DOCUMENT RESUME

ED 076 675 TM 002 683

AUTHOR Hecht, James T.
TITLE Usability of Scores Obtained from Repeated IQ Test Administrations.
PUB DATE 73
NOTE 18p.; Paper presented at annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Intelligence Differences; *Intelligence Quotient; Intelligence Tests; Scores; Standardized Tests; Technical Reports; *Testing; *Test Validity; *Test Wiseness; *Time Factors (Learning)

ABSTRACT

The relationship of test wiseness to I.Q. and the usability of I.Q. scores are discussed. Test wiseness involves the examinee's ability to obtain a high score on a standardized achievement test as a result of utilizing test-taking experience. Usability of I.Q. scores refers to the value of I.Q. scores to educators in making educational decisions. A primary reason for conducting the present investigation was to study the effects of repeated testing over an eighteen month interval. When I.Q. tests are administered over a short term, temporary sources of variance may be, at least in part, responsible for the increase in I.Q. Remembering specific items and practice effect provide plausible explanations for the short term gains. Gains found over periods of four months or less were not present over the longer time interval of eighteen months. (Author)

USABILITY OF SCORES OBTAINED FROM REPEATED IQ TEST ADMINISTRATIONS

By

James T. Hecht, Ph.D.
Coordinator of Certification
American Nurses' Association

presented at

Americal Educational Research Association
New Orleans, Louisiana
February 25 - March 2, 1973

The "intelligence" test, commonly abbreviated to "IQ" has acquired an important position in the decision-making processes employed by our nation's school systems. The IQ test is used widely for predictive purposes. An example would be admission officers using IQ test scores as criteria for evaluating students seeking admission to college. In the lower grades, IQ test scores have served as a basis for assigning pupils to special classes and advanced or experimental curricula.

Since so much importance is attached to IQ test scores, such scores should be reliable and usable. A general definition of test-retest reliability indicates a test is reliable if subjects tested with a particular test are able to maintain approximately the same position with respect to the mean on different test administrations. A student's score may fluctuate without seriously affecting reliability as long as the other students' scores fluctuate in a like manner. However, fluctuations in scores from one testing session to another may have adverse effects upon usability. Decisions based on test scores that fall below a particular cutoff point on one occasion, while falling above the cutoff point on another occasion, are not sound. As a result, reliable test scores are not necessarily usable.

A common type of fluctuation found with respect to IQ test scores occurs on repeated measures. Most school systems administer several IQ tests to students during their primary and secondary school years. An important decision that educators must make concerns which test to consider as being representative of students' abilities. Because test scores may fluctuate without necessary being unreliable or invalid, some criteria of usability must be established. In the present investigation, the criteria of GPA predictability

and stability of IQ scores have been selected to serve as a basis of evaluating usability of IQ scores. In this case, GPA predictability refers to the amount of common variance obtained between the variables of IQ and GPA. Stability refers to the extent to which mean IQ scores fluctuate. Small fluctuations in mean IQ scores from one testing session to another would be more desirable than large fluctuations in mean IQ scores. Research on the effects of repeated measures may offer insight into problems concerning usability.

Related Research

Research on the results of repeated IQ testing indicates that scores increase from one testing session to another. While IQ test scores increase significantly from the first to second (time intervals range from a few minutes to five months) test administration (Lewis, 1971; Eichelberger, 1970; Kreit, 1968; Vernon, 1954; Thorndike, 1922), they do not increase significantly between the second and third (time intervals range from a few minutes to five months) IQ test administrations (Lewis, 1971; Eichelberger, 1970; Mann, et. al., 1970; Kreit, 1968; Peel, 1972; Watts, et. al., 1952).

Three alternative explanations have been advanced to account for the gain found on repeated IQ testing. The first explanation deals with a specific variance theory. If the same form of a test is administered repeatedly, it is possible that examinees may remember items from the first testing session (Eichelberger, 1970). The second explanation concerns a

practice effect theory. Practice effect is an effect that results from students actually having practice taking particular tests (Thorndike, 1922). The third explanation concerns test-wiseness. Test-wiseness involves test-taking experience and strategy not related to test content (Vernon, 1962).

