were not modified in any way other than to move in an overhead projector and the 45 minute training session represented no deviation from the school's ordinary class schedule.

All groups took both the verbal and nonverbal batteries of the Lorge-Thorndike Intelligence Tests on October 2, 1970. The testing session lasted approximately 90 minutes and was conducted by the classroom teacher in the classroom to maintain as normal a set of conditions as possible. All groups took the verbal battery first and then the nonverbal battery. The December 1, 1970 and January 26, 1971 sessions followed exactly the same format as the October 2 testing session.

A student's score on the test consisted of the number of items answered correctly. All test scoring was done by machine.

Results

Before an appropriate analysis could be performed, it was necessary to determine if individuals had been randomly assigned to the various groups involved in the study. Analysis indicated marked differences in the average abilities of the individual students in the various classrooms. In addition, the results indicated that the mean IQ scores of the classrooms were significantly different. Therefore, it was necessary to consider the number of classrooms in each group as being the number of observations for that group which limited the degrees of freedom in the various comparisons made in this study.

Tables 2 and 3 present the means and standard deviations of both the verbal and nonverbal IQ tests for each of the three testing sessions. A general observation that can be made from this descriptive data is that nearly every group experienced an increase in mean IQ from the first to second administration of the test, regardlesss of the form that was employed in the testing session and completely independent of whether or not the group had received instruction in test-taking behavior.

The results reported in Tables 4 and 5 are the mean difference scores for the groups. Although some significant increases were obtained on the verbal IQ test, the greatest number of significant increases occurred on the nonverbal test. On the nonverbal test, the groups which showed a significant increase from time 1 to time 2 were given the same form of the test on both administrations. Groups 3,4,7, and 8, which received different forms, did not show a statistically significant change. However, Group 8 did show a statistically significant change from testing session 1 to testing session 2 when separate forms of the verbal test were used. In general, almost all of the comparisons from testing time 1 to testing time 3 indicate an increase in the average score for the various groups represented.

Tables 6 and 7 indicate that the correlations between testing periods 1 and 2, 2 and 3, and 1 and 3 were all quite high. The only exceptions can be found in Group 3 and Group 6 where, because of the small number of classrooms involved in each group, a slight change in the positions of only two classrooms could result in a significant reduction in the correlation coefficient. In general, however, the correlation coefficients were quite high and there seems to be little difference as to whether the same form of the test was used or

parallel forms of the test were used. It is interesting to note that although these correlations are quite high, there were significant mean increases in many cases.

The results in Tables 8 and 9 indicate that the Trained Group and Nontrained Group showed no significant differences in change scores over the various time periods used in this study. In general, the results of both the verbal and nonverbal analyses indicate that the training that occurred in the first four groups had very little effect, if any, on the groups in comparison with the nontrained groups. The only significant differences obtained between a trained and nontrained group occurred in the nonverbal gain scores of Groups 1 and 5 between testing sessions 1 and 3 and between testing sessions 2 and 3. In both cases, Group 5, the group which had <u>not</u> received training in test-wiseness, experienced the largest gain in mean nonverbal IQ scores.

Tables 10 and 11 present the results of comparisons between groups which received parallel forms of the test and groups which received the same form. It is obvious from Table 10 that no significant differences were found between groups which received the same form of the verbal test and groups which received different forms of the verbal test. Alternating forms seemed to have no effect on the changes in scores that occurred over the testing sessions. The nonverbal test results reported in Table 11 are similar to the verbal results with the exception of the comparisons that were made in the nontrained groups. In the nontrained groups, the groups which had the same form over the various testing periods showed a greater mean increase than those groups which received parallel forms.

Conclusions

As was stated earlier, the primary purpose of this study was to investigate the nature of gains which occur in intellegence test scores over repeated testing. Specifically, this particular study was concerned with training children in principles of test-wiseness and determining if that training has any effect on the gain scores from one testing period to another. Since the nontrained groups experienced mean IQ score gains as large as those of the trained group, the conclusion can be drawn that test-wiseness training sessions employed in this study seemed to have no effect on the change scores of the students.

A second conclusion that can be drawn from the results of this study is that repeated testing with the same instrument does result in an increase in mean IQ scores. From the standpoint of usability of test scores, the results of this study seem to indicate that the scores obtained from the second testing session were more stable and representative of the child's ability than those scores obtained from the first testing session. This is based on the fact that the changes from test periods 2 and 3 were not nearly as marked as the changes from 1 to 2, and 1 to 3. Therefore, a third conclusion that can be drawn from this study is that if test scores are to be used in the decision-making process, one should consider the possibility of assessing students more than one time with respect to the same trait in order to obtain the most appropriate test score for the individual child.

