THE CONSTRUCTION, VALIDATION AND STANDARDIZATION OF CRITERION REFERENCED TEST

IWEKA FIDELIS
Department of Educational Psychology, Guidance and Counseling
University of Port Harcourt
Port Harcourt

Abstract

The paper is on the construction, validation and the standardization of criterion – referenced test. It highlighted an introduction into criterion – referenced test. The researcher took time to explain the different methods of constructing – criterion referenced test, established five methods of determine reliability of criterion referenced test and also the different methods of determining the validity of the test. The use of table of specification in determining content validity of criterion – referenced test was fully discussed. Finally conclusion of the paper was given.

Introduction

Test may be seen as an integral aspect of feedback process that helps in quality assurance and control at any level of educational system. Test according to Onunkwo in Iweka (4014;1) in defined as an instrument which can be utilized in detecting some qualities, traits, characteristics, attributes etc. possessed by a person, an object or a thing, while Ukwuije (2019;5). Saw it as series of questions given to the tastes or examinees to be answered.

Orluwene (2012) states that test can be regarded as an instrument used to determine the relative presence or absence of the trait measured for or it can be a measurement instrument or device administered to someone to determine the relative value of the traits or skill to which the test relates. Ukwuije (2009; 6) states that test in education is used for administrative, instructions, guidance and counseling, research, prediction, diagnosis and evaluation purposes. It therefore follows that test is used by different people in different ways which will enable them take some form of action. Criterion Referenced test according to Joap, Cees and Sally (2003) is used to give clear indication about which educational content is mastered when a particular score and it is based on performance standard.

The scores gotten from the test could be used for different decision making, therefore, there is a strong call to validate test instrument to enable the test measure what it ought to measure. In order to get this done, estimation of reliability and validity of the test become very imperative. Iweka (2014) identified five methods of establishing reliability to include: test-retest split half, parallel form, interpreter, and interpreter agreements. To get this done, one is expected to calculate the Proportion of Observed Agreement (A), Proportion of Chance Agreement (AC) and the Proportion of Agreement Uncontaminated by chance (K). Method of calculating A is given thus:

1) Total Subjects/Examines Consistently Classified on two Administration

The formula used in establishing total subject agreement is given

Thus: A= a+d/N

Where A- proportion of agreement coefficient

a- Number of students consistently classified as masters

d- Number of students consistency classified as non-masters on the two administrations of the test

N- Total number of tastes.

E.g.: in a test given in order to test student' ability to mount an air conditioner given a cut-off point of 70%, 10 students got the activity right in two cases while 15 students got it wrong in two occasions available to them. In the same activities, 6 students got it right in the first but not in the second, while 4 got it wrong in the first and not in the second occasion. Determine coefficient of agreement, by chance and uncontaminated.

Solution:

i) Coefficient of Administration is given thus:

A= a+d/N

10- Masters

15- Non masters

A= 10+ 15/35=0.71

This means that 71% of the students were consistently classified as masters and non-masters in the test.

ii) Calculation of A with proportion is given thus: A = a+d

a- Proportion of masters

b- Proportion of non-masters

iii) Method of Calculation of the proportion of chance agreement (AC)

It is therefore, imperative to determine those who have scores as a result of chance. This could be done using total subjects classification method or by using the proportion of subjects' classification method. Using total subjects classification, we can apply this formula

AC = $\frac{(a+b)(a+c)+(c+d)(b+d)}{N2}$ - total subjects classification method

AC = (a+b)(a+c) + (c+d)(b+d) - proportion of subjects classification method

AC = chance agreement

a = masters of two administration of test

b = number that are masters in second administration but not in first

c = number that are masters in first administration but not in second

d = number of students who showed non-mastery in the two administrations

N = total number of subjects from our example above

AC = (.29+.17) (.29+.11) + (.11+.43)(.17+.43)

AC = 0.46 x 0.4 + 0.54 x 0.6

AC = 0.184 + 0.324 = 0.51

This implies that 51% of the observed agreements were attributed to chance.