Three variables manipulated in the past on repeated measures include time intervals between testing sessions, test forms administered, and item types. Research indicates that, regardless of the size of the time interval employed, significant differences between first and second testing sessions result (Vernon & Parry, 1949; Green, 1928). Forms of the test administered have been manipulated in an attempt to isolate specific and general sources of variation. When alternate forms are employed, variance specific to a particular test is eliminated (Snedden, 1931). Item type refers to verbal and nonverbal test items. Test results indicate that mean IQ gains on repeated testings are larger on the nonverbal section than the verbal portion of the tests (Derner, et. al., 1950).

Method

The purpose of this study was to investigate usability of IQ scores. Specifically, the effect of a nineteen-month interval between first and last testing sessions was examined. In addition, the results were analyzed for both verbal and nonverbal IQ tests.

Subjects for this investigation included approximately 1200 seventh-grade students from three middle schools in the Springfield, Illinois area. The subjects were tested by Lewis (1971) three times with Level 3 of the Lorge-Thorndike Intelligence Tests. Also, Lewis (1971) attempted to train one-half the subjects to be test-wise. Group designations and order of testing appears in Table I.

TABLE I

GROUP DESIGNATIONS AND ORDER OF TESTING WITH LORGE-THORNDIKE INTELLIGENCE TEST

		Level 3 Oct. 2, 1970	Level 3 Dec. 1, 1970	Level 3 Jan. 26, 1971	Level 4 May 2, 1972
Trained Group	Group 1	Form A	Form A	Form A	Form A
	Group 2	Form A		Form A	Form A
	Group 3	Form B		Form A	Form A
	Group 4	Form B	Form A		Form A
Nontrained Group	Group 5	Form A	Form A	Form A	Form A
	Group 6	Form A		Form A	Form A
	Group 7	Form B		Form A	Form A
	Group 8	Form B	Form A		Form A

Criterion instruments employed in this investigation include the verbal and nonverbal batteries of the Level 4 Lorge-Thorndike Intelligence Tests. Level 4 of the Lorge-Thorndike Intelligence Tests was designed for seventh to ninth-grade students. Only Form A was employed in the present investigation.

Three steps were involved in the process of gathering data for the present investigation. Scores from three previous IQ testings conducted by Lewis (1971) were collected and matched. In May, 1972, the verbal and nonverbal batteries of the Lorge-Thorndike Intelligence Tests were administered. And in June, 1972, when the 1971 - 72 school year had ended, seventh-grade GPA's were calculated.

Results

Repeated IQ testings of middle school subjects yielded increases in mean IQ that are generally consistent with literature on repeated testing. Mean verbal IQ increased consistently until the fourth testing session, which occurred nineteen months after the original testing session. However, mean nonverbal IQ increased consistently through the fourth testing session. Results presented in Table II summarize findings from the present investigation concerning mean IQ score gains found on repeated measures.

TABLE II

IQ MEANS AND STANDARD DEVIATIONS FOR SUBJECTS RECEIVING ALL FOUR TESTS

N = 191	Testing Session I		Testing Session II		Testing Session III		Testing Session IV	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Verbal	108.84	14.53	110.86	13.99	111.27	13.90	108.50	12.98
Non-Verbal	103.82	14.88	110.77	13.01	111.94	15.17	115.53	16.85

The relationship of gain scores over different time intervals was investigated. In the case of verbal scores, the gain between first and second test administrations was larger than the gain between first and fourth test administrations. For nonverbal scores, the gain between first and fourth test administrations was larger than the gain between first and second test administrations.

The relationship between mean verbal IQ gains and mean nonverbal IQ gains was examined. Results indicate that obtained mean nonverbal IQ gains over a nineteen-month interval appear to be greater than obtained mean verbal IQ gains. Thus, in the long run, subjects gain more on non-verbal repeated testing than verbal repeated testing.

Lewis (1971) did not find a statistically significant difference between subjects trained to be test-wise and nontrained subjects on mean verbal or nonverbal IQ gains over the short run (four months). Results from the present investigation confirm the findings of Lewis (1971). An analysis of the results did not reveal a statistically significant difference in mean verbal or nonverbal IQ gains over the long run (nineteen months) between subjects trained in test-wiseness and nontrained subjects.