Another conclusion which can be drawn from the results of this study is that the gains which occur in mean verbal IQ scores result from a practice effect rather than from students remembering specific items in the test. With respect to nonverbal IQ tests, the groups in the study which received parallel

forms of the test showed almost no increase or a slight decrease in mean IQ scores while groups which took the same form of the test a second time showed a relatively large increase in mean IQ. Consequently, it seems that the increases which occurred in the nonverbal mean IQ scores might have resulted from students learning specific items on the test.

In general, the results indicate that students do increase their test scores upon repeated testing with a test that is designed to measure a stable trait. This would seem to imply that those individuals who are using tests for the purpose of decision-making with respect to students need to obtain more than a single measure of the same trait on any child in order to obtain a more stable and usable measure for the individual student. The results of this study would seem to indicate that there is a need to use parallel forms of a test in the assessment of nonverbal intellectual ability.

REFERENCES

Derner, G. F., Abden, M., & Canter, A. H. The Reliability of the Wechsler-Bellevue Intelligence Test. Journal of Consulting Psychology, 1950, 14, 172-179.

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.

Hamister, R. G. Test-retest reliability of the Wechsler Bellevue. Journal of Consulting Psychology, 1949, 13, 39-43.

Heim, A. W., & Wallace, J. G. The effects of repeatedly retesting the same group in the same intell gence test: Part I: Normal Adults. Quarterly Journal of Psychology, 1949, 1, 151-159.

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

Mann, L., Taylor, T. G., Proger, B. B. Dungan, R. H., & Tidey, W. S. The effect of serial retesting on the relative performance of high- and low-test anxious seventh grade students. Journal of Educational Measurement. 1970, 7, 97-104.

Odell, C. W. Some data as to the effect of previous training upon intelligence test scores. Journal of Educational Psychology, 1925, 16, 482-486.

Peel, E. A. Practice effects between three consecutive tests of intelligence. British Journal of Educational Psychology, 1951, 22, 196-199.

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

Vernon, P. E. Practice and coaching effects in intelligence tests. The Educational Forum, 1954, 269-280.

Watts, A. F., Pidgeon, D. A. & Yates, A. Secondary school entrance examinations. Newnes, London, 1952.

TABLE 2

VERBAL IQ MEANS AND STANDARD

DEVIATIONS FOR THREE TESTING SESSIONS

	Group	Testing Session 1		Testing Session 2		Testing Session 3	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Trained group	1	111.33	6.68	114.74	6.96	114.71	5.80
	2	108.77	11.97			112.34	10.64
	3	106.04	1.03			107.38	1.38
	4	93.70	9.87	95.99	11.58		
Nontrained group	5	103.72	7.12	105.17	5.71	107.37	6.01
	6	108.55	4.65			110.18	2.47
	7	107.61	6.58			108.91	6.34
	8	98.88	7.43	103.53	9.64		

TABLE 3

NONVERBAL IQ MEANS AND STANDARD

DEVIATIONS FOR THREE TESTING SESSIONS

	Group	Testing Session 1		Testing Session 2		Testing Session 3	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Trained group	1	107.94	5.49	112.42	6.23	112.72	6.88
	2	105.87	8.98			111.44	8.02
	3	103.21	5.76			108.20	2.52
	4	94.83	12.25	97.11	9.55		
Nontrained group	5	101.18	5.40	106.94	5.53	109 94	5.13
	6	107.24	3.46			111.64	3.98
	7	107.80	6.86			107.74	5.93
	8	102.74	8.00	102.16	9.49		

TABLE 4

MEAN DIFFERENCE SCORES ON VERBAL IQ

TEST FOR THREE TESTING SESSIONS

Group	TS 2 - TS 1	TS 3 - TS 2	TS 3 - TS 1	df
1	+3.41*	-0.03	+3.38	3
2			+3.57	2
3			+1.34	4
4	+2.29			2
5	+1.45	+2.20*	+3.65*	4
6			+1.63	4
7			+1.30	4
8	+4.65*			3

* Significant at the .05 level using a one-tailed correlated t-test.

Note. - Groups 1,2,5, and 6 received the same form each time they were tested while Groups 3,4,7, and 8 received alternate forms.

