AN ASSESSMENT OF THE RELATIONSHIP BETWEEN RELIABILITY AND VALIDITY OF SOME SELECTED PSYCHOLOGICAL TESTS

Jonathan A. Odukoya
West African Examinations Council
P.M.B. 1076
Yaba - Lagos.

ABSTRACT

This study investigated the relationship of reliability and validity. Three standardised psychological tests were administered on 268 Nigerian secondary students spread over forms one, three and five. The tests were administered twice within a period of six weeks to allow for the computation of test-retest reliability. The Cronbach's r, the Split-half r, the Spearman-Brown r and the Kuder-Richardson (KR) 20 r were also derived. Content and construct validities were established through Z-test, t-test, factor analysis and correlation coefficient. It was concluded that when it is said that valid tests are reliable, it is the internal consistency reliability that is being referred to rather than the test-retest reliability. In essence, not all valid tests will necessarily furnish indices of reliability on all types of reliability coefficients. Apart from raising some issues for further study, the finding, has important implications for psychological tests' development and usage.

Introduction

The problem of inaccurate measurements has been a persistent one in the fields of education and psychology. It is towards finding solution to this critical problem that the concepts of validity and reliability emanated. The common conception has been that validity is all that matters for any one test. Reliability is regarded as being subsumed under validity (Salvia and

Yasseldyke, 1978). But gradually, the concept of reliability began to acquire more significance, such that it is now virtually being treated on the same pedestal as validity. According to Salvia & Ysseldyke (1978),

> ... all valid tests are reliable, but no unreliable test may be valid. Reliable tests may or may not be valid (p. 104).

This study was designed to test this statement. One may therefore ask, is it true that, at all times, a valid test is reliable while an unreliable test is never valid?

Tests and measurements are sensitive issues as the results of measurements, especially psychological ones, to a large extent, determine the fate of the respondent(s). The seriousness of this issue is further confirmed with reports on students who commit suicide because of unexpected examination results. Many are 'forced' into 'wrong' professions, while a host of others become frustrated and maladjusted. These issues have some bearing on the psychometric properties of the instruments in question.

A psychological test is essentially an objective and standardized measure of a sample of behaviour. No psychological test can do more than measure behaviour. Nunnally (1972) described a test as a standardized situation that provides an individual with a score. Taking a different dimension. Bakare (1988) defined a psychological test as a stimulus presented to an individual so as to elicit a response on the basis of which a judgement is made on certain attributes and abilities possessed by that individual.

Psychological tests are classified in many ways, depending on the criterion for testing. One of the commonly used classification is that of psychomotor (capacities involving motor abilities), cognitive (capacities dealing with knowledge and the acquisition and utilization of information), and affective (feelings and values). Another acceptable classification is that of Aptitude, Achievement and Personality tests. Aptitude and Achievement tests are conventionally grouped as Ability tests.

Reliability and validity are indispensable constructs in psychological testing. According to Ghiselli (1964), the reliability of measurement is the extent of unsystematic variation in the scores of an individual on some traits, when that trait is measured over a number of times (p. 218). In other words

test results would have little meaning if they fluctuate widely from one testing situation to another. For a psychological test to be useful, it must measure what it is measuring consistently. In its broadest sense, test reliability indicates the extent to which individual differences in test scores are attributable to chance error of measurement (Anastasi, 1961). In other words, reliability refers to the extent an examinee would be expected to earn similar scores taking the same test again on different days, or occasion. This highlights the different types of reliability indices: the test-retest reliabilities, the parallel test reliabilities and internal consistency reliabilities. For the purpose of this study, the test-retest reliability, the alternate-form reliability and the internal consistency reliabilities were derived.

The subject of test validity, however, concerns what the test measures and how well it does so. In the process of validation, one examines the soundness of all the interpretations one can make from a test score. According to Cronbach (1971), validation is a process for developing sounder interpretations of observations. In essence, one validates, not a test, but an interpretation of data arising from a specified procedure. So, a test may be valid for one use, but this is no assurance that it will be valid for other uses. Salvia and Ysseldyke (1978) appear to support this view when they explained that a test's validity is not measured; rather a test's validity for various uses is judged on the basis of a wide array of information. Fundamentally, all procedures for determining test validity are concerned with the relationship between performance on the test and other independently observable facts about the behaviour characteristic or construct under consideration. The American Psychological Association (APA, 1974) classified the various methods of estimating test validity under four categories. Content validity, predictive validity, concurrent validity, and construct validity. In actual practice, these approaches complement one another.

As mentioned earlier, this study has been designed primarily to assess the relationship between reliability and validity.

