1973

Test-Retest Reliability Of The Peabody Picture Vocabulary Test With Fourth-Grade Public School Children In Albuquerque, New Mexico

Alan Lowell LaFon

This thesis, directed and approved by the candidate's committee, has been accepted by the Graduate Committee of The University of New Mexico in partial fulfillment of the requirements for the degree of

MASTER OF ART

TEST-RETEST RELIABILITY OF THE PEABODY PICTURE VOCABULARY TEST WITH FOURTH-GRADE PUBLIC SCHOOL CHILDREN IN ALBUQUERQUE, NEW MEXICO

Title

ALAN LOWELL LAFON

Candidate

SPECIAL EDUCATION

Department

John W. Zeppa

Dean

May 2, 1973

Date

Committee

Billy L. Watson

Chairman

Glen Van Etten

Gary Hanson

TEST-RETEST RELIABILITY OF THE PEABODY PICTURE VOCABULARY TEST WITH FOURTH-GRADE PUBLIC SCHOOL CHILDREN IN ALBUQUERQUE NEW MEXICO

BY

ALAN LOWELL LAFON

B.U.S., University of New Mexico, 1971

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF ART IN SPECIAL EDUCATION
in the Graduate School of
The University of New Mexico
Albuquerque, New Mexico

MAY, 1973

ACKNOWLEDGEMENTS

I would like to thank Mary Patino for her enthusiasm and encouragement at a critical time. I wish to thank Dr. Gary Adamson for his confidence, high expectations and guidance from the start. I would also like to thank Dr. Adamson for the many intangible things that made this last year the most educational and enjoyable one I've ever spent. I wish to thank Dr. Billy Watson for giving of his time and knowledge so generously, and for putting up with me for a whole year.

TEST-RETEST RELIABILITY OF THE PEABODY PICTURE VOCABULARY TEST WITH FOURTH-GRADE PUBLIC SCHOOL CHILDREN IN ALBUQUERQUE NEW MEXICO

BY

ALAN LOWELL LAFON

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF ART IN SPECIAL EDUCATION
in the Graduate School of
The University of New Mexico
Albuquerque, New Mexico

MAY, 1973

ABSTRACT

The purpose of this study is to determine the test-retest reliability and standard error of measurement of the Peabody Picture Vocabulary Test (PPVT) based on a sample of fourth-grade public school children.

The PPVT is a test designed to provide a standardized estimate of a person's verbal intelligence through measuring his receptive vocabulary. The PPVT is a frequently used test in the appraisal and evaluation of children in the state of New Mexico and the Albuquerque Public School System. The PPVT was standardized on a sample of white children from Nashville, Tennessee. No cultural, socio-economic or geographic precautions were used in standardizing the test, and no test-retest reliability data on normal children was published with the standardization data. The assumption is that children from a different culture, socio-economic level or geographic area will perform differently on many of the standardized tests that are frequently the basis for their educational placement. Thus, the reliability of standardized tests such as the PPVT for different populations of children is of growing concern. Equivalent form reliability data published with the PPVT suggest the test is not highly reliable with nine year old children.

The sample consisted of 40 fourth-grade public school children, between the ages of 9-0 and 9-11 years, who had never been referred for Special Education. Two examiners administered Form A of the PPVT to the subjects, and in not less than seven nor more than ten days, a second administration of Form A by the same examiner was given.

This study found coefficients of test-retest reliability of r=0.795 for raw scores, r=0.769 for IQ scores, r=0.751 for Percentile scores, and r=0.807 for MA scores. Also calculated, using the test-retest reliability coefficients and the actual standard deviations of the sample, was a standard error of measurement for raw scores of ±3.89, for IQ scores ±6.52, for Percentile scores ±12.21, and for MA scores ±9.79.

