Evaluation of Mathematics Achievement Test: A Comparison Between Classical Test Theory (CTT) and Item Response Theory (IRT)

Eluwa, O. Idowu

Department of Educational Foundation, Guidance and Counselling
University of Calabar, Calabar, Cross River State, Nigeria

Akubuike N. Eluwa

Department of Rural Sociology and Extension
Michael Okpara University of Agriculture, Umudike-Umuahia, Abia State,Nigeria

Bekom K. Abang

Department of Educational Foundation, Guidance and Counselling
University of Calabar, Calabar, Cross River State, Nigeria

Abstract *Item bias is critical to the process of evaluating the quality of an educational assessment in terms of reliability and validity. This study applied the Classical Test Theory and Item Response Theory to evaluate the quality of an assessment constructed by the researchers to measure National Certificate of Education (NCE) students' achievement in Mathematics. The sample for this study consisted of the junior and senior Mathematics and English major teacher-education student from the Abia State College of Education, Arochukwu. A sample of 80 students was drawn for this study. The Mathematics Achievement Test (MAT) for College students developed by the authors was used. Data was analyzed in two dimensions. First, the psychometric properties of the instrument were analyzed using CTT and IRT and the detection of item bias was performed using the method for Differential Item Functioning (DIF). The results showed that although Classical Test Theory (CTT) and Item Response Theory (IRT) methods are different in so many ways, outcome of data analysis using the two methods in this study did not say so. Items which where found to be "bad items" in CTT came out not fitting also in the Rasch Model.*

Keywords: Evaluation, Reliability, Validity, Quality and Test.

Introduction

Mathematics achievement is the proficiency of performance in any or all mathematics skills usually designated by performance on a test. After over 20 years of educational research on the issue of mathematics achievement amongst students, deficiencies in the academic subjects of mathematics and science persist. (Thiessen & Blasius, 2008). Between 1970 and 1990, mathematics assessment benefitted from wider techniques and tasks to match broader teaching objectives. With the exception of increasing use of information and communication technology, recent changes have been in purposes rather than techniques. Emphasis on accountability has led to more central testing which is narrowing the curriculum and demanding strategies to reliably assess rich tasks.

Formative assessment has become an important element in improving teaching and raising standards, but more practical support for teachers is needed for successful implementation in mathematics classrooms. Students perform differently on measures of achievement in mathematics and science depending on the type of test, its content and the population of students being considered. Students of education have different attitudes towards mathematics and science, which may grow out of their distinct experiences. As such, studying their mathematics achievements continues to be an area of great interest to educational researchers.

In order to contribute to the growing interest in mathematics achievement amongst educational researchers, a mathematics achievement test was developed by the researchers in an effort to study the dismal performance of the teacher-education graduates in the mathematics portion of the National Certificate of Education's examination for teachers (NCE). The developed assessment is in line with the objectives of the mathematics curriculum of the colleges of education and in consonance with the mathematics ability required in the NCE programme. It is hoped that the performance of our students in this achievement test will predict their performance in the mathematics portion of their NCE examination.

Classical test theory (CTT) and item response theory (IRT) are commonly perceived as representing two very different measurement frameworks. Although CTT has been used for most of the time by the measurement community, in recent decades IRT has been gaining ground, there by becoming a favorite measurement framework. The major arguments against CTT are its rather weak theoretical assumptions which make CTT according to Hambleton & Jones (1993), easy to apply in many testing situations. In their views, the person statistic is item dependent and the item statistics such as item difficulty and item discrimination are sample dependent. On the other hand, IRT is more theory grounded and models the distribution of examinees' success at the item level. As its name implies, IRT mainly focuses on the item-level information in contrast to CTT's principal focus on test-level information. The IRT framework includes a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the practicality of different theoretical assumptions about the test items.