METHOD

Subjects

Two hundred and sixty-eight (268) subjects consisting of 123 females and 145 males were initially sampled for this study. The subjects were drawn

from classes 1, 3 and 5 of the United High School, Ijokodo, (an urban school) and Moniya-Aponmode High School, Moniya (a rural school). The age of the subjects ranges between 11 years and 20 years.

Instruments

Though eight psychological tests spreading over mental ability, achievement and personality tests were originally used in this study, only three of these were used in testing the hypothesis of this study. These are the Study Habit Inventory (SHI), the Academic Need Achievement Scale (ANAS) - both by Bakare (1977, 1976), and the Standard Progressive Matrices - which was validated for Nigeria. The SHI is designed to identify defective or poor study habits in students of secondary school age. It is a self-report inventory which enables the individual student to describe the situations, habit and conditions which affect his use of study time. There are 8 sections in this 45-item Likert-type scale. Scoring is done by assigning a score ranging from 1 to 5 to each response depending on whether the item is positively or negatively stated. The higher the total score the more pronounced the study problem. The test-retest reliability of the SHI was reported to be .83 (for $N = 58$, $p < .05$ with 3 weeks interval).

The ANAS measures students' level of motivation towards academic achievement. The scale consists of 36 items to which the subjects respond on a five-point Likert scale. Scoring is done by assigning a score ranging from 1-5 to each response depending on whether the item is positively or negatively stated in terms of enhancing academic achievement motivation. High score indicates high achievement motivation. Hassan (1982) offered evidence of the construct validity of ANAS when he observed that the instrument discriminates significantly between passing and failing students ($t = -2.57$, $p < .05$, $df = 120$).

The SPM was originally compiled by Ravens (1958). It is a non-verbal culture-fair test of intelligence. Respondents are expected to select, out of a group of designs, the design that best fit the empty space within a larger design that is serving as the stem question. There are 5 sections of 12 items each in the test. According to Burke and Bingham (1969), the evidence of the construct validity of the test is given by the fact that it correlated .76 with the Wechsler Adult Intelligence Scale. bakare (1976) also validated this

instrument for Nigeria. Scoring is done by simply adding together the number of items answered correctly based over 60. The higher the score, the more 'intelligent' the respondent is assumed to be.

These tests were administered twice within a period of 6 weeks to allow for the computation of the Pearson-Product Moment Correlation Coefficient and consequently the test-retest reliability. Other related reliability coefficients as well as validity indices were also established in the process.

RESULTS AND DISCUSSION

In order to establish the nature of the relationship between reliability and validity and, in particular to ascertain whether 'all valid tests are always reliable', further evidence of the validity and reliability of the SPM, the SHI and the ANAS were sought. To this end, it was decided that:

(i) If for every test sufficiently judged to be valid, there are sure evidence(s) of its reliability, then the submission that 'all valid tests are reliable' will be held.

(ii) However, if there was found anyone valid test not having sufficient evidence of reliability, then the commonly held belief that 'all valid tests are reliable' will be debunked. From the results furnished in Tables 1 through Table 7 below, coupled with the psychometric data furnished in their respective Manuals, it was ascertained that the SPM, the SHI and the ANAS are valid psychological instruments.

Table 1
Validity and Reliability Estimates for the Standard Progressive Matrices (SPM)

	URBAN						RURAL					
	CLASS 1			CLASS 5			CLASS 1			CLASS 5		
	r	N	P	r	N	P	r	N	P	r	N	P
Validity Construct (with AP)	.108	23	ns	.167	29	ns	.372	30	.05	.233	29	ns
RELIABILITY i. Split-Half	.51	31	.01	.78	38	.01	.63	43	.01	.96	31	.01
ii. Spearman-Brown	.68	31	.01	.88	38	.01	.77	43	.01	.978	31	.01
iii. KR 20	.92	31	.01	.92	38	.01	.75	43	.01	.93	31	.01
iv. Cronbach's	.94	31	.01	.947	38	.01	.84	43	.01	.95	31	.01
v. Retest (1 week)	-	-	-	-	-	-	.874	31	ns	.94	14	.01
vi. Retest (1 month)	.627	28	.01	-	-	-	.85	31	.01	.69	14	.01

NOTE: 'AP' implies Academic performances; 'ns' implies non-significant at .05 level.

Table 2
Effect of Class-in-School on Intelligence Using the SPM on Urban Pupils

Class	N	Mean	SD	df	t
1	31	18.07	11.47	67	4.25
5	38	30.42	12.45		

Table 3
Intercorrelation among Subsections of the SPM for Rural Class 5 Pupils (N=14)

	A	B	C	D	E
A	1,000				
B	0.615*	1.000			
C	0.155	0.670*	1.000		
D	0.344	0.754*	0.747*	1.000	
E	0.663*	0.705	0.298	0.709*	1.999

Note: The asterisked is/are significant at $P < .05$.

