IMPROVING VALIDITY OF TEST ITEMS THROUGH CREDIBLE AND ROBUST TRIAL TESTING EXERCISE :NECO APPROACH

Being a paper presented at the Annual Conference of International Association of Educational Assessment held between October 11–15, 2015 in University of Kansas Lawrence, Kansas, USA

By

MOSES OLADIPUPO

Bayelsa State coordinator

National Examinations Council, Nigeria

oladipupodele@yahoo.com

Abstract

The underlying concept of whether an item, is testing what is supposed to test and does the item satisfy the constructor yearning , underscores the treatise of validity of test item, individual and examining body tend to adopt the most convenient and universally accepted method in ensuring validity of test items ,National Examinations Council, Nigeria over the years conduct trial testing of her test items which a view of establishing the difficulty index, discriminating index, reliability and validity of test items.

this exercise integrate psychometric ingredients of test items, present a standardized item ,structured in multiple choice form to ensure robust content coverage, the security network through which this sensitive materials passes is more tight than the normal examinations ,of course the supervision of the exercise is solely the responsibility of the examining body to the extend that the selected schools are not permitted to have access to the items before ,during and after the exercise.

Since the guiding principles of improving validity such as (a)reliable measurement of each facet through the use of multiple ,alternate form item, (b)accurate articulation of facets within and between content domain, (c) examination of incremental validity ,(d)empirical examination of whether there is a broad construct or combination of separate construct,(e)use of items that represent single facets rather than combination of facets the trial testing exercise conducted by the national examinations council satisfies these traits.

This paper diligently looked at the procedure, process, pattern, personality, prospect and problems of trial testing in Nigeria schools with the view of establishing the relevance of trial testing for the improvement of validity and reliability of a test item in assessment industry.

INTRODUCTION

National Examinations Council was established nearly two decades ago, to conduct examinations for Nigeria Secondary Schools at both school-based level and private candidates with the view of assessing candidates from thirty six states and Federal Capital Territory Abuja in order to determine their eligibility for higher education and job placement. It behooves the Organization therefore to evolve items that are not only credible but standard, and in conformity with universally accepted norms having in mind that the products of this assessment will travel outside the shores of Nigeria.
In a bid to actualize this, the Council recruited Item Writers to generate items for use for each subject, these items are constantly subjected to editing before usage, to further validate this, items are placed in the public domain in term of trial testing, these trial tested items are therefore stored in the question bank where the Council falls back to yearly. this exercise has not only afforded the Council the privilege of determining the difficulty index of these items but help in knowing the challenges of conducting examinations in different terrains which serve as input during the planning system

SURVEY

In the recent survey l carried out among selected principals of secondary schools in Bayelsa State of Nigeria on their assessment of National Examinations Council conducted trial testing shows that 98% adjudged Council's items to be standard and tough, the academic performance of these schools does not reflect the toughness of the items as these schools score between 50 to 100 percent in the last three years of Council conducted examinations this might be partly a product of exam fraud.(no correlation between the Council's difficulty index and the performance of schools)

DEFINITION OF TERMS

- **SENSITIVE MATERIALS:-** These are questions that have not been made public which must be removed from printing points en route the Council ' state office and eventually to the intended examinees.
- **NECO:- National** Examinations Council. Nigeria

- **TRIAL TESTING:-** Placing a generated standard item before intending examinees with the view of determining the psychometric ingredients(discriminating index and difficult value) of such items.
-
- **EXTERNAL CANDIDATES:-**These are candidates who had passed out of secondary school but could not make a requisite minimum grades to enable them secure admission to higher institution who now decide to remedy the deficiencies by re-enrolling for school-based exam.