Obtained correlations between verbal IQ and GPA were slightly higher than obtained correlations between nonverbal IQ and GPA. However, since no particular correlation of IQ testing session with GPA was consistently higher than other correlations of IQ testing sessions with GPA, a best predictor of GPA could not be determined. The obtained correlations between IQ and GPA are presented in Table III.

TABLE III

CORRELATIONS BETWEEN IQ TESTS AND GPA AND ~~z~~'S BETWEEN CORRELATIONS

	Correlation Coefficients				~~z~~
Verbal	$r_{I,GPA}$	.60	$r_{II,GPA}$	.63	.91
	$r_{I,GPA}$	.60	$r_{III,GPA}$	.64	1.41
	$r_{I,GPA}$	.60	$r_{IV,GPA}$	.67	1.71
	$r_{II,GPA}$	.63	$r_{III,GPA}$	.64	.16
	$r_{II,GPA}$	.63	$r_{IV,GPA}$	.67	1.58
	$r_{III,GPA}$	.64	$r_{IV,GPA}$	.67	.80
Non Verbal	$r_{I,GPA}$	.56	$r_{II,GPA}$	.56	.00
	$r_{I,GPA}$	.56	$r_{III,GPA}$	.62	1.55
	$r_{I,GPA}$	.56	$r_{IV,GPA}$	.63	1.54
	$r_{II,GPA}$	.56	$r_{III,GPA}$	.62	2.73*
	$r_{II,GPA}$	.56	$r_{IV,GPA}$	.63	2.00*
	$r_{III,GPA}$	.62	$r_{IV,GPA}$	.63	.27

* $p < .05$ using a dependent test

Conclusion

While many researchers (Mann, Taylor, Proger, Dungan & Tidey, 1970; Kreit, 1968; Vernon, 1954; Yates & James, 1953; Watts, Pidgeon, & Yates, 1952; Peel, 1951; Odell, 1925) have indicated that gains in mean IQ scores result from repeated testing; others have sought to provide explanations for those gains (Lewis, 1971) Eichelberger, 1970; Slakter & Koehler, 1969). A common explanation employed by most researchers in the area of repeated testing concerns attributing the gain to a temporary general source of variation such as practice effect. Eichelberger (1970) attributed gains resulting from repeated measures to a temporary specific source of variation which he interpreted as remembering specific test items. Lewis (1971) attributed gains on repeated measures to remembering specific test items and practice effect. Explanations provided by Lewis (1971) and Eichelberger (1970) appear reasonable in light of the fact that their research was confined to relatively short intervals between repeated testings of less than four months. However, explanations of gains occurring on repeated measures over time intervals of less than four months may not be adequate for long-term gain (more than one year) situations.

In the present investigation, subjects employed by Lewis (1971) were retested nineteen months after the first testing session administered by Lewis (1971). While Lewis (1971) reported continual gains in mean verbal and nonverbal IQ scores over the four months employed in his research, the present investigator, employing a nineteen-month interval, found gains only in mean nonverbal IQ scores. In the present investigation, mean verbal IQ gains obtained by subjects over the short-run (four months) were lost in the long-run (nineteen months). Nonverbal IQ test results indicated a continual gain in mean IQ scores over the short run and the long run.

Since researchers (Lewis, 1971; Eichelberger, 1970) have indicated that temporary specific and temporary general (less than six months) (Cronbach, 1960) sources of variance may be responsible for short-term gains in mean IQ (less than six months), it would seem reasonable to expect lasting (more than one year (Cronbach, 1960) sources of variance to be responsible for long-term (nineteen months) gains in mean IQ. Since researchers (Slakter & Koehler, 1969; Vernon, 1954) have indicated that test-wiseness may represent a lasting source of variation, the present investigator considered test-wiseness as a possible factor responsible for long-term (nineteen months) gains in mean IQ.

Since Lewis (1971) attempted to train approximately one-half of the subjects to be test-wise, an analysis of the long-term (nineteen months) training effects was made. Subjects trained in test-wiseness did not gain more or less statistically over the long run (nineteen months) than nontrained subjects. Although the present investigator expected test-wiseness to be at least in part responsible for long-term (nineteen months) gains found on repeated testing, the expectation was not supported.