This study indicates that the PPVT does not have a high enough test-retest reliability to warrant its use as the major placement instrument with nine year old children in the Albuquerque area. This study also found that there is no significant difference between the reliability coefficients published with the PPVT and the coefficients reported for the Albuquerque sample. At their best the results of the PPVT should be cautiously interpreted, especially when the test is being used as a placement instrument.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	viii
Chapter	
1. INTRODUCTION	1
2. PROCEDURE	6
Sample	6
Instrument	7
Methods	7
Data Analysis	7
3. RESULTS	8
4. DISCUSSION	12
APPENDIX	18
Review of the Literature	18
REFERENCES	22

LIST OF TABLES

Table | Page

1. Ethnic Representation of Sample Elementary School 6

2. Means and Standard Deviations of Raw Scores, IQ Scores, Percentile Scores, and MA Scores for Two Administrations of the PPVT (Form A) 9

3. Coefficients of Reliability and SE_m for Raw Scores, IQ Scores, Percentile Scores, and MA Scores for Two Administrations of the PPVT (Form A) 10

Chapter 1

INTRODUCTION

The purpose of this study is to determine the test-retest reliability and standard error of measurement of the Peabody Picture Vocabulary Test (PPVT) based on a sample of fourth-grade public school children from an Albuquerque elementary school located in the northeast heights.

The PPVT is a test designed to provide a standardized estimate of a person's verbal intelligence through measuring his receptive vocabulary. The PPVT is widely used for diagnosis and placement in the state of New Mexico and specifically in the Albuquerque Public School System. The State Standards for Special Education, Revised 1972, list the PPVT as a required test for the appraisal and evaluation of physically handicapped individuals. The PPVT is a recommended test for the appraisal and evaluation of emotionally handicapped, learning disabled, trainable mentally handicapped and speech impaired individuals.

Controversy has arisen in the last few years (Williams, 1970; Milgram, 1970; Wikoff, 1970; Newland, 1970) concerning the use of various testing instruments

with populations of children different from those on which a test was standardized. The assumption is that children from a different culture, socio-economic level or geographic area will perform differently on many of the standardized tests that are frequently the basis for their educational placement. The reliability of standardized tests for different populations of children is an important aspect of educational research.

Lyman (1965, p. 820) emphasized the significance of meaningful reliability when using standardized instruments and concluded that the PPVT is a "highly usable test of moderate reliability." Lyman warned that considerable caution needs to be exercised when interpreting the norms, especially in communities other than Nashville, Tennessee. No test-retest reliability data on regular classroom children are cited in the manual (Dunn, 1965). When using any testing instrument, it is important to have at least short-term test-retest reliability data. When the test is administered once for placement purposes, high reliability is vital because the examiner is concerned with the level at which the subject is presently functioning and how accurate the test score is. Equivalent form reliability is necessary for situations utilizing pre and post tests, but in most instances the PPVT is given only once for the child's initial appraisal and placement, and it is critical for the examiner to know how much confidence can be placed in the test scores.

The standardization manual says that certain precautions were taken so that the norms would be useful throughout the United States. These precautions centered around subject selection which was divided into three areas: pre-school, early elementary, and upper elementary and high school. At the pre-school level four elementary schools in Nashville, Tennessee were selected which provided a composite normal probability curve of IQ's based on the Kuhlmann-Finch Intelligence Test. The pre-school subjects were taken from the school zones served by these four schools. The early elementary level sample was obtained by random selection from the same four schools. The upper elementary and high school sample was taken from two junior high and two high schools where pupils approximated a composite normal probability curve of IQ's on the Kuhlmann-Finch Intelligence Test. Aside from these precautions no cultural, socio-economic or geographic precautions were taken in standardizing the test. The PPVT's standardization was conducted in 1958 and consisted only of white children from Nashville, Tennessee. The normative data was broken down into nineteen age levels and was based on a total sample of 4,012 children.