Measurement is central to the construction of a quality student assessment even in the case of a classroom-designed or non-standardized assessments instrument. Measuring variables are one of the necessary steps in the research process. Therefore, the main objectives of the present study was to analyze the psychometric properties of the mathematics achievement instrument developed and administered on two different groups of students in order to established the validity and reliability of the instrument using CTT and IRT framework. The study also determined the Differential Item Functioning (DIF) for each item. The test that measured achievement in college mathematics is criterion-referenced so that test scores directly conveyed level of competence in defined mathematics domain.

Theoretical Framework

Basic Tenets of IRT, CTT, and DIF

The tenets of Item Response Theory (IRT) are based on two basic assumptions. First, a more able person should have greater probability of success on assessment items than a less able person. Secondly, any person should always be more likely to do better on an easier item than on a more difficult one. IRT assumes item difficulty is the characteristics influencing a person response, and person ability is the characteristics influencing item difficulty estimates (Linacre, 1999). Thus, careful considerations should be given to the construction of assessments. Items should be written clearly and concisely such that they are not vulnerable to guessing. On the other hand, Classical Test Theory (CTT) emphasizes that item parameter should form the basis of assessing academic achievement.

CTT depends largely on the characteristics of the testes and thus item difficulty fluctuates depending on whether the population taking the test possesses certain ability.

In evaluating the quality of an assessment tool, a discussion of reliability and validity is essential. The reliability is the degree to which an instrument consistently measures the ability of an individual or group while validity is the degree to which an instrument measures what it is intended to measure. The CTT provides a very simple way of determining the validity and reliability of a test. By subjecting the whole test results to simple statistical tests, one can determine the validity and reliability of the test. In the same view, IRT offers a more complex but more reliable way of determining validity and reliability of test. Thus, if the focus of CTT is on the test as a whole, IRT focuses on each item and each individual test taker.

Furthermore, latent trait models in test construction are utilized for purposes of constructing equivalent test forms, developing tests that discriminate between ability levels, and improving customized test system. If a test item has different connotative meanings for different groups, then examinees' performance on that item may be subject to sources of variation that are unrelated to ability level. This refers to differential item function and can cause item bias (Crocker & Algina, 1986). IRT can also be used to investigate item bias. A set of items is considered unbiased if all subpopulations are equally affected by the same sources of variance, thus producing similar ICCs for both groups (Cole & Moss, 1985). In order words, set of items is considered unbiased if a source of irrelevant variance does not give an unfair advantage to one group over another (Scheuneman, 1979).

Unfortunately, the investigation of item bias is not that clear cut. IRT, as well as chi-square and item difficulty, can flag items as biased even if they are not (Park & Lautenschlager, 1990).More so, multidimensionality can be mistaken for item bias with IRT as a result of differences among ICCs. ICC differences can occur even when item bias does not exist. This distinction can indicate that items are not unidimensional.

Differential Item Functioning (DIF) detection procedures can investigate the effects achievement tests have on different sub- populations (Zwick, Thayer, & Mazzeo, 1997). Some research has evaluated DIF analysis methods that involve matching examinees' test scores from two groups and then comparing the item's performance differences for the matched members (Zwick et al., 1997; Ackerman & Evans, 1994). Such nonparametric detection methods include the Mantel-Haenzsel procedure and Shealy and Stout's simultaneous item bias (SIBTEST) procedure. These procedures, however, lack the power to detect non-uniform DIF which may be even more important when dealing with polytomous items due to the multiple ways in which item scores can interact with the total score (Spray & Miller, 1994). There is also the newer procedure of detecting item bias, the Item Response Theory Likelihood-Ratio Test for Differential Item Function (IRTLRDIF). Of all of the procedures available for DIF detection and measurement, IRT-LR procedure posits several advantages over its rivals. IRT-LR procedures involve direct tests of hypotheses about parameters of item response models, and may detect DIF that arises from differential difficulty, differential relations with the construct being measured, or even differential guessing rates (Thissen, 2001). This is the reason why the researchers used this method in the detection of item bias.