3) Method of calculation of proportion of agreement uncontaminated by chance (k)

Proportion of uncontaminated agreement by chance using Kappa coefficient (K) Harbors the observed Agreement (A) and Agreement by chance (AC)

The kappa coefficient (k) by Iweka (2014; 146) corrects for chance agreements uncontaminated and is given by the formulae:

$K = \frac{A-AC}{1-AC}$ where

A = proportion of agreement or agreement coefficient
AC = proportion chance agreement/chance agreement coefficient i.e., one has to solve for observed agreement and chance agreement before uncontaminated chance agreement.

$AC = \frac{0.73-0.51}{1-0.51}$ where

Where 0.73 = proportion of subject agreement
0.51 = agreement by chance

$K = \frac{0.22}{0.49} = 0.45$

The result shows that 45% of the agreements were uncontaminated by chance

Let us fall back to the five procedures of establishing kappa (k) reliability coefficient for a criterion referenced test.

1) Test – retest techniques of Estimating Reliability:

Orluwene (2012;84) states that test-retest method of reliability estimation involves the correlation coefficient by obtaining two sets of scores for the same students through the administration of the same test on two different occasion. Iweka (2014; 149) asserts that in criterion referenced testing, test-retest has to do with comparing the proficiency testing, test-retest has to do with comparing the proficiency or mastery of a particular skill or objectives being measured when they are administered with a particular test on two separate occasions in a 2-3 weeks intervals, from example above

Administration 1	**Administration 2**			Total
	Masters	10 .29 (a)	6 .17 (b)	(a + b) 29 + .17 16
	Non-Masters	(c) 4 .11	(d) 15 .43	(c + d) 11 + 43 19
Total		(a + c) .29 + .11 14	(b + d) .17 + .43 21	a + b + c + d 35 1.00

Remember the various formulas

1) Proportion of observed agreement (A) from test-retest administration is given:

	Second administration			
First administration		Mastery	Non	Total
	Mastery	10 (a)	6 (b)	16 (a + b)
	Non mastery	4 (c)	15 (d)	19 (c + d)
Total		14 (a + c)	21 (b + d)	(a + b + c + d) 35

Computation with proportion of subjects or examines consistently classified on two classifications.

To get this done, the various scores are converted into proportions using number of tastes. This gives the formula:

A = a + d

A – Proportion of agreement

a – proportion of students consistently classified as masters

d – Proportion of students of non-mastery based two administrations from our example above, a is given as

i) A = a + d

A = 0-.29 + 0.43 = 0.72

By implication, it shows that 72 % of the agreements were uncontaminated by chance

Proportion of chance agreement (AC)

AC = (a + b)(a + c) + (c + d)(b + d)

AC = (.29 + .17)(.29 + .11) + (.11 + 43)(17 + .43)

AC = .51

By interpretation, 51% of the Observed Agreement for mastery and non-mastery was founded on chance.

Agreement Uncontaminated

$$K = \frac{A\text{-}AC}{1\text{-}AC}$$

$$K = \frac{.73\text{-}.51}{1\text{-}.51} = 45$$

This result shows that 45% of the decision or agreement was uncontaminated by chance; the Kappa (K) acceptable proportion is above .50 therefore, the test is not reliable.

2) Split half technique of estimating criterion referenced test reliability

Kpolovie (2010) defines split half reliability as the extent to which two halves of a test taken once truly and consistently measures the same trait. It there means that must be a cutoff point where one falls under mastery and non-mastery. In both cases the odd and even splits are grouped as shown in the table below.

	Second half even			
First half odd		Mastery	Non mastery	Total
	Mastery	10 (a)	6 (b)	16 (a + b)
	Non mastery	43 (c)	15 (d)	19 (c + d)
Total		14 (a + c)	21 (b + d)	(a + b + c + d) 35

Note that all the scores will be converted to proportion using the total number of testes or students. Therefore after, one establishers Observed Agreement (A), compute Agreement by chance (AC) and Kappa Agreement of uncontaminated chance (K).

3) Parallel form techniques of establishing the reliability of a criterion referenced test.