Another basis for questioning the reliability of the PPVT is the high standard error of measurement (SE_m) data for certain age level children. The normative data for fourth-grade (nine year old) children was based on

259 children and an equivalent forms reliability coefficient of r=0.74 was obtained. SE_m for IQ scores was ±7.65. A SE_m was calculated by using equivalent forms reliability data, which assumes that the two forms are equivalent. If test-retest reliability coefficients were used instead, and indeed were lower than the equivalent forms reliability coefficients, then the already high SE_m would increase even more. The equivalent forms reliability coefficients of the total sample ranged from r=0.67 to r=0.84. SE_m ranged from ±6.00 to ±8.61 for the total standardization sample. The reliability coefficient of the nine year old group (r=0.74) placed it in the lower third of the standardization sample relative to the other reliability coefficients. When the PPVT is used as an initial appraisal and placement instrument, test-retest reliability and SE_m data are essential if accurate interpretations and predictions are to be made from the test scores.

The original standardization sample contained no Spanish-American children, or children raised in the cultural climate of the Southwest. It is reasonable to question whether a test standardization would yield equivalent data based on a sample of children living in a bilingual and bicultural environment such as is found in Albuquerque. Even without the obvious language variable, it is equally reasonable to question whether the children in Albuquerque, New Mexico, or the Southwestern

part of the United States in general (regardless of ethnic consideration) share a common vocabulary with the Southern standardization sample.

Another factor which must be considered is the fifteen year time lapse since the original test construction and standardization. After this length of time, word meanings and word usage have inevitably changed, even for the original standardization sample. Any one or combination of these factors could render the PPVT unreliable with certain populations.

The specific purpose of this study will be to determine the test-retest reliability coefficients and SE_m of the PPVT based on a sample of 40 fourth-grade public school children in one elementary school in Albuquerque, New Mexico.

Chapter 2

PROCEDURE

Sample

The sample was selected from fourth-grade students at an elementary school in Albuquerque, New Mexico. The school has a total population of 401 students in grades one through six. Three fourth-grade classrooms have a total population of 77 children. The sample consisted of 40 children in the fourth-grade who are between 9-0 and 9-11 years of age, and who have never been referred for Special Education. The mean chronological age of the sample is 9 years 6 months. The school is located in a lower middle class area with a variety of ethnic representation (see Table 1).

Table 1

Ethnic Representation of Sample Elementary School

Ethnic Group	Frequency	
	Number	%
Native American	4	1.0
Negro	8	1.9
Spanish Surnamed	68	17.0
Oriental	2	.6
Other	319	79.5
Total	401	100.0

Instrument

The Peabody Picture Vocabulary Test consists of a booklet containing 150 plates with four illustrations per plate. The subject is read a word and his task is to identify which illustration is appropriate for that word. Basal and ceiling levels are established and based on raw scores and chronological age of the subject, an IQ score, a Percentile score, and a Mental Age (MA) score are determined from standardization data published in the manual.

Methods

The subjects were tested by two examiners. One examiner tested 33 subjects and the second examiner tested seven subjects. Form A of the PPVT was administered to a subject and in not less than seven days nor more than ten days, a second administration of Form A by the same examiner was given.

Data Analysis

The study utilized PPVT (Form A) raw scores and the converted IQ, Percentile, and MA scores from both administrations of the test. Means, standard deviations, and SE_m of raw scores, IQ scores, Percentile scores, and MA scores were calculated for the total sample.

Chapter 3

RESULTS

The specific purpose of this study is to determine the test-retest reliability coefficients and SE_m of the PPVT based on a sample of 40 fourth-grade public school children in one elementary school in Albuquerque, New Mexico.

Means and standard deviations of raw scores, IQ scores, percentile scores, and MA scores based on two administrations of the PPVT (Form A) are presented in Table 2. Coefficients of test-retest reliability and SE_m for raw scores, IQ scores, Percentile scores, and MA scores based on two administrations of the PPVT (Form A) are presented in Table 3. SE_{m1} was calculated using the mean of the two actual standard deviations, and SE_{m2} was based on the theoretical standard deviation of 15 IQ points.