Method

Participants

A total of 80 students (34 mathematics majors, 46 English majors) completed the mathematics achievement test during the ending period of the 2nd semester, school year 2008-2009.

Measure

The mathematics achievement test, a multiple-choice assessment designed to measure college students' mathematics ability was administered. The instrument comprised of 40 multiple choice items with five answer choices. The achievement test was trial tested with two groups of junior and senior teacher-education students who were not participating in the study. Mathematics majors comprised the first group while the second group was all English majors. The items on the achievement test were categorized into five content domains: Patterns and relations, equations and distances, geometric and trigonometric, shapes, areas and volumes and combinatory and probability. For all domains, the underlying construct of teacher-education mathematics remains the same; thus, the theoretical framework of unidimensionality is upheld. The test was content validated by a mathematics professor in the college of education in the same school. Suggestions were taken and the test was revised accordingly.

Procedure

The teachers in the colleges administered the test for the senior students while teachers of the junior students were the ones who conducted the test for the junior level. The students were given the test after receiving specific instruction for the test. The test was administered simultaneously for the two groups of students. The students completed the test for two hours under the supervision of their teachers. The purpose of the teacher-proctors monitoring of the test was to minimize measurement errors that could arise during the actual test.

Data Analysis

Two sections of analysis were done to establish psychometric properties. First is using the classical test theory steps which include the item analysis. Microsoft Excel was used for the analyses and computations involved in the CTT analysis. SPSS software was also used to determine reliability of the test. Secondly, item response theory method was employed to calibrate for item and person difficulties. WINSTEPS' Bigsteps software was used for this analysis. The third and last part is the presentation of the Differential Item Functioning as a result of the IRTLRDIF analysis. To detect for item biased with regards to different groups of students, DIF test was conducted using software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio Tests. The software was downloaded from the website of L.L. Thurstone Psychometric Laboratory based at University of South Carolina, Chapel Hill.

Results

This section is divided into three parts. First is the presentation of the psychometric properties of the mathematics achievement test. The validity and reliability analyses presented here were done following both Classical Test Theory (CTT) and Item Response Theory (IRT). The statistical package for Social Sciences (SPSS 15) was used to perform the analyses according to CTT. Secondly, the presentation of IRT analyses, where the software WINSTEPS' BIGSTEPS was utilized to estimate students' abilities and item difficulty for the test as well as the goodness of fit of the items. What follows are the statistical tools to analyze the data. Thus, the interpretation of data analyses can only be as good as the quality of measures (Bond & Fox, 2001). Although many testing and measurement textbooks present classical test theory as the only way to determine the quality of an assessment, the IRT offers a sound alternative to the classical approach. Because CTT is rooted in a process of dependability rather than measurement, it does not rely on item difficulty variable for precision and calibration or on total score for indicating the measured ability (Sirotnic, 1987). Thus, the weaknesses of CTT have caused IRT to gain the attention of researchers since it makes allowances where CTT does not (De Ayala, 1993; Welch & Hoover, 1993).

Reliability

The internal consistency of the test was found to be high with a Cronbach's alpha value of .77. This value indicates a good reliability for the achievement test. Aside from internal consistency, Split-half method was also performed resulting to a Guttman coefficient of .72, a value that indicates internal consistencies of the responses in the test. Finally, Kuder-Richardson, KR20 was also used to determine internal consistency with a value of .90.

Item Difficulty and Discrimination

Each item's difficulty and discrimination index were determined using the classical test theory. It shows that

27 (73%) of the items are average items. The remaining 27% belong to difficult and easy items. It could be implied from this result that the achievement test was fairly difficult because more than half of the students got the most of the items correctly. But, considering that the examinees were mathematics and English majors, the result could also mean that they really have the ability to answer even difficult items. English and Mathematics majors have rigid qualifying test to proceed with their field of specialization. Thus, to be able to major in Mathematics or English, the students must have attained an above average score in the university entrance examination test. Of the 37 items considered in the test, only 3 or 8% come up to be poor items. These items were rejected. Only two items (marginal items) need to be improved. Thirty or 81% of the items were either good or very good items. This only means that generally the items for the achievement test truly represent the learning ability of the test takers because most of the items can discriminate well between the high and low performing groups.

