Abacus (Mathematics Education Series) Vol. 44, No 1, Aug. 2019

ASSESSMENT OF WINSTEPS AND BILOG MG3 SOFTWARES IN DETECTING ITEM PARAMETERS IN PHYSICS ACHIEVEMENT TEST

By

Aliyu, R. Taiwo & Akuche, Ukamaka E.

Department of Science Education (Measurement and Evaluation Unit),Faculty of Arts and Education, Lead City University, Ibadan. aliyutaiwo2013@gmail.com,

Abstract

This study used Winsteps and Bilog-MG3 software to assess item Parameters of Physics Achievement Test (PAT). A 50 items instrument with a reliability value of 0.82 was developed by the researchers using Classical Test Theory (CTT) principle with a sample of 755 selected through multi-stage sampling technique. The infit and outfit of mean square score (MNSQ) and standardized score (ZSTD) of fitness of Winsteps were used while chi-square and probability values of Bilog-Mg3 were used to assess how well the two software perform in selecting item parameter in the PAT. Eventually, forty-three (43) items whose parameters were known scaled through the Winsteps software and were confirmed to measure the same construct (uni-dimensionality) following the scientific software international (SSI) prescription while 11 items were only recognized by Bilog-MG3 which were not statistically significant and fit into the prescribed model at $p < 0.05$. This shows that a great disparity occurs between winsteps and Bilog-MG3 software. Therefore, recommendation for the use of winsteps over Bilog-MG3 was made since item parameters show unidimensionality of the test while items of Bilog-MG3 shows variation between subpopulation of test takers.

Keywords*: Item Response Theory (IRT), Rasch model, 3-PL model, Physics Aptitude Test (PAT),*

Introduction

Winsteps is Windows-based software which assists with many applications of the Rasch model, particularly in the areas of educational testing, attitude surveys and rating scale analysis. It is designed to construct Rasch measurement from the responses of a set of persons to a set of items. Responses may be recorded as letters or integers and each recorded response may be of one or two characters. Alphanumeric characters, not designated as legitimate responses, are treated as missing data (Linacre, 2012). This causes these observations, but not the corresponding persons or items, to be omitted from the analysis. The responses to an item may be dichotomous ("right"/"wrong", "yes"/"no"), or may be on a rating scale ("good"/ "better"/"best", "disagree"/"neutral"/"agree"), or may have "partial credit" or other hierarchical structures. The items may all be grouped together as sharing the one response structure, or may be sub-groups of one or more items which share the same response structure.

WINSTEPS begins with a central estimate for each person measure, item calibration and response-structure calibration, unless pre-determined, "anchor" values are provided by the analyst. An iterative version of the PROX algorithm is used reach a rough convergence to the observed data pattern. The Joint Maximum Likelihood Estimation (JMLE) method is then iterated to obtain more exact estimates, standard errors and fit statistics. Output consists of a variety of useful plots, graphs and tables suitable for import into written reports. The statistics can also be written to data files for import into other software.(Linacre, 2012; Aliyu, 2015).

Measures are reported in Logits (log-odds units) unless user rescaled. Fit statistics are reported as mean-square residuals, which have approximate chi-square distributions. These are also reported t standardized, N(0,1).(Linacre, 2012; Aliyu, 2015)

The person and item total raw scores are used to estimate additive measures. Under Rasch model conditions, these measures are item-free (item-distribution-free) and person-free (person-distribution-free). So that the measures are statistically equivalent for the items regardless of which persons (from the same population) are analyzed, and for the items regardless of which items (from the same population) are analyzed. Analysis of the data at the response-level indicates to what extent these ideals are realized within any particular data set.

The Rasch models implemented in Winsteps include the Georg Rasch dichotomous, Andrich "rating scale", Masters "partial credit", Bradley-Terry "paired comparison", Glas "success model", Linacre "failure model" and most combinations of these models. Other models such as binomial trials and Poisson can also be analyzed by anchoring (fixing) the response structure to accord with the response model. The estimation method is JMLE, "Joint Maximum Likelihood Estimation", with initial starting values provided by PROX, "Normal Approximation Algorithm". There is more information at: www.winsteps.com.

Estimation of Item Parameters and Standard Errors in BILOG-MG 3

The estimation of item parameters in BILOG-MG 3 uses an approach efficient for short and long tests called MMLE (Bock & Aitken, 1981; Harwell & Baker, 1991; Harwell, Baker, & Zwarts, 1988; Mislevy, 1986), which was developed by Bock and Aitkin (1981) and extended by Mislevy (1986) to include prior probability distributions for both ability and item parameters. In general, BILOG-MG 3 is a program for multiple group analysis of dichotomously scored data with the 1PL, 2PL, and 3PL models. The approach used in BILOG-MG 3 for estimating item parameters and standard errors are ***prior ability distribution, Gaussian quadrature, MMLE estimation equation***.