Table 4

Validity and Reliability Estimates for the SHI

	URBAN						RURAL					
	CLASS 1			CLASS 5			CLASS 1			CLASS 5		
	r	N	P	r	N	P	r	N	P	r	N	P
Validity Construct (with AP)	.35	23	ns	.29	.29	ns	011	30	ns	.25	29	ns
RELIABILITY i. Split-Half	.50	26	.01	.83	33	.01	.95	40	.01	.68	30	.01
ii. Spearman-Brown	.67	26	.01	.91	33	.01	.97	40	.01	.81	30	.01
iii. KR 20	-	-	-	-	-	-	-	-	-	-	-	-
iv. Cronbach's	.56	26	.01	.62	33	.01	.67	40	.01	.75	30	.01
v. Retest (1 week)	.141	4	ns	-	-	-	.175	26	ns	.706	9	ns
vi. Retest (1 month)	.157	4	ns	.005	19	NS	.379	26	ns	.573	9	ns

Table 5

Intercorrelations among Subsections of the SHI for Urban Class 5 Pupils (N = 42)

	A	B	C	D	E	F	G	H
A	1.000							
B	0.849	1.000						
C	0.818	0.919	1.000					
D	0.829	0.910	0.935	1.000				
E	0.654	0.774	0.812	0.778	1.000			
F	0.776	0.870	0.856	0.828	0.683	1.000		
H	0.805	0.853	0.816	0.839	0.605	0.879		1.000

Table 6

Academic Achievement Motivation of Passing and Failing Students for Urban Class 1 Pupils

Group	N	Mean	SD	t	df	P
Passing	10	109.8	8.848	2.248	18	.05
Failing	10	99.0	10.54			

Table 7

Validity and Reliability Estimates for the ANAS

	URBAN						RURAL					
	CLASS 1			CLASS 5			CLASS 1			CLASS 5		
	r	N	P	r	N	P	r	N	P	r	N	P
Validity Construct (with AP)	.375	23	.05	-.375	29	.05	.265	30	ns	.272	29	ns
RELIABILITY i. Split-Half	.55	34	.01	.49	21	.05	.80	41	.01	.84	33	.01
ii. Spearman-Brown	.709	34	.01	.66	21	.01	.89	41	.01	.91	33	.01
iii. KR 20	-	-	-	-	-	-	-	-	-	0	-	-
iv. Cronbach's	.65	34	.01	.72	21	.01	.56	41	.01	.55	33	.01
v. Retest (1 week)	-	-	-	-	-	-	.072	31	ns	.678	12	.05
vi. Retest (1 month)	.624	20	.01	.326	5	NS	-.103	31	ns	-.21	12	ns

Evidence for the validity of the SPM, SHI and ANAS was sought by establishing their construct and content validity. For SPM, it was hypothesized that, as an intelligence test, the instrument should correlate positively with Academic Performance (AP); respondents from the senior class 5 should perform significantly better than the junior respondents (that is, those in class 1), on the SPM, and that the intercorrelation amongst the subsections of the instrument should be fairly high – being an homogeneous test. These were confirmed as in Tables 1 to 3.

For the SHI, it was hypothesized that the more study problems a student is experiencing, the poorer he is likely to perform academically; and since a high score on the SHI depicts less study problems, then the higher the score obtained on this instrument, the better the respondent is expected to perform academically. Hence the SHI is expected to correlate positively with academic performance. It was also hypothesized that if a test can correlate fairly well with an already standardized test measuring similar construct then the new instrument is very likely to be valid; and also, as an instrument having fairly homogeneous items, it was expected that the intercorrelations among its subsections should be predominantly positive. Tables 4 and 5 confirmed these thus lending support to the validity of the SHI.

For the ANAS, it was assumed that, being an effective test measuring pupils' motivation towards academic achievement, it should be able to discriminate significantly between passing and failing students. It should also correlate positively with respondents' academic achievement in school. Tables 6 and 7 above furnished enough evidence to confirm these hypotheses, hence the validity of the ANAS. It is therefore concluded from the indices, of the Split-half, Spearman-Brown, KR-20, and Cronbach's reliability coefficients obtained (see Tables 1, 4 and 7), that valid tests are reliable.

In discussing this finding, the following two premises are considered. First, the expression that 'all valid tests are reliable' could imply that, for any test to be valid, it must be reliable. The second premise is that it could also imply that 'no valid test could be unreliable', that is, no valid test could have a reliability coefficient that is not significant irrespective of the type of reliability.