One Parameter-Rasch Model

The Rasch model was applied to the responses of 80 students to the achievement test in its original form of forty multiple-choice items. First, the item and person separation and reliability were examined prior to any interpretations of the data. The person separation and reliability values for the pilot data were 1.84 and 0.77 respectively. This person separation indicates the number of groups the students can be separated into according to their abilities. So, in this case there are approximately two different levels of performance in the sample. Likewise, the item separation and reliability for the trial data was 4.4 and 0.95, respectively as shown in Table 1. Considering the moderate sample size, person and item reliabilities are acceptable for the analysis to continue.

Table 1. Summary of Measure Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	18.6	37.0	.01	.40	1.00	.0	.97	-.2
S.D.	5.4	.0	.83	.02	.20	1.1	.33	1.0
MAX.	30.0	37.0	1.88	.47	1.49	2.5	2.11	2.5
MIN.	7.0	37.0	-1.88	.38	.59	-2.5	.41	-2.2
REAL RMSE .41 ADJ.SD .72 SEPARATION 1.75 PERSON RELIABILITY .75 MODEL RMSE .40 ADJ.SD .73 SEPARATION 1.84 PERSON RELIABILITY .77 S.E. OF PERSON MEAN .09								
SUMMARY OF 37 MEASURED (NON-EXTREME) ITEMS								
	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	40.1	80.0	.00	.28	1.00	.1	.97	-.1

S.D.	18.4	.0	1.26	.04	.07	.7	.15	.8
MAX.	72.0	80.0	2.82	.44	1.16	1.9	1.50	1.9
MIN.	6.0	80.0	-2.43	.24	.84	-1.3	.74	-1.6

REAL RMSE .28 ADJ.SD 1.23 SEPARATION 4.33 ITEM RELIABILITY .95

MODEL RMSE .28 ADJ.SD 1.23 SEPARATION 4.40 ITEM RELIABILITY .95

S.E. OF ITEM MEAN .21

WITH 3 EXTREME ITEMS = 40 ITEMS MEAN .41 S.D. 1.88

REAL RMSE .47 ADJ.SD 1.82 SEPARATION 3.82 ITEM RELIABILITY .94

MODEL RMSE .47 ADJ.SD 1.82 SEPARATION 3.84 ITEM RELIABILITY .94

MAXIMUM EXTREME SCORE: 3 ITEMS

All items fit the expectations of the Rasch model. In other words, all items had ZSTD infit and/or outfit statistics between -2 and 2 as shown in Table 2. The item map [on which stems are indicated on the left side and students are indicated by their number] was examined for gaps where a number of students were located along the continuum without items targeted at that ability level (see Figure 1 for circles indicating gaps).

Inserting items reflecting corresponding levels of difficulty provides more accurate measures of student abilities at these levels. Notice there are gaps between item 27 and item 23 with five students falling in this ability range. Similarly, 12 students fall in the gap between items 31 and 38, and so on. Addition of items at these difficulty levels will provide more precise measures for students at this ability levels.

