Using IRT in Determining Test Item Prone to Guessing

A.D.E. Obinne
Department of Educational Foundations and General Studies
University of Agriculture, Makurdi
Benue State, P.M.B 2373, Nigeria
Tel: 234-805-059-0409 Email: amatheldaya@yahoo.com

Received: November 12, 2011 Accepted: December 10, 2011 Published: February 1, 2012
doi:10.5430/wje.v2n1p91 URL: http://dx.doi.org/10.5430/wje.v2n1p91

Abstract

The 3-parameter model of Item Response Theory gives the probability of an individual (examinee) responding correctly to an item without being sure of all the facts. That is known as guessing.

Guessing could be a strategy employed by examinees to earn more marks. The way an item is constructed could expose the item to guessing by the examinee.

A study on comparison of the Psychometric properties of the Biology Examinations conducted by West African Examination Council and National Examination Council, in the year 2000, identified items that were prone to guessing.

1. Introduction

Item Response Theory (IRT) is one of the Latent Trait theories. It has three parameter models namely:

1) 1-prameter model (also known as the Rasch Model) which ascribes only the difficulty level of an item as the trait level required to correctly answer a question.

2) 2-parameter model which deals with the discrimination parameter of an item in addition to the items difficulty parameter.

3) 3-parameter model which gives the probability of an individual with ability, responding correctly to an item with a difficulty index, discrimination index and a guessing index. The model assumes that the three parameters (difficulty, discrimination and guessing) are necessary for an estimate a valid relationship between the probability of a correct response of an item and the trait level (ability) of an individual.

Guessing means giving an answer or making a judgment about something without being sure of all the facts.

Guessing is a standard test-taking strategy presented to examinees taking a multiple choice assessment. This strategy provides an opportunity to have an item counted correct even when the examinee has insufficient knowledge of the subject matter. If test scores are based simply on the number of questions answered correctly, then a random guess increases the chance of a higher score. Formula scoring introduces an adjustment, in the attempt to recapture ability estimation undiluted by chance responses. With formula scoring, guessing is only advisable when the choices are not completely random, but based instead on partial knowledge (Talento-Miller & Guo 2009).

Guessing is seen by many students and teachers as a major factor that determines the scores of an examinee in an objective test. This explains why most people refer to the multiple choice test as “multiple guess test”. Guessing is a serious problem in examinations.

According to Gao & Stokes (2008), there are three models two of which the likelihood of guessing or low effort is related to the location of the item in the test. The first of these, the IRT threshold guessing model (IRT-TG), may be appropriate for a low-stakes test in which the examinees are motivated at first by curiosity or interest in the test, and then abruptly abandon effort and begin to select responses at random after a certain point.

This model may, also, be appropriate for speeded tests. The second model, the IRT continuous guessing model (IRT-CG), assumes a gradual change of examinee motivation that does not result (necessarily) in guessing, but just low effort that results in a smaller probability of correct response. This can occur on a low-stakes test under the same conditions as discussed above for the IRT-TC model, or due to fatigue. Both of these patterns result in more distortion to items late in the test, especially low-difficulty ones. This observation can be applied in the design of a test. If a researcher wants to

investigate the degree of this item-location-related guessing behaviour in the test, he or she should put more easy test items near the end. The discrepancy of the correct response rates on these easy test items will help identify the presence and type of guessing or low motivation behaviour.

The third model is the IRT difficulty-based guessing model (IRT-DG). It assumes the guessers only answer the relatively easy test items and guess on the remaining items. It differs from the other two in that the chance of guessing on an item is not related to its location. Unlike the "switch-only once" strategy in the IRT-TG model, the guessers can switch multiple times between solution behaviour and guessing behaviour in the test. All three guessing models assume that the guessers employ a certain homogeneous guessing strategy. Guessers apply similar methods when guessing during examinations.

2. Why Do Examinees Guess?

Examinees guess because they do not have adequate knowledge or ability to provide correct answer. Mehrens & Lehmann (1984) identified two types of guessing namely:

1) **Blind guessing:** Where an examinee chooses an answer at random from among the alternatives offered.
2) **Informed guessing:** Where the examinee draws upon all his knowledge and abilities to choose the answer most likely to be correct.