Explanations by Lewis (1971) and Eichelberger (1970) of short-term (less than six months) IQ gains on repeated measures concerning practice effect and remembering specific items appear reasonable. However, it does not seem reasonable to expect temporary sources of variance such as remembering specific items and practice effect to be as important in long-term (more than one year) gain situations as lasting sources of variation. Although

test-wiseness was expected to account for a statistically significant amount of the long-term (more than one year) gains resulting from repeated testing, support for that expectation was not obtained in the present investigation. In order to develop an adequate explanation of long-term (more than one year) IQ gains on repeated measures, further research is necessary.

Educators frequently employ IQ scores as a basis for assigning pupils to classes. If more than one set of IQ scores is available, educators must decide which set to employ. Given the changes which occur in IQ scores upon repeated measures, this can be a difficult decision for educators. Lewis (1971) attempted to provide an answer to the foregoing problem by analyzing data from three administrations of Level 3 of the Lorge-Thorndike Intelligence Tests given to approximately 1,000 middle school children. If three IQ scores are available, which one should be used to make educational decisions? Since mean IQ scores increased significantly from administration one to two, but did not increase significantly from administration two to three, Lewis (1971) suggested that the scores from the second testing session may be more usable. Lewis' (1971) suggestion that IQ scores from the second testing administration may be the most usable was based on the criterion of stability. Since the first set of scores changed significantly on the second administration, they were considered to be more stable than the first set of scores. The results from Lewis (1971) imply that two IQ tests should be administered in order to obtain a set of scores that are usable.

Although research by Lewis (1971) provided important information for the area of repeated testing, the greatest contribution was to stimulate further research. A problem that was not resolved by Lewis (1971) concerned the effect of repeated testing on IQ scores when the testing sessions were separated by more than a year. Since Lewis (1971) suggested that a follow-up study be conducted to determine the long-range effects of repeated testing, the present investigator retested Lewis' (1971) subjects nineteen months after the original test administration. This provided the present investigator with four sets of IQ scores. Stability, as employed by Lewis (1971), provided the first criterion, while correlations between IQ tests and GPA provided the second criterion.

Mean verbal IQ increased from testing session one to testing session two, and from testing session two to testing session three. Mean verbal IQ declined on the fourth testing session, conducted nineteen months after the original test administration, to the level established on testing session one. Gains found on verbal IQ test administrations two and three may have been caused by temporary sources of variation such as practice effect and remembering specific items as indicated by the research of Lewis (1971) and Eichelberger (1970). Since the mean verbal IQ from testing session one was not statistically significantly different ($p.05$) from the mean verbal IQ from testing session four, the first verbal IQ testing session scores appear to be more stable than the others.

Lewis (1971) found the second verbal IQ testing (conducted two months after testing session one) administration scores to be more stable than the first. In the present investigation, the first verbal IQ testing administration scores were found to be more stable than the second or third testing administration scores. After a time interval of seventeen months elapsed between Lewis' (1971) second testing session, a testing session administered by the present investigator indicated that the mean verbal IQ scores had dropped back to the level established on the first testing session.

Although correlations between verbal IQ scores and GPA were relatively high (ranging from .60 to.67), none of the four sets of verbal IQ test scores correlated statistically significantly higher (p.05) with GPA than any of the others. On the basis of a relationship with GPA, all of the four sets of verbal IQ testing session scores appear to be usable. However, none of the four sets of verbal IQ scores were found to be more usable than any of the others on the basis of a relationship with GPA.

Mean nonverbal IQ tests scores increased from testing session one to two, two to three, and three to four. On the basis of stability, the first three nonverbal IQ testing session scores should not be considered usable. It is difficult to determine whether the fourth nonverbal IQ testing session scores are stable, because a fifth set of nonverbal IQ scores is not available for examination. Whether or not any of the four sets of nonverbal IQ test scores are usable cannot be determined on the basis of stability.