Table 2. Item Statistics Misfits Order

Entry Number	Raw score	Count	Measure	Model error	Infit		Outfit		Ptbis Corr.	
					Mnsq	Zstd	Mnsq	Zstd		Items
7	72	80	-2.43	.38	1.10	.3	1.50	1.0	A-.01	Item 7 ; Item 7 : 7-7
5	45	80	-.29	.24	1.16	1.9	1.22	1.9	B.08	Item 5 ; Item 5 : 5-5
26	11	80	2.09	.34	1.12	.5	1.19	.5	C.08	Item 26 ; Item 26 : 26-26
10	39	80	.05	.24	1.09	1.1	1.14	1.4	D.17	Item 10 ; Item 10: 10-10
35	36	80	.23	.24	1.09	1.0	1.14	1.3	E.18	Item 35 ; Item 35 : 35-35
16	67	80	-1.84	.31	1.05	.3	1.13	.4	F.11	Item 16 ; Item 16 : 16-16
8	20	80	1.26	.28	1.10	.7	.98	-.1	G.21	Item 8 ; Item 8 : 8-8
11	39	80	.05	.24	1.06	.7	1.09	.9	H.21	Item 11; Item 11 : 11-11
25	52	80	-.71	.25	1.06	.7	1.05	.4	I.18	Item 25 ; Item 25 : 25-25
32	44	80	-.23	.24	1.06	.8	1.03	.2	J.21	Item 32 ; Item 32 : 32-32
12	31	80	.53	.25	1.06	.6	1.05	.4	K.22	Item 12 ; Item 12: 12-12
30	54	80	-.84	.25	1.02	.2	1.05	.3	L.21	Item 30; Item 30 : 30-30
38	47	80	-.41		1.04	.5	1.01	.-.1	M.23	Item 38 ; Item 38 : 38-38
20	53	80	-.77	.24	1.03	.4	1.01	-.1	N.21	Item 20; Item 20 : 20-20
36	39	80	.05	.26	1.03	.4	1.03	-.2	O.25	Item 36 ; Item 36 : 36-36
17	56	80	-.97	.25	1.01	.3	.99	-.3	P.21	Item 17 ; Item 17 : 17-17
34	35	80	.29	.24	1.01	.1	.99	-.6	Q.28	Item 34 ; Item 34 : 34-34
6	25	80	.91	.26	1.00	.1	.99	-.2	R.28	Item 6 ; Item 6 : 6-6
31	52	80	-.71	.29	1.00	.0	.97	-.3	S.26	Item 31 ; Item 31 : 31-31
23	58	80	-1.10	.26	1.00	.0	.94	-.6	r.25	Item 23 ; Item 23 : 23-23
15	17	80	1.50	.29	.99	.0	.87	-.2	q.30	Item 15 ; Item 15 : 15-15
9	71	80	-2.29	.36	.99	.0	.93	-.3	p.16	Item 9 ; Item 9 : 9-9

40	22	80	1.11	.27	.97	-.1	.94	-.7	o.30	Item 40 ; Item 40 : 40-40
29	24	80	.97	.26	.93	-.3	.90	-.3	n.34	Item 29 ; Item 29 : 29-29
14	47	80	-.41	.24	.96	-.9	.97	-.3	m.35	Item 14 : Item 14 : 14-14
33	27	80	.78	.25	.95	-.4	.95	-.7	l.33	Item 33 ; Item 33 : 33-33
22	15	80	1.68	.30	.95	-.3	.83	-.4	k.32	Item 22 ; Item 22 : 22-22
27	64	80	-1.56	.29	.95	-.3	.89	-.4	j.27	Item 27 ; Item 27 : 27-27
21	15	80	1.68	.30	.92	-.3	.85	-.6	i.33	Item 21 ; Item 21 : 21-21
2	30	80	.59	.25	.94	-.8	.94	-.5	h.37	Item 2 ; Item 2 : 2-2
3	15	80	1.68	.30	.94	-.3	.77	-.9	g.36	Item 3 ; Item 3 : 3-3
4	66	80	-1.74	.31	.93	-.3	.75	-.9	f.30	Item 4 ; Item 4 : 4-4
24	6	80	2.82	.44	.91	-.2	.80	-.4	e.26	Item 24 ; Item 24 : 24-24
19	52	80	-.71	.25	.90	-1.1	.84	-1.2	d.38	Item 19 ; Item 19 : 19-19
13	65	80	-1.65	.30	.90	-.6	.76	-1.0	c.34	Item 13 ; Item 13 : 13-13
1	52	80	-.71	.25	.90	-1.1	.81	-1.4	b.40	Item 1 ; Item 1 : 1-1
37	22	80	1.11	.27	.84	-1.3	.74	-1.6	a.49	Item 37 ; Item 37 : 37-37
MEAN	40.	80.	.00	.28	1.00	.1	.97	-.1		
S.D.	18.	0.	1.26	.04	.07	.7	.15	.8		

The item map was also used to examine whether the difficulty of items were spread across all five content domains: Patterns and relations, equations and distances, geometric and trigonometric, shapes, areas and volumes and combinatory and probability. It can be deduced from the resultant item map that the difficulty of the items are well-distributed across the domains.