PROCEDURE FOR TRIAL TESTING IN NECO

- Items are generated for each subject in multiple choice type

- Generated items are professionally edited and moderated by seasoned Evaluators before typesetting.
- There are more than two sets of items for each subject from subject officer.
- The selection of the item to be used is at the high echelon of the Council. this reduces the number of staff that will be privy to the selected items.
- Selected items are made photo ready for printers.
- Printing of sensitive materials is given to a credible and reliable printer (printing outfit).
- Monitoring the routine and labeling of the sensitive materials is exclusive duty of reliable and honesty very senior staff.
- Movement of the materials is always done based on the selected states and centers
- Nigeria is divided into six (6) zones from which schools are selected based on an established criteria that take care of public, private, rural, urban and allied schools.
- There is always a pre-information of schools nearly twelve weeks to the trial testing exercise.
- The materials from the printing outfit reside in each state office of the council under a heavy security.
- The sensitive materials released to the assigned senior staff for each school to administer.
- The administration of trial testing exercise is mainly for council's staff as schools staffs are exempted from it.
- The question papers are always collected back after each subject including the answer scripts (OMR).
- Student's eligibility is restricted to only the final year students who are due to take their final exam in the school.
- Trial testing question is presently a multiple choice format with four scrambled types ranging from A to D.
- The whole exercise last for fourteen (14) days of rigorous assessment exercise.
- The answer booklets are packaged and sent directly to the ICT office for scanning.
- The ICT section must have programme her system to accommodate the keys to each test item.

PRODUCT OF TRIAL TESTING FROM THE COUNCIL

- The results from the scanned answer booklets leave the council with further determination of psychometric indexes-i.e. one of such exercise for a set of 50 mathematics items conducted on 2000 students of senior secondary schools students in 2002 yielded a final version with 43 items. The 43 so retained after analysis yielded an average difficulty (P) value of 0.58 and mean discrimination (D) index of 0.92.These are good test items, in terms of Plausibility of the distracters for the test result examined no single option for all the retained 43 items was selected by less than 5% of the candidates while the key to each of the items in the test were balanced (ten for each option prior to the trial-testing) among the five options (A-E) provided. After the trial testing, option A was made the key 7 times (16.28%), (option B,C,D and E) were each made key 9 times (20.93%).This corroborated the assertion of Ellsworth et al (1990) that balancing the key among the alternatives provided in multiple choice test is a desirable quality. The final version of such are kept in items banks and on the highly restricted computer files in the council.

CHALLENGES FACING TRIAL TESTING IN NIGERIA

- Huge financial demand in hiring permanent staff and printing of materials couple with paucity of fund allocated to public examination body
- Challenge of securing consent of the school(s) that have been selected.
- Challenge of negative response from Items Writers especially Ad hoc.
- Challenge of defective test items.
- Laborious task involved in item analysis.
- Challenge of pooling and sampling of items from the item banks.
- School desperation in retaining a copy of question paper for each subject.
- The challenge of external candidates.
- Bad state of educational system in Nigeria where most public schools are in a dilapidated state, inadequate teaching staff, insufficient laboratory equipment or outright absent these facilities and the evil effect of examination malpractice which snowball into poor performance of the examinees.
- Most schools don't cover the syllabus thereby putting the students (examinees) at the risk of inadequate content coverage.
- Where council staff compromise the ethics of examination faulty representation of candidate intelligent capability will be recorded.

LIMITATIONS

- Trial testing is done on a sample of students, if sample selection is faulty then the whole exercise follows the same direction.
- Trial testing that is multiple choices should be interpreted objectively bearing in mind a good guesser.
- selection of the schools done subjectively may lead to faulty outcome

BENEFITS

- It prepares the school as well as the students ahead of the main exam.
- It helps in eliminating or re-structuring of ambiguous question.
- It assists examining body in knowing areas of lapses in the conduct of exam.
- The administration of examination becomes easier

V **v**aluable, accessible,
A **a**cceptable, attainable
L **l**iberal accommodate all facets
I Incisive Point blank, Straight to the point
D **d**etailed, comprehensive, unambiguous
I Interpretable, informed
T Thorough, Tactful
Y Yield the desire result

PERSONALITY

The personality involved in trial testing are mainly Evaluators, who know the rudiment of test and measurement couple with this, he/she must be a very senior member of staff of the Council who must have stay for more than half a decade.

Apart from trial testing exercise standards of test items are maintained by ensuring:

- Assurance of content validity through the usage of test blue print which is always emphasis during accreditation.
- Generating items exclusively based on secondary school curriculum for each subject.
- Assurance of good and acceptable psychometric properties of the items to be administered.

CALCULATION OF ITEM DIFFICULTY INDEX "P VALUE"

This statistics serve as an indicator for detecting items that could be removed from delivery. They set threshold for items that are too easy and too difficult.