Analysis of raw scores indicated twenty subjects (50% of sample) scored higher on the second administration of the test. Fifteen subjects (37.5%) scored lower on the second administration and five subjects (12.5%) scored the same on both administrations. The range of raw scores for the 20 subjects with improved performance

Table 2

Means and Standard Deviations of Raw Scores, IQ Scores, Percentile Scores, and MA Scores for Two Administrations of the PPVT (Form A)

Test Admins.	Scores							
	Raw		IQ		Percentile		MA	
	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.
$PPVT_1$	79.42	9.24	109.13	14.52	66.33	25.85	129.27	24.03
$PPVT_2$	80.22	7.95	110.08	12.61	70.15	23.08	131.02	20.56
$\bar{X}$	79.82	8.59	109.60	13.56	68.24	24.46	130.14	22.29

Table 3

Coefficients of Reliability and SE_m for Raw Scores, IQ Scores, Percentile Scores, and MA Scores for Two Administrations of the PPVT (Form A)

Statistical Analysis	Scores			
	Raw	IQ	Percentile	MA
r	0.795	0.769	0.751	0.807
SE_{m1}	±3.89	±6.52	±12.21	±9.79
SE_{m2}	—	±7.22	—	—

was from 1 to 12 (raw score increase), with a 5.00 mean increase. The range of raw scores for the 15 subjects with declining performances on the second administration was also 1 to 12, with a 4.53 mean decline.

Chapter 4

DISCUSSION

The purpose of this study was to determine the test-retest reliability coefficients and SE_m of the Peabody Picture Vocabulary Test based on a sample of fourth-grade public school children in one elementary school in Albuquerque, New Mexico.

The coefficient of reliability for the Albuquerque sample of $r=0.769$ provides additional data that the PPVT can be used as a general screening instrument with moderate reliability, but other supportive instruments must be used for an accurate diagnosis and evaluation. The high SE_m for IQ scores (± 6.52 and ± 7.22 using actual and theoretical standard deviations) and for MA scores (± 9.79) suggests that extreme caution be exercised when using the PPVT for placement with nine year old fourth-grade children in Albuquerque.

Comparisons of reliability coefficients and SE_m between the Albuquerque sample and the Nashville standardization sample should be made cautiously. Caution is necessary for three reasons. First, the original coefficients of reliability were calculated for equivalent forms reliability, and this study utilizes test-retest

reliability. Secondly, the original standardization SE_m data was based on the theoretical standard deviation of 15 IQ points, and this study used both actual and theoretical standard deviations in calculating SE_m. Thirdly, the original standardization sample was selected to approximate a normal probability curve of IQ's, and this study is based on a sample whose mean intelligence, based on their PPVT scores, is ten points above the theoretical one hundred.

However, despite these differences, some conclusions can be drawn. In the original standardization an equivalent form reliability coefficient of $r=0.74$ for raw scores and a SE_m of ± 7.65 for IQ scores was calculated for the nine year old age level. This is comparable to the test-retest reliability of $r=0.795$ for raw scores and the SE_m of ± 6.52 for IQ scores found for the Albuquerque sample. This SE_m of ± 6.52 was based on the actual mean standard deviation for IQ points (± 13.56) of the Albuquerque sample. If the theoretical standard deviation of 15 IQ points was used in this study as it was in the original standardization data, a SE_m of ± 7.22 for IQ scores would be obtained. This would make the Albuquerque SE_m (± 7.22) for IQ scores even closer to the Nashville, Tennessee SE_m data (± 7.65). This study clearly indicates that there is no significant difference in equivalent form and test-retest reliability coefficients, and SE_m between the Nashville and Albuquerque samples for nine

year old children.