Equivalent forms reliability of the correlation coefficient of two alternative or parallel forms of the same test at about the same time to examines (Kpolovie 2010; 554). Adding to this view, Iweka (2014: 115) submitted that this method of reliability involves administering an equivalent form or exactly the same test on a group of students on either one or two closely spaced occasions. In the reliability estimate methods, we are concerned with determining the degree of agreement in classifying students as masters and non-masters on a particular objective. In a situation of perfect agreement, the student are arrayed in mastery (a) cell and non-mastery (d) cell but in a situation of perfect disagreement the testes fall under b and c. see the table

		Form two test		
First half odd		Mastery	Non mastery	Total
	Mastery	10 (a)	6 (b)	16 (a + d)
	Non mastery	4 (c)	15 (d)	19 (c + d)
Total		14 (a + b)	21 (b + d)	(a + b + c + d) 35

Note that all the score will be converted to proportion using the total score. The proportion of agreement, agreement by chance and uncontaminated by chance using the various suitable formulas stated above.

4) Inter rater agreement reliability for criterion-referenced test

This involves two different raters classifying the students as masters and non-masters on the same or different occasions. The data are analyzed using proportion of agreement uncontaminated by chance (K).

5) Intra-scorer agreement reliability of criterion-referenced test scores

Here, the same rater rates the students two times to group them into mastery and non-mastery giving them two separate scoring.

The Observed

Agreement (A) is gotten using this formula: $A = \frac{A + D}{A+B+C+D}$

Proportion by chance (AC) is gotten thus: $AC = \frac{(A + B)\ (A + C) + (C + D)}{N^2}$

Calculation of the agreement uncontaminated by chance (K) = $K = \frac{A\text{-}AC}{1\text{-}AC}$

Validity of a Criterion Referenced Test

This is the ability of the best to measure what it is design to measure accurately. Most of the times validity does not only measure the degree of accuracy of an instrument but also the accuracy in the interpretation of the scores in relation to what it wants to measure. Techniques for establishing validity of a test

1) **Decision validity techniques:**
This deals with the ability to accurately classify people into mastery or non-mastery independent of the test (true classification) and with the rest (test classification). The double entry classification table gotten will be calculated using classification outcome probabilities (C) as estimated by Baumgartner, Jackson, Mahar & Rowe in Iweka (2014). Another double entry classification table called phi can be used to estimate the validity of the rest

2) **Domain/logical-referenced validity techniques**
Domain here means the criterion ability the test is meant to measure. It is also the same with content validity which covers the subject matter content specified. Ways of establishing domain reference validity are:

i) **Judgment by content specialists:** in this test, the items of the must measure the intended instructional objectives judgment is meant using

a) **Ratings scale method.** Here the test is organized into bad=1, good=2 better=3 and best=4. The content specialist be more than two and instructed to rate and provide the relevance of each item in the test to the traits in the objectives thereafter the mean of each content specialist is calculated which shows whether an item is retained modified or discarded based on a criterion mean given.

ii) **Index of item-objective congruence method:**
Some content experts are requested to independently identify if each of the items in the test is a measure of the make up to the trait in the objects or not. The raters are instructed to assign +1 to any item that measures an objectives while -1 when it is not the measure of the objective, 0 when the measure is not certain Then index of objective congruence is computed for each test item.

Expert raters	Objectives			
	1	2	3	S'O
1	-1	+ 1	-1	
2	0	+ 1	-1	
3	-1	0	0	
4	-1	+ 1	-1	
SO	-3	+ 3	-3	-6

$1_{10} = \frac{(M-1)\ SO-S'O}{2N\ (M-1)}$ FORMULA for index of item –objectives congruence

M-number of objectives
N-number of content specialist (expert raters)
SO = sum of the ratings assigned to objectives 0 i.e. 2 (sum of positive values)
S'O = sum of the table above (sum of negative values)

110 = (3-1)(+3)-(-6)/2*4(3-1) = 2*3+6/8*2 = 12/16 =.75
75% shows that the raters rated item 2 as a measure of the objectives 2 and none rated it as not being a measure of objectives 2

iii) Judgment by content specialist method

Here two experts are requested to rate each test item independently based on the objective under study. A point scale as in our bad =1, good =2, better =3 and best =4 can be used but for a test of 30 items rated by two content specialists.