The amount of blind guessing that occurs in normal circumstances is very small. Logic and evidence suggest that informed guessing is what takes place more in test situations. Students who are motivated to do their best will utilize every bit of information at their disposal to seek out the right answer, will eliminate implausible alternatives and will hardly find themselves in situations where blind guessing is all they can do.

For two major reasons, some educators who see guessing as a problem try to discourage guessing. First is the ethical/moral belief that guessing is wrong and/or sinful because it is a form of gambling. Secondly, guessing can affect the psychometric properties of the test. Guessing should be discouraged by means of instructions given on the test and by scoring the test in such a way as to penalize those who guess incorrectly by the use of formula scoring (correction for guessing).

These procedures have been sources of controversy for many years.

$$S = R = \frac{W}{A - 1}$$

Where: S = Corrected score

R = number of right answers

W = number of wrong answers

A = number of alternatives per item

The purpose of formula scoring, according to Ebel (1979), is to make the score a student could expect to get by guessing blindly on certain questions not higher than the score of a student who omits those items in preference to guessing blindly on them. Some of which are:

1) If a guessing penalty is applied, students should omit questions if they judge that their probability of success is at chance level. Students' probabilities of success are usually greater than they believe, and it is in their interest to attempt all items, regardless of their perceived likelihood of success.
2) The penalty penalizes only guessers. This is also not correct. If a person answers a question wrongly in the genuine belief that it is correct, not only does he fail to score on that question but he loses a fraction of a mark from somewhere else (a penalty for a guess he did not make).

Explanations have been offered for non-random guessing. One of such explanations as given by Lord in Warm (1978) is that item writers are very clever in writing distracters that are very attractive to a low ability examinees. Thus, when they do not know the answer, they are attracted more to distracters than to the correct answer and so gets the item wrong more often than if they guessed randomly.

An item writer has a tendency to try to hide the correct choice. In a four-choice item there are only 2 places to hide it – choice B, or choice C. Therefore, he writes many more items, keyed B or C than A or D, and in fact there seems to be a much stronger tendency towards C (I have verified this tendency with many item writers). This also seems to be true for 5 – choice items.

Most testing organizations have a requirement that there should be about equal number of items with the keyed choice in

ISSN 1925-0746 E-ISSN 1925-0754

each of the 4 or 5 possible positions.

We now know that test items prone to guessing can be determined and correction for guessing is not the best. The way a test is constructed can expose it to guessing. In a study in 2008, those test items that were prone to guessing among biology test items constructed by West African Examination Council (WAEC) and National Examination Council (NECO) of Nigeria were determined (Obinne, 2008).

This study specifically determined the level of guessing parameter of the test items in Biology examinations conducted by WAEC and NECO examination bodies of Nigeria.

One thousand eight hundred (1,800) students formed the sample for this study. The multi-stage stratified sampling technique was used for the study. The maximum likelihood technique of the BILOG MG computer programme was used to analyze the data.

3. Results

Table 1 shows the guessing (Asymptote) values of the test items of Biology examination conducted by NECO for the year 2000. Typically, guessing values (C-values) range from 0.00 to 0.40.

The results show that the C-values of test items for the year 2000 ranged from 0.01 of item 7 to 0.41 of item 48. most of the item C-value were between 0.10 (58%) to 0.20, (22%) a few (10%) were within 0.00, very few (3%) within 0.30 while only one item (25) had C-value of 0.41.

<Table 1 about here>

The guessing (Asymptote) values of Biology test items for 2000 WAEC ranged from 0.09 of item 59 to 0.50 of item 29. Within this range, most items had their C-values between 0.20 and 0.30.

<Table 2 about here>

4. Discussion

Most C-values of items range from 0.00 to 0.40. Items with C-values of .30 or greater are not very good items. It is desirable to have the C-value at .20 or less (Warm, 1978). For this study, the results showed the C-values of items for the year 2000 ranging from 0.01 to 0.41. The items C-values were within the recommended C-value range with about 53 (90%) out of the 60 test items having C-values of .20 and less which is the most desirable C-value. However, three items (items 27, 32 and 48) had C-values of .3 and above indicating that those items were not very good items as they are prone to guessing.

Item of 2000 WAEC test had C-values ranging from 0.09 to 0.50 with only 40 items 68% having the desirable C-value of .20 and less. This showed that most (68%) of the items in the test were good items. However, a few of them were not good enough like items 3,4,29 and others.