The smaller standard deviation of IQ scores (13.56) for the Albuquerque sample can be explained in part because of the local sample selection criterion. All subjects from either extreme of the intelligence range were excluded from the sample. The sample represented a narrower range of abilities between subjects, hence the standard deviation was less than the theoretical 15 points used in the Nashville standardization.

The mean raw score of the nine year olds on the original standardization was 71.29, with a standard deviation of 9.03. The mean raw score of the Albuquerque sample when averaged for both administrations of the PPVT was 79.42, and the standard deviation was 8.59. The higher mean raw score and smaller standard deviation for the Albuquerque sample is again attributable to the elimination from the local sample of all children at the lower end of the intelligence range.

In analyzing the data it is important to note how the subjects performed on the second administration in comparison to their performance on the first administration. Fifteen subjects declined in performance on the second administration, twenty subjects improved performance, and five subjects stayed the same. If all of the subjects had increased or decreased in performance in no discernible pattern other than a general direction, the accuracy of the reliability data would be questionable.

But, in this study, the average increase (5.00) was very close to the average decrease (4.53). Further, the range of increase and decrease was identical (1 to 12 raw score points), and approximately the same number went up as went down. This superficial analysis would indicate that the scores are evenly distributed.

Due to the objective nature of the PPVT, and because the same examiner gave both administrations to each subject tested, inter-scorer variability is not significant. The two examiners made every attempt to administer the tests in exactly the same manner to each subject tested. The subjects received the second test, whenever their attendance made it possible, on the same day of the week, under as close as possible physical conditions, and with the same level of positive reinforcement that he received on the first administration.

Other researchers including Dunn and Brooks (1960) have also found higher reliability coefficients for MA scores than for IQ scores. In this study the higher reliability coefficient may be due in part to the range of MA scores within the sample. The mean standard deviation for MA scores was a fairly large 22.29, and the wider the range of scores within a sample the higher resulting reliability coefficients.

In many standardized tests such as the PPVT, raw score data can often be the most informative. Raw scores

are important because they are usually most reflective of a change in a subject's performance. IQ scores, percentile scores, and MA scores are all directly based on the subject's raw scores. Changes in IQ scores, percentile scores, and MA scores are less accurate because they may reflect not only changes in the subject's performance, but also clustering within the standardization. The clustering is reflected in the lower reliability coefficients for IQ scores (r=0.769) and percentile scores (r=0.751). The slightly higher reliability coefficient for MA scores (r=0.807) than for raw scores (r=0.795) is due to the wider range of MA scores.

Further research is necessary to determine if test-retest reliability coefficients on children at other levels of the intelligence range are similar to the coefficients found in this study. The subjects of this sample were approximately ten IQ points above the mean based on the results of the PPVT, and it is reasonable to question whether children who are ten IQ points below the mean would perform as consistently on a second administration of the test.

This study indicates that the PPVT does not have a high enough test-retest reliability to warrant its use as the major placement instrument with nine year old children in the Albuquerque area. Furthermore, the use of the PPVT as a required test in the appraisal and evaluation of physically handicapped individuals

is highly questionable. Further research on the test-retest reliability of the PPVT on physically handicapped children is needed. This study also found that there is no significant difference between the reliability coefficients published with the PPVT and the coefficients calculated for the Albuquerque sample. At their best the results of the PPVT should be cautiously interpreted, especially when the test is being used as a placement instrument.

APPENDIX

Review of the Literature

In three studies using regular school children as subjects only one dealt with test-retest reliability. Raskin & Fong (1970) investigated the temporal stability of the PPVT in normal and educable-retarded children. Using 41 normal subjects a test-retest reliability of r=0.72 was obtained for IQ scores and a reliability of r=0.87 was obtained for MA scores. The coefficients for the educable-retarded sample of 42 subjects were r=0.80 for IQ scores and r=0.89 for MA scores.