TABLE 5

MEAN DIFFERENCE SCORES ON NONVERBAL

IQ TEST FOR THREE TESTING SESSIONS

Group	TS 2 - TS 1	TS 3 - TS 2	TS 3 - TS 1	df
1	+4.48*	+0.30	+4.78*	3
2			+5.57*	2
3			+4.99	4
4	+2.28			2
5	+5.76*	+3.00*	+8.76*	4
6			+4.40*	4
7			-0.06	4
8	-0.58			3

*Significant at the .05 level using a one-tailed correlated t-test.

Note.- Groups 1,2,5, and 6 received the same form each time they tested while Groups 3,4,7, and 8 received alternate forms.

TABLE 6

CORRELATIONS BETWEEN CLASSROOM VERBAL MEAN IQ

SCORES FOR THREE TESTING SESSIONS

Group	TS 1 with TS 2	TS 2 with TS 3	TS 1 with TS 3
1	+.970	+.916	+.929
2			+.989
3			+.107
4	+.999		
5	+.911	+.989	+.902
6			+.246
7			+.999
8	+.992		

TABLE 7

CORRELATIONS BETWEEN CLASSROOM NONVERBAL

IQ SCORES FOR THREE TESTING SESSIONS

Group	TS 1 with TS 2	TS 2 with TS 3	TS 1 with TS 3
1	+.979	+.996	+.965
2			+.993
3			-.309
4	+.999		
5	+.951	+.980	+.993
6			+.881
7			+.808
8	+.951		

TABLE 8

COMPARISONS BETWEEN VERBAL MEAN DIFFERENCE SCORES OF TRAINED GROUPS AND CORRESPONDING NONTRAINED GROUPS

Group	TS 2 - TS 1	TS 3 - TS 2	TS 3 - TS 1	F	df Numerator	df Denominator
1*	+3.41			1.044	1	7
5	+1.45					
1*		-0.03		2.1140	1	7
5		+2.20				
1*			+3.38	0.0157	1	7
5			+3.65			
2*			+3.57	0.3418	1	6
6			+1.63			
3*			+1.34	0.0015	1	8
7			+1.30			
4*	+2.29			1.4156	1	5
8	+4.65					

* Indicates group which received training.

TABLE 9

COMPARISONS BETWEEN NONVERBAL MEAN

DIFFERENCE SCORES OF TRAINED GROUPS

AND CORRESPONDING NONTRAINED GROUPS

Group	TS 2 - TS 1	TS 3 - TS 2	TS 3 - TS 1	F	df Numerator	df Denominator
1*	+4.48			1.1438	1	7
5	+5.76					
1*		+0.30		11.9249@	1	7
5		+3.00				
1*			+4.78	6.6110@	1	7
5			+8.76			
2*			+5.57	0.6385	1	6
6			+4.40			
3*			+4.99	1.5687	1	8
7			-0.06			
4*	+2.28			1.1438	1	5
8	-0.58					

* Indicates group which received training.

@ Significant at .05 level using one-tailed F-test

TABLE 10

COMPARISONS BETWEEN VERBAL MEAN DIFFERENCE SCORES OF GROUPS RECEIVING SAME FORM OF IQ TEST AND GROUPS RECEIVING ALTERNATE FORMS DURING TWO TESTING SESSIONS.

Group	TS 2 - TS 1	TS 3 - TS 1	F	df Numerator	df Denominator
1*	+3.41		0.5209	1	5
4	+2.29				
2*		+3.57	2.0547	1	6
3		+1.34			
5*	+1.45		2.2607	1	7
8	+4.65				
6*		+1.63	0.0187	1	8
7		+1.30			

* Indicates group receiving same form during both sessions.

TABLE 11

COMPARISONS BETWEEN NONVERBAL MEAN DIFFERENCE SCORES

OF GROUPS RECEIVING SAME FORM OF IQ TEST AND GROUPS RECEIVING

ALTERNATE FORMS DURING TWO TESTING SESSIONS.

Group	TS 2 - TS 1	TS 3 - TS 1	F	df Numerator	df Denominator
1*	+4.48		1.3638	1	5
4	+2.28				
2*		+5.57	0.0149	1	6
3		+4.99			
5*	+5.76		11.7156@	1	7
8	-0.58				
6*		+4.40	3.9733	1	8
7		-0.06			

* Indicates group receiving same form during both sessions

@ Significant at .05 level using two-tailed F-test.