Prior ability distribution: According to Shah,(2002) To estimate item parameters in BILOG-MG 3 an approach is invoked where examinees represent a random sample from an assumed prior population ability distribution $g(\theta|\tau)$, where τ is the vector containing the parameters, $\mu\theta$ and $\sigma\theta$, of the examinee population ability distribution. In this approach ability is removed from the estimation process and item parameters are estimated in the marginal distribution. In essence, estimation of item parameters is not dependent upon estimation of each examinee's ability estimate, but is dependent on the ability distribution specified a priori. The specification of the prior ability distribution is based on a researcher's knowledge of the distribution of ability for the test and examinees of interest. By invoking this approach an assumption is made that the prior ability distribution is the same for all examinees (Baker & Kim, 2004; du Toit, 2003). The prior ability distribution is important in the item estimation process because an incorrect specification could potentially lead to inaccurate item parameter estimates and standard errors (i.e., the true ability distribution does not match the prior ability distribution, Shah, 2002) Note that BILOG-MG 3 also provides the option of concurrently estimating the population ability distribution along with the item parameters instead of specifying a *fixed* prior ability distribution (du Toit, 2003). The basic idea behind this latter approach is that once the test has been administered observational data

is collected (i.e., examinees responses to each item that are scored 0, 1) on each examinee and based on these data the prior distribution is modified to incorporate observational data about each examinee. The modified distribution is now called the posterior distribution (Harwell et al., 1988).

Gaussian quadrature: Before going on, it is important to point out that the MMLE procedure used in IRT applications for estimating item parameters is usually presented in integral form, however, integration is difficult to evaluate by a computer (Harwell & Baker, 1991). As a result, the MMLE method used in BILOG-MG 3 for estimating item parameters makes use of numerical integration (quadrature), which is better known as Gaussian quadrature, for approximating the integral (Baker & Kim, 2004). In BILOGMG 3, a simple histogram technique is used to make Gaussian quadrature work. This is done by making the assumption that examinees are randomly sampled from some continuous ability distribution in the population. Typically, a standard normal prior ability distribution, $g(\theta|\tau)$, is assumed with q equally spaced standard-normal histograms used over the ability range -4 to +4 (Harwell & Baker, 1991). This means the continuous ability distribution can be approximated by using a discrete ability distribution consisting of q histograms over this range and can be more closely approximated by including more histograms. Each histogram will have a midpoint, which is known as a quadrature point (node), Xq ($q = 1, 2, \ldots, Q$). Each quadrature point will have an associated weight, $A(Xq)$, that reflects the height of the function (i.e., probability of occurrence), $g(\theta|\tau)$, around Xq. The quadrature weight is found by multiplying the width of each rectangular histogram by its height. That is, the probability density at Xq multiplied by ($Xq - Xq+1$) gives $A(Xq)$ (Baker & Kim, 2004).

Priors used in estimating item parameters in BILOG-MG 3. In BILOG-MG 3 a prior component is imposed on each item parameter during the estimation of item parameters. The term prior comes from Bayesian statistics, often referred to as the prior probability distribution, and provides information about a variable in the absence of data. Essentially, Bayesian statistics is based on the idea that each parameter of interest has its own distribution, whereas most typically view parameters as fixed characteristics of the population. The function of the prior distribution in Bayesian statistical inference is for a researcher to specify their assumption about the distribution of the parameter(s) of interest (Baker & Kim, 2004).

In the IRT literature, many authors have advocated that priors be used in estimating item parameters so reasonable or identifiable parameter estimates may be found (Harwell & Baker, 1991; Mislevy, 1986; Swamminathan & Gifford, 1985). As a result, prior distributions and their hyper parameters (e.g., μ and σ of the distribution) are utilized in BILOG-MG 3 in estimating item parameters along with their respective standard errors (Baker & Kim, 2004). By imposing prior distributions on the items BILOG-MG 3 is utilizing a Bayesian approach and the MMLE approach in BILOG-MG 3 is then referred to by others as the marginalized Bayesian item parameter estimation procedure (Baker & Kim, 2004; Harwell & Baker, 1991). However, it is easier to consider the marginalized Bayesian model as an extension of MMLE (Baker & Kim, 2004). To keep things simple, only the prior distributions imposed on the item parameters in BILOG-MG 3 are discussed. In BILOG-MG 3 the default prior

discrimination (*a*) distribution is believed to be lognormal over the range 0 to ∞ (Baker & Kim, 2004). As Mislevy (1986) describes, the rationale for this prior distribution is that most IRT applications have *aj* that are greater than 0, suggesting a positively skewed distribution like the lognormal distribution.