P. stands for proportion of participants who got the item correct; i.e. if 100 participants answered the item and 70 of them answered the item correctly. Then, the P value is 0.70; the P value takes value from 0.00 and 1.00. Higher value represent easier item and lower values represent harder item m

Table 2.Guidelines for improving and Reporting the Psychometric Soundness of Instruments			
Psychometric concept and definition	**Statistical test**	**Accepted standard**	**Common errors**
Validity			
Construct(overall)			
Extent to which an instrument measures the construct under study	Exploratory Factor Analysis(EFA);	Eigen values>1.0;	Finding same number of factors as items on the tool;
	Confirmatory Factor Analysis(CFA);	Generally, factor loadings >.40;	Too few participants for number of variables
	Principal Components Analysis	Approximately five subjects per variable or number of subjects exceeds number of variables by 50	
Translational 1.Face: the instrument, on the face of it, appears to measure the construct 2.Content: Extent to which items in the tool sample the complete range of the attribute under study	None, tool "looks" like valid measure of the construct Content Validity Ratio or Content validity Index:$CVR=\frac{ne-N/2}{N/2}$	None: not considered a "true" measure of validity. Tool is accepted at face value Depends on the number of expert reviewers.	Subject measure; no criteria for acceptance Tools with low CVR or CVI are called content valid; Inflated CVI is possible with high levels of agreement
Criterion Relationship linking the attributes in a tool with the performance on a criterion	Pearson Product Moment Correlations $r=\frac{\sum ZxZy}{N}$ must have continous data	Substantial and high:r≥.45 is recommended by many authors	Tools with low CVR or CVI are called content valid; Inflated CVI is possible with high levels of agreement
1.Concurrent Scores on the measurement tool are co-related to a related criterion at the same time	High Pearson Product Moment Correlations	Substantial and high:r≥.45	Tools with low Correlations are labeled criterion valid
2.Predictive The degree to which test scores predict performance on some future criterion	High Pearson Product Moment Correlations	Substantial and high:r≥.45	Tools with low Correlations are labeled criterion valid
3.Convergent	High Pearson Product Moment Correlations	Substantial and high:r≥.45	Tools with low Correlations are labeled criterion valid

Extent to which different measures of the same construct co-relate with one other			
4.Discriminant Extent to which measures of different constructs correlate with one other	Low Pearson Product Moment Correlations	≤.45	Tools with high Correlations are labeled criterion valid
Reliability 1.Internal Consistency Extent to which performance on one item in an instrument is a good indicator of performance on any other item in the same instrument	Coefficient alpha (Cronbach's alpha)	≥..90 for clinical tools; ≥.70 for research tools; Guideline based on underlying dimensions of the construct	Tools with lower Coefficient alphas are called reliable
2. Test-retest Extent to which an instrument measures stable characteristics at two separate times	Interclass Correlation Coefficients; Pearson Product Moment Correlations; *t t*est	High correlations; generally r≥.70; No staistically significant difference in scores from pre to posttest	No report of statistical test, level of significance, or confidence intervals
3 Alternative forms Extent to which different forms of an instrument yield comparable results when given to the same sample during a single administration	Pearson Product Moment Correlation coefficient; Spearman Brown if test length has been changed(Both versions of the instrument must have equal means; variances, and alpha coefficients)	High correlations; generally r≥.70	No report of statistical test, level of significance, or confidence intervals

REPORT ON ITEM PERFORMANCE

- Does the item has an acceptable level of difficulty?
- Does the item has an acceptable discrimination index
- Where does the greater proportion of participants fall (higher, middle or lower limited?

INTERPRETATION OF RESULT ON TRIAL TESTING

consequences of test use and interpretation are hinged on the validity evidence which focuses on intended and unintended benefits or consequences of taking a survey, interpreting the results, and utilizing the result to make informed decisions.the result from trial testing always assist the council in making a meaningful decision on the best way to prosecute examination.

OTHER METHODS OF IN DETERMINING THE PSYCHOMETRIC SOUNDNESS OF ITEMS

Trial testing is a veritable tool in ascertaining the soundness of an item nonetheless this does not foreclose the existence of other methods hence the below mentioned forms of determining the difficulty index of an item

1. Factor analysis
2. Inter item correlation.

Types of validity

There are four types of validity evidence.