It is probably in this perspective that Salvia and Ysseldyke (1978) opined that unreliable tests are measuring errors. The questions, however, are: would

it be appropriate to unequivocally declare that a test is measuring error (that is, not valid) simply on the ground that it is not having significant index of a particular reliability, in this case, the test-retest r'. And what if such an instrument is having significant index of other type of r's (example, internal consistency r's) as observed in the present study?

Answers to these questions were partly provided by Anastasi (1961) and Stanley (1971). Anastasi submitted that, for the large majority of psychological tests the retest r is not suitable; it presents difficulties when applied to most psychological tests. In the same vein, Stanley observed that, in reality; there is no single, universal and absolute reliability coefficient for a test. There could be as many varieties of test reliability as there are conditions affecting it. Going by this submission, it could then imply that, once a test is showing evidence of reliability on any of the reliability estimates, to that extent it is reliable. In essence, the tests used in this study, that are ascertained to be valid and are showing evidence of at least internal consistency reliability, are indeed reliable. And if this is so, then the common statement that 'all valid are reliable' will hold.

On the other hand, this appear to contradict the second premise or perspective that 'no valid test could have a reliability coefficient that is not significant'. Most of the tests used here show no evidence of test-retest reliability. This is in part explained by Anastasi's (1961) view that for tests involving reasoning and ingenuity, its nature change with repetition. This is because once the respondent has grasped the working principle, he automatically responds the same way when the test is repeated. But for affective tests (such as the SHI and ANAS), this is hardly so. On the contrary, the respondents somehow find it difficult making accurate recall of their previous response since there are no basic underlying principle to follow – hence the unduly low, erratic and non-significant retest r's observed for these instruments. Field observations also revealed that respondents were generally unco-operative and bored with repeated tests, especially for the verbal affective tests. These points serve to confirm that it might be erroneous to conclude that a test is not reliable (hence, not valid), simply on the ground that it is not having a significant test-retest r while having ample evidence of reliability on other reliability estimates. The work of Jaworska & Szustrowa (1993) are very relevant to this.

CONCLUSION AND RECOMMENDATION

It will be reiterated that nothing should be taken as absolute about the common statement that: all valid tests are reliable and that no unreliable test could be valid. A lot depends on the perspective from which one is interpreting this statement. Indeed, from a theoretical standpoint, for any test to be valid, it ought to measure what it purports to measure consistently; it is however not true that a valid test could not have a reliability coefficient that is insignificant. The relationship between reliability and validity cannot be held in absolute terms. As Guilford submits, psychological tests' developers and users need be relativists when dealing with problems of reliability and validity.

References

American Psychological Association (APA) (1974) *Standards for Educational and Psychological Tests.* Washington DC: APA.

Anastasi, A. (1961): *Psychological Testing* (2nd ed.) New York: Macmillan.

Bakare, C.G.M. (1976): *Academic Need Achievement Scale (ANAS)* Ibadan: University of Ibadan Press.

Bakare, C.G.M. (1977): *Study Habit Inventory (SHI).* University of Ibadan.

Bakare, C.G.M. (1988): *The Use of Psychological Tests in Guidance and Counselling.* Background Paper Prepared for Lagos State Workshop on Use of Psychological Tests.

Burke, A. and Bingham, T. (1969): In Vehaza (1974) *The Relationship Between Some Personality factors and Academic Performance Among Some Nigerian Secondary School Students.* Ph.D. Dissertation, University of Ibadan.

Cronbach, L. (1971): In Thorndike, P.L. (ed.) *Educational Measurement.*

Washington: American Council of Education.

Ghiselli, E. (1964): *Theory of Psychological Measurement.* New York: McGraw Hill Book Co.

Guilford, J.P. (1956): *Psychometric Methods.* New York: McGraw-Hill Book Co.

Guilford, J.P. (1956): *Fundamental Statistics in Psychology and Education* (3rd ed.) New York: McGraw-Hill Book Co.

Hassan, Titi (1982): *Effect of Response* set on the Psychometic Properties of some Personality Inventories Unpublished Ph.D. Dissertation, Department of Guidance and Counselling, University of Ibadan.

Jaworska, A. & Szustrowa, T. (1993). Polish Standardization of Raven's Progressive Matrices *Polish Psychological Bulletin* Vol.24 (4) 303-307.

Nunnally, J. (1972): *Educational measurement and Evaluation* New York: McGraw-Hill Book Co.

Ravens, J. (1958): *Progressive Matrices* London: H.K. Lewis.

Salvia, O. and Ysseldyke, A. (1978): *Assessment in Special and Remedial Education.* Boston: Houghton-Mifflin Co.

Thorndike, R. and Hagin, E. (1977): *Measurement and Evaluation in Psychology and Education* (4th ed.). New York: John Wiley and Sons.