Nonverbal IQ test scores were examined for usability on the basis of a relationship with GPA. Because the correlation between GPA and nonverbal IQ scores from testing session one was not found to be statistically significantly different (p.05) from the correlation between GPA and nonverbal

IQ testing session three scores and the correlation of GPA with nonverbal IQ testing session two scores the three correlations were considered to be the same statistically. Since correlations of GPA with nonverbal IQ testing sessions three and four were statistically significantly higher ($p<.05$) than the correlations of GPA with nonverbal IQ testing session two, nonverbal IQ scores from testing sessions three and four appear to be more usable than nonverbal IQ scores from testing sessions one and two on the basis of a relationship with GPA.

In terms of a statistical analysis, verbal IQ scores from administration one appear to be more usable in terms of stability than verbal IQ scores from the other three testing sessions, while nonverbal IQ scores from administrations three and four appear to be more usable in terms of predictability than nonverbal IQ scores from testing sessions one and two. Although scores from verbal IQ testing session one were found to be more usable than scores from the other three verbal IQ testing sessions on the basis of stability, a more usable set of verbal IQ test scores could not be determined on the basis of a relationship with GPA. Although nonverbal IQ scores from testing sessions three and four appear to be more usable than nonverbal IQ scores from testing sessions one and two on the basis of a relationship with GPA, a more usable set of nonverbal IQ scores could not be determined on the basis of stability.

In summary, several conclusions were reached as a result of the present investigation. Conclusions reached by Lewis (1971) and Eichelberger (1970) indicating that remembering specific items and practice effect may be responsible for short-term (less than five months) gains resulting from repeated

testing appear reasonable. Support for the theory that test-wiseness is an important source of variability in long-term (more than a year) gain scores resulting from repeated testing was not found. Further research is indicated to provide an adequate explanation of long-term (more than one year) gains resulting from repeated testing. Verbal and nonverbal IQ scores were examined for usability on the basis of stability and a relationship with GPA. Verbal IQ scores from administration one were found to be more usable than verbal IQ scores from the other three testing administrations on the basis of stability. Although nonverbal IQ scores from administrations three and four appear to be more usable than nonverbal IQ scores from administrations one and two on the basis of a relationship with GPA from a statistical standpoint, nonverbal IQ scores from administration one would be more usable on the basis of practicality. By using nonverbal IQ scores from administrations three or four, instead of nonverbal IQ scores from administratione one or two, approximately four percent of additional variance is accounted for. From a practical standpoint, a gain of four percent of the variance accounted for may not provide justification for administering two or three more nonverbal IQ tests. However, if a school system has a policy of administering three or four nonverbal IQ tests, it would seem reasonable to use test results from administrations three or four for basing educational decisions.

JH:ma

REFERENCES

Cronbach, Lee J. Essentials of Psychological Testing. New York: Harper and Brothers, 1960.

Derner, G.F., Aborn, M., and Canter, A.H. The reliability of the Wechsler-Bellevue subtests and scales. 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.

Greene, K.B. The influence of specialized training on test of general intelligence. 27th Yearbook: National Society for the Study of Education, 1928, 421-428.

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

Lewis, Ernest L. The effects of practice and specific item learning on verbal and nonverbal ability assessment. Unpublished doctoral dissertation, Southern Illinois University, Carbondale, Illinois, 1971.

Mann, L. Taylor, R.G., Proger, G.B., Dungan, R.H., and Tidey, N.J. 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. A note on practice effects in intelligence tests. British Journal of Educational Psychology. 1951, 21, 122-125.

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.

Snedden, D. Practice effects. Journal of Educational Research, 1931, 24, 376-380.

Thorndike, E.L. Practice effects in intelligence tests. Journal of Experimental Psychology, 1922, 5, 101-107.

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

Vernon, P.E. The determinants of reading comprehension. Educational and Psychological Measurement, 1962, 22, 269-286.

Vernon, P.E., and Parry, J.B. Personnel Selection in the British Forces. London: University of London Press, 1949.

Watts, A.F., Pidgeon, D.A., and Yates, A. Secondary School Entrance Examinations. London: Newnes, 1952.

Yates, D.A., and James, W.S. Symposium on the effects of coaching and practice on intelligence tests. British Journal of Educational Psychology, 1953, 23, 147-162.