	Content specialist 1			
Content specialist 2		Mastery 1 or 2	Non mastery 3 or 4	Total
		10 (a)	6 (b)	16 (a + b)
	non mastery 3 or 4	4 (c)		
Total		14 (a + c)	21 (b + d)	(a + b + c + d) 35

Cell a represents the number rated 1 and 2 by the first and second raters, cell b shows The number rated 3 and 4 by the first specialist and 1 and 2 for the second specialist, cell c represents the number of items rated 1 and 2 by the first specialist and 3 and 4 by the second specialist and finally Cell d shows the number of items rated 3 and 4 by the two raters. Then inter raters agreement of a, Ac and K is done. Or Computation of the content validity index could be done using solely Observed Agreements (OA) and Uncontaminated by Chance (K). Authorities recommended OA above .8 and greater than or equal to 0.25 before the computation of CVI. The formula for CVI is given thus: CVI = a+b+c+d/N or d/N for this cell, if 50% is gotten, it shows the 50% of the items were rated 3 or 4 by the content specialists. The 50% shows that 18 items out of 35 items were rated good and best as the case may be.

The Use of Table of Specification In Determining Content Vanity of Criterion-Referenced Tests.

The major assertion here is that this method allows the various content or level of an objective to be represented in the test. In the cognition domain where there is knowledge, comprehension, application, analysis, synthesis and evaluation, if only knowledge is to be measured it only requires some sets of items on knowledge, but if all the levels are to be measured, then the table of specification of blue print is to be used to represent the other levels and major emphasis on topic and behavior which ensures reliability and validity. Kpolovie (2002, 33) listed four main reasons for the construction of test of blueprint to include:

i) To ensure that the test adequately covers only those objectives which were actually implemented during the instructional process?

ii) To ensure that the test items accord a proportional emphasis on each of the objectives in relation to how it was emphasized during instruction.

iii) To guarantee that no important or content matter of the subject to which students were exposed is inadvertently omitted in the test.

iv) To ensure that the test has content validity.

Table 1: Table of specification for a test on motion on physics

Content areas	Kno w	Compre	Applic	Analy.	Synt h	Evalu.	Tota1
Definition of motion	(1) 2	(1) 3	(1.5) 3	(.5) 1	(.5) 1	(.5) 1	(5) 10
Types of motion	(3) 6	(3) 6	(4.5) 9	(1) 3	(1.5) 3	(1.5) 3	(15) 30
Formula for motion	(2) 4	(2) 4	(3) 6	(1) 2	(1) 2	(1) 2	(10) 20
Uses of motion	(2) 4	(2) 4	(3) 6	(1) 2	(1) 2	(1) 2	(10) 20
Disadvantages of friction in motion	(2) 4	(2) 4	(3) 6	(1) 2	(1) 2	(1) 2	(10) 20
Total	(10) 20	(10) 20	(15) 30	(15) 10	(5) 10	(15) 10	(50) 100

Table one shows the relevant weights of relative emphasis in terms of hours (in brackets spent on teaching each unit of the content of area (in physics) at each level of cognitive behavioural objectives. Definition of motion was taught in 5 hours and 10 items covered this unit. The proportion assignment of weights and items is done by multiplying every column total by each row total and dividing the product by the overall total. For example, in definition of motions the teacher's time for the first unit is obtained by:

10*5/50 = 1

The item for the same unit is

20*10/100. This is how all units and hours were assigned.

Item Difficulty Analysis:

After you created your objectives assignment of items theme is needed for you to know the items are appropriate. You need to effectively differentiate between students who do well on the overall test and those who did not, this assertion can only be made through determining the difficulty level of test items, (Difficulty index) this measurement allows teachers to calculate the proportion of students who answered the test items accurately. e.g. if you give a multiple choice test four options (A,B,C,D,). To 35 students as follows: the item difficulty analyses can be done

Question	A	B	C	D
1	0	1	29*	5
2	23*	2	7	3

***correct answer**

From question I, we can see that A was not a very good distractor- no one selected that answer. We shall compute the difficulty of the item by dividing the number of students who chose the correct answer (29) by the number of total students (35). Therefore, the difficulty index of question 1 using this formula.

D = U + L / N is given as 29/35 .83, where U- is the best able, while L- is the least able students. N- is the number of students both in the upper and lower group. Onunkwo (2002) recommended an index of 0.50. A rough "rule-of- the thumb" is that if the item difficulty is more than .71, it is an easy item; if the difficulty is below .25, it is a difficulty item. Given this parameter, this item could be regarded moderately easy as (83%) of the students got it correct, the same method is applicable for question 2.