Norris, Hottel, & Brooks (1960) administered alternate forms of the PPVT under group and individual conditions to 60 fifth-grade children, and concluded that the two forms were indeed equivalent. The raw scores were so close for the two administrations of the test that reliability coefficients were not reported.

Tempero & Ivanoff (1960) administered alternate forms of the PPVT under group and individual conditions to 150 seventh-grade students, and calculated a reliability of r=0.75 for the equivalency of IQ scores between the two forms.

Bashaw & Ayers (1967) conducted research to evaluate the norms and reliability of the PPVT for pre-school subjects. They found test-retest reliability coefficients of r=0.69 for raw scores, r=0.69 for MA scores and r=0.56 for IQ scores using Form A of the PPVT, given twice with a four month interval between testings. In the same study equivalent forms reliability coefficients of r=0.75 for raw scores, r=0.71 for MA scores and r=0.37 for IQ scores were obtained for 131 subjects.

Moed, Wight, & James (1963) administered the PPVT to 29 crippled children after one year in a hospital. A coefficient of r=0.88 was reported for the stability of IQ scores.

Dunn & Harley (1959) administered both forms of the PPVT, one week apart, to 20 cerebral palsied children aged 7-1 to 16-2. An equivalent forms reliability of r=0.97 for MA scores was calculated.

Mandel & McLeod (1970) conducted a longitudinal investigation of the stability of IQ's on the PPVT with high and low socioeconomic subjects. Form B of the PPVT was individually administered four times over a five year period to 253 pre-school subjects. At the end of the five year period 199 subjects were still available for testing. Sixty-nine subjects were selected from the socioeconomic extremes and for the total group an average retest correlation, over the four testing

periods, of r=0.53 was obtained. Correlations for the high socioeconomic group ranged from r=0.23 to r=0.60, and for the low socioeconomic group between r=0.53 and r=0.79.

Goldstein, Coller, Dill, & Tilis (1970) in an extended study with disadvantaged children used the PPVT as one of three instruments used to calculate the stability-reliability coefficients for experimental and control subjects tested over a five year period. Four administrations were given with the initial testing begun prior to the subjects beginning kindergarten. The PPVT was found to have reliability coefficients of r=0.66, r=0.61, and r=0.71 for the three subsequent administrations following the initial pre-kindergarten pre-test. These coefficients were obtained for the control group and offer valuable data on the stability of PPVT scores over a long period of time.

Kimbrel (1960) used 62 mentally retarded pupils aged 10-5 to 15-8, IQ's 40 to 62, from a state residential facility and calculated an equivalent forms reliability of r=0.86 for IQ scores.

Dunn & Hottel (1961) had teachers administer both forms of the PPVT one week apart to 220 trainable retardates. A reliability coefficient of r=0.84 for MA scores was reported.

Budoff & Purseglove (1963) using 46 institutionalized retardates, aged 16 to 18 years, calculated

a coefficient of r=0.85 for the equivalency of MA scores and a coefficient of r=0.87 for the stability of MA scores after one month.

Blue (1969) investigated the one year temporal stability and alternate form reliability of the PPVT (Form A) using 116 subjects, aged 6-6 to 32-8, in a school for the trainable mentally retarded. Blue concluded that a high raw score reliability was demonstrated in both alternate form testing (r=0.92) and one year interval test-retest (r=0.92) regardless of the form of scores employed or age groupings.

Dunn & Brooks (1960) administered both forms of the PPVT to 371 educably mentally retarded pupils. Forms were given one week apart in counter-balanced order by the same examiner, and a reliability of r=0.83 was reported for MA scores, with a coefficient of r=0.61 for IQ scores.

Kahn (1966) investigated the long-term reliability of the PPVT with adolescent and young adult retardates. The subjects were tested once a year for four years. Reliability coefficients of r=0.71, r=0.77, and r=0.80 were calculated for the three administrations following the initial first year testing. Kahn concluded that long-term reliability coefficients were approximately equal to the coefficients reported for short-term reliability.