1. **Test content**

   With content validity, the reasoning is that an instrument can only be interpreted if it effectively reflects the construct of interest. There are two specific threats to content validity: Construct-irrelevant content and Construct under – representation. Construct-irrelevant content is spurious or redundant items that are unrelated to the construct of interest. These items can introduce statistical "noise" into the analysis and detract from the precision and accuracy of survey scores. Construct under-representation means that specific content areas that are relevant towards measuring for the construct are not accounted for in the survey.

2. **Internal structure**

   The internal structure of the survey instrument should reflect the underlying structure of the construct of interest. The construct specification begins the process of formulating this internal structure and the "factors" yielded from exploratory and confirmatory factor analyses should reflect the construct as it exists within the literature and natural environment.

3. **Processes used to respond to items**

   The psychological, cognitive, emotional, and affective processes that respondents use to answer questions is also pertinent when assessing validity. Test instructions, the reading level of the language used in the survey, and the survey items must be written in a fashion that respondents can understand and logically respond to the survey.

4. **Association with other scores and variables**

This form of validity is focused on assessing the associations between the survey instrument and other theoretically, conceptually, and empirically similar constructs. Significant associations between survey instruments and other existing instruments provides evidence of validity.

Forms of validity

- Construct validity is the aggregate form of validity evident related to assessment.
- Concurrent validity provides evidence of how a survey instrument predicts for another measure or construct at the same time.
- Predictive validity provides evidence of how a survey instrument predicts for outcomes and events that will occur in the future.
- Content validity relates to the representation of the content areas that exist in the body of literature.
- Known-groups validity generates evidence related to the ability of a survey instrument to differentiate between existing independent groups.
- Convergent validity shows evidence that a survey instrument positively correlates with other survey instruments measuring for theoretically or conceptually similar constructs.
- Divergent validity shows evidence that a survey instrument negatively correlates with other survey instruments that it should have a negative association with theoretically or conceptually.
- Incremental validity generates high-level multivariate evidence that a survey instrument accounts for new and unique variance above and beyond what has been accounted for up until that present time in the body of literature.
- Face validity is the weakest form of validity evidence showing that at face value, an instrument does what it is suppose to do.
- Confirmatory factor analysis (CFA) allows you to validate the internal structure of the survey instrument yielded from an EFA.

CONCLUSION

The concept of trial testing as seen in this paper provides a good avenue for any serious examining body to determine the validity and reliability of items before presenting it to the general public ,more so it enables public as well as private examination bodies to get a first hand information of the challenges that lie ahead in the conduct of examination,this germane utility of trial testing make it convenient to advocate its adoption for every result driven examination body to embrace.

REFERENCES

Holli A D, Michelle E. Block, Patricia Moyle-wright, Diane M. Ernest (2007) A psychometric toolbox for testing validity and reliability journal of nursing scholarship 39:2, 155-164

Mwachili, J.M. and Wasanga, P.M. (2004) *Establishing and Maintenance of Standards in Kenya Certificate of Secondary Education (KCSE) Examination*. Paper Presented at the 22nd Annual Conference of the Association for Educational Assessment in Africa, Gaborone, Botswana between 13-17 September, 2004.

National Examinations Council (NECO) (1999) *Facts about NECO*. Minna (Nigeria):

. **Odusola O. Dibu-Oyerinde (2005)** The challenges of standards in the test development process for the senior school certificate examination in Nigeria.

Okpala, P.N.; Onocha, C.O. and Oyedeji; O.A. (1993) *Measurement and Evaluation in Education.* Ibadan (Nigeria): Stirling – Horden Ltd.

Taylor, A.K. (2005). Violating Conventional Wisdom in multiple choice test Construction. *College StudentJournal*,March,2005.available at: http://www.findarticles.com/p/articles/mi_mOFCR/is_1_39/ai_n1362.../prin (viewed in 25/05/2005). Ultwang, A. and Jeremiah, B. (2004). *The Challenges of Standard Setting in Botswana's National School Examinations.* Paper Presented at the 22nd Annual Conference of the Association for Educational Assessment in Africa, Gaborone, Botswana between 13-17 September, 2004.