Discrimination Index

It refers to how well an assessment differentiates between high and. Low scorers. Therefore, high performing students should be able to select the correct answer for each question more than the low performing students. Positive discrimination index indicates that students who had a high total score chose the correct answer for a specific item more often than the students who had a lower overall score. Note, the reverse is the case for a negative discrimination index.

Students	**Total scores (%)**	**Questions**		
		1	2	3
A	91	1	0	1
B	91	1	0	1
C	80	0	0	1
D	70	1	0	1
E	60	1	0	1
F	50	1	0	0
G	50	1	1	0
H	40	1	1	0
I	45	1	1	0
J	45	0	1	0

***1 indicates correct answer, 0* indicates incorrect answer**

1) Arrange the students with highest overall scores at the mastery to the mastery to the non-mastery. Count the number of students in the mastery and non-mastery group who got each item correct. For question I, there were four students in the top half who got the item collect and four at the bottom half who got the item wrongly.

2) In the item get the difficulty index. The index for questions I is gotten thus 8/10 = 80

3) Discrimination index is gotten by subtracting the number of students in the non-mastery group who got the item comet from the number of students in the mastery group who get the item correctly and dividing by the half of the total students. The formula is given thus:

DR = (U-L)/1/2'N. Four students got the item correct in the master.

Group while four students got it wrong in the non-master group. Therefore, the discrimination index is given thus:

1) Distracter index (DI)

The distracter index is used calculate the effectiveness of each distracter in the test item.

Question	A	B	C	D	N
1	0	3	29*	4	34
2	23*	2	6	3	34

In question 1, option B, 2 students in the lower group answered this option while 1 student answered this option. The distracter index is given thus:

DI = (L-U) ½ N therefore, 2-1/17 = .05

For distracter A the index is ineffective because it distracted the non-mastery and mastery students the same way. In the option B, the index is .05 shows that the distracter is effective since it is chosen by more of the non-mastery group than the mastery group.

Setting Cut-Off Point

In order to decide whether a student has mastered an objective an objectives or not, there is bound to be basis for the judgment. This can be done using the following methods: as identified by Orluwene (2012) to include:

1. **Professional Judgment:**
   This is widely used by the content teachers in the school setting who are familiar with the objectives to be measured and is the best position to draw a performance level that will be acceptable in classifying students into mastery and non-mastery. This method is closely for each objective.
2. **Needlessly Method:**
   This method is used by panel of qualified experts in the content area. They are instructed to examine every item on the test and eliminate every option on the item which the minimally competent student will eliminate thereafter calculate the probabilities across the items. Example in a 4 option multiple choice items. The probability of guessing correctly on an item is ¼. But if the raters eliminate 2 options, then the probability of guessing is ½. This task is to be done independently by the number of raters and the various probabilities gotten is summed and average is used as the cutoff point.
3. **Angoff Method**
   Here the experts examine every item on the test and estimate the percentage of those in a group of minimally competence who would answer the item correctly. Sum the percentages across the items to establish minimally acceptable scores for all the experts. The averages of their scores give cut off point.

Conclusion

In the criterion-reference test construction, it is worth knowing that the validity, reliability and estimation of cutoff point are vital and paramount parameters which will qualify a test to adequately measure it is meant to measure

References

Jaap, S., Cees, G., & Sally, MT. (2003). Educational evaluation, assessment and monitoring. Swets & Zeitlinger Publishers.

Kpolovie, P.J. (2002). Test, measurement and evaluation in education. Emhai Printing and Publishing Co.

Iweka, F. (2014). Comprehensive guide to test administration. CHIFAS, Rivers State, Nigeria.

Onunkwo, G.I.N. (2002). Fundamentals of educational measurement and evaluation Cape Publishers International, Owerri, Onitsha.

Orluwene, G.W. (2012). Introduction to test theory and development process. Chris-Ron Integrated Services. Port Harcourt, Nigeria.

Ukwuije, R.P.I. (2009). Test and measurement for teachers. Chadic Printing Press, Port Harcourt, Nigeria.