Most of the items had C-values of .2 and less which indicated that the items were good, with low probability of getting the answers correct by the mere guessing of the low ability examinees. This result agrees with the findings of Akindele (2003).

When low ability examinees do not know the answer, they are attracted more to destructors than the correct answer and so get the item wrong more often than if they guessed randomly.

5. Implication

This study has implication on item writing and construction. The way an item is written can influence guessing on the item. Item writers should be conscious of guessing and not write item that could be prone to guessing. Upon construction, IRT method of item analysis should be employed to eliminate those items prone to guessing, so that when guessing occurs it will not be blamed on the item.

6. Conclusion

The conclusion is that more of the Biology test items constructed by the WAEC were found to be prone to guessing than those items of NECO. Therefore, it is recommended that WAEC engages the services of experts in Educational Measurement and Evaluation to review examination items before publishing them.

References

Akindele, B.P. (2003). The development of an item bank for *selection tests into Nigerian Universities:* An exploratory study (Unpublished doctoral thesis) University of Ibadan, Ibadan, Nigeria.

Ebel, R.L. (1979). *Essential of Educational Measurement*. (3rd ed.) Englewood Cliffs New Jersey, Prentice-Hall Inc.

Gao, J. & Stokes, S.L. (2008). Bayessian IRT guessing models for partial guessing behaviours *Psychometrika* , 73(2).

Mehrens, W.A & Lehmann, I.J (1984). Measurement and evaluation in education and psychology (3rd ed.) New York : Holt, Rinehart and Winston.

Obinne, A.D.E (2008). Comparison of the psychometric Properties of senior Certificate Biology examinations conducted by West African Examinations Council and National Examinations Council (Unpublished doctoral Thesis) University of Nigeria, Nsukka, Nigeria.

Talento-Miller, E. & Guo, F. (2009). Guess what? Score differences with Rapid Replies Versus Omissions on a Computerized Adaptive Test. GMAC Research Report RR 409 – 04 Retrieved from www.gmac.comwww.mba.com.

Warm, T.A. (1978). A primer of Item response theory. Technical report with no 941078. Oklahoma City: USA Coast Guard Institute.

Table 1. Guessing Parameters of NECO Items for the year on 2000 Based on the Three -Parameter Model of IRT

ITEM	ASYMPTOTE	ITEM	ASYMPTOTE	ITEM	ASYMPTOTE
	2000	21	0.15	41	0.17
1	0.20	22	0.16	42	0.20
2	0.19	23	0.22	43	0.26
3	0.08	24	-	44	0.29
4	0.06	25	0.22	45	0.21
5	0.17	26	0.14	46	0.20
6	0.14	27	0.34	47	0.10
7	0.01	28	0.24	48	0.41
8	0.13	29	0.18	49	0.20
9	0.16	30	0.25	50	0.11
10	0.11	31	0.11	51	-
11	0.23	32	0.36	52	-
12	0.16	33	0.24	53	0.12
13	0.13	34	0.19	54	0.09
14	-	35	0.15	55	0.20
15	0.17	36	0.13	56	0.15
16	0.11	37	0.12	57	0.06
17	0.20	38	0.15	58	0.21
18	0.19	39	0.16	59	0.18
19	0.19	40	0.14	60	0.28
20	0.18				

ISSN 1925-0746 E-ISSN 1925-0754

Table 2. Guessing Parameters of WAEC Items for the Years 2000, Based on the Three-Parameter Model of IRT

ITEM	ASYMPTOTE (Lower)	ITEM	ASYMPTOTE (Lower)
	2000		2000
1	0.27	31	0.22
2	0.31	34	0.21
3	0.23	35	0.10
4	0.27	36	0.25
5	0.34	37	0.29
6	0.47	38	0.21
7	0.32	39	0.37
8	0.30	40	0.13
9	0.36	41	0.29
10	0.48	42	0.27
11	0.34	43	0.19
12	0.31	44	0.21
13	0.18	45	0.32
14	0.28	46	0.21
15	0.24	47	0.20
16	0.14	48	-
17	-	49	0.22
18	0.18	50	0.24
19	0.22	51	0.21
20	0.10	52	0.24
21	0.31	53	0.26
22	0.31	54	0.22
23	0.30	55	0.26
24	0.30	56	0.24
25	0.44	57	0.19
26	0.28	58	0.25
27	0.50	59	0.09
28	0.29	60	0.2