Differential Item Functioning Analysis

The result of the IRTLRDIF procedure for all the achievement test items is shown in Table 2. The significant tests for items 3, 4, 7, 8 11, 36, 38 and 40 indicated DIF. English majors are more likely to respond in the lower score categories of item 3 as evidenced by the chi-square value (*χ2*= 5.5, df =3) greater than the critical value of X^2 = 3.84. Similar significant values can be observed on items 4, 7, 8, 11, 36, 38 and 40 with computed chi-square values of 6.4, 4.9, 4.6, 3.9, 8.1, 14.11, and 5.5, respectively. This result indicates that the difficulty of the items functions differentially across the two groups, and as a result, the English and Mathematics major examinees may have different probabilities of getting the same scores. On the other hand, upon close examination of the items, it could possibly mean that these particular items' concepts were not discussed in depth for the English majors. Nevertheless, all of these items with DIF are flagged for revision or rephrasing in a way that should be balanced for both groups of students.

Discussion and Conclusion

Based on the test results, the researchers revisited all items flagged for review in the IRT analysis. Item 9 on the achievement test belong to the easiest items, yet, no students were able to answer it. The item will be rejected or it will be revised thoroughly and make it the first item in an effort to place an easier item first on the student assessment. The item will be reworded because the author felt students were overanalyzing the question. The item with the negative item-total correlation (item 7) will be deleted because the item in general was confusing.

Overall results of analysis showed that the achievement test in its generality was a good test. Although there are items removed, revised, and rephrased, most of the items came out to be good items. While classical test theory (CTT) and item response theory (IRT) methods are different in a so many ways, results of the analyses using these two methods do not say so. Items which were found to be "bad items" in CTT came out be not fitting also in the Rasch Model. Items 7, 9, 16, 24 and 26 were found to be marginal if not poor items in CTT. These were also the items that turned out to have extreme logit measures qualifying it to be unfitting in the latent trait model.

Surprisingly, some of the items came out to be biased as detected in the DIF analysis. Items 3, 4, 7, 8, 11, 36, 38 and 40 will be subjected to revision to remove its bias that is in favor to Mathematics majors (Table 2). Although it could be said that Mathematics majors have the advantage in taking the test, it should not stop there. The test was made to measure the knowledge that was supposedly acquired by a student regardless of his/her field of specialization. Besides, most of the items were patterned from the Mathematics items in the General Education part of the NCE examination where there is no biasness in its items.

References

Bond, T. G., & Fox, C. M. (21). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Cole, N. S., & Moss, P. A. (1993). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (201-219). Phoenix, AZ: Oryx Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Spray, J., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items.* (RR 941). Iowa City, IA: American College Testing Program.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 535-556.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Linacre, J.M. (2002) 'What do Infit and Outfit, Mean-Square and Standardized Mean?', *Rasch Measurement Transactions, 16 (*2), 878.

Spray, J.A., & Miller, T.R. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportions correct when population ability distributions are incongruent*. Research Report 92–1. Iowa City, Iowa: ACT.

Thissen, D. (2001). *IRTLRDIF user's guide: software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning* [computer program]. L.L. Thurstone Psychometric Laboratory. University of North Carolina.

Thiessen, V. & Blasius, J. (2008). Mathematics achievement and mathematics learning strategies: Cognitive competences and construct differentiation*. International Journal of Educational Research*, 47(6), 362-371.