Accordingly, BILOG-MG 3 implements the transformation $\alpha j = \log aj$ to produce a normal distribution for each αj with probability density function that is proportional to with default $\mu\alpha = 0$ and $\sigma\alpha = 0.5$, which result in $\mu a = 1.13$ and $\sigma a = 0.6$ (Mislevy, 1986; du Toit, 2003). To keep in line with the marginalized Bayesian model utilized in BILOG-MG 3, this prior component is appended to the likelihood component to produce the two components of the marginalized Bayesian item parameter estimation equation (Baker & Kim, 2004). Similarly for the *bs*, a normal prior distribution can be requested with $\mu b = 0$ and $\sigma b = 2$ (Zimowski et al., 2003). This prior distribution is selected because the distribution of *bs* in IRT applications typically follow a normal distribution and vary between -4 to +4 (Harwell & Baker, 1991). For the *c*s a prior Beta distribution is assumed with parameters ALPHA = 5 and BETA = 17. These parameters are defined as ALPHA = $mp + 1$ and BETA = $mp + 1$, where *p* is the mean of the Beta distribution and *m* is an a priori weight of 20 observations of respondents who are marking randomly (Zimowski et al., 2003). The use of a Beta prior distribution for the *c* parameters pertains to interpreting *p* as the mean probability of a correct response for an examinee with low ability. In this case $p = 1/k$, where *k* is the number of response options. By default *k* is 5 in BILOG-MG 3, so $p = .2$. The central idea behind ALPHA and BETA values is to find values that give a desired *p* value (Baker & Kim, 2004; Harwell & Baker, 1991).

Since, it is generally recognized that examinations determine the extent to which educational goals have been achieved as well as the extent to which educational institutions have served the needs of community and society (Shah, 2002). Wikipedia 2017 described test or examinations as alternative terms of assessment and defined it as; test or an examination (or exam) is an assessment indeed to measure a test-takers knowledge, skill, aptitude, physical, fitness or classification in many other topics. The psychometric methods that allow the scores of test-takers attempting different sets of items to be compared directly are based either on the Classical Test Theory model using logistic regression, (Aliyu, 2018), Rasch model (Odili, Osadebe, & Aliyu, 2015) or on item response theory (IRT) models (Wagner-Menghin & Mater, 2013).

Winsteps software is used to handle Rasch Model. This is also known as one parameter model which uses only a single parameter, namely item difficulty to estimate an unobservable trait of a particular examinee. Bilog MG3 software handles the two-parameter and three-parameter models which are widely used especially in large scale assessment (Downing, 2003; Odili, Osadebe, & Aliyu, 2015). The table 1 shows the similarities and differences between the model when anchored by the two different softwares; Winsteps and Bilog MG3.

Table 1: Rasch Dichotomous Model vs. One-parameter Logistic Model (1PL 1-PL)

For most practical purposes these models are the same, despite their conceptual differences.

Aspect	Rasch Dichotomous Model	Item	Response	Theory:

		One-Parameter Logistic Model
Abbreviation	Rasch	1-PL IRT, also 1PL
For practical purposes	When each individual in the person sample is parameterized for item estimation, it is Rasch.	When the person sample is parameterized by a mean and standard deviation for item estimation, it is 1PL IRT.
Motivation	Prescriptive: Distribution-free person ability estimates and distribution-free item difficulty estimates on an additive latent variable	Descriptive: Computationally simpler approximation to the Normal Ogive Model of L.L. Thurstone, D.N. Lawley, F.M. Lord
Persons, objects, subjects, cases, etc.	Person n of ability B_n, or Person ν (Greek nu) of ability β_n in logits	Normally-distributed person sample of ability distribution θ, conceptualized as N(0,1), in probits: incidental parameters
Items, agents, prompts, probes, multiple-choice questions, etc.: structural parameters	Item i of difficulty D_i, or Item ι (Greek iota) of difficulty δ_i in logits	Itemi of difficulty b_i (the "one parameter") in probits
Nature of binary data	1 = "success" – presence of property 0 = "failure" - absence of property	1 = "success" - presence of property 0 = "failure" - absence of property
Probability of binary data	P_{ni} = probability that person n is observed to have the requisite property, "succeeds", when encountering item i	$P_i(\theta)$ = overall probability of "success" by person distribution θ on item i
Formulation: exponential form e = 2.71828	$P_{ni} = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}}$	$P_i(\theta) = \frac{e^{1.7(\theta - b_i)}}{1 + e^{1.7(\theta - b_i)}}$
Formulation: logit-linear form $\log_e$ = natural logarithm	$\log_e\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i$	$\log_e\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right) = 1.7(\theta - b_i)$
Local origin of scale: zero of parameter estimates	Average item difficulty, or difficulty of specified item. (Criterion-referenced)	Average person ability. (Norm-referenced)
Item discrimination	Item characteristic curves (ICCs) modeled to be parallel with a slope of 1 (the natural logistic ogive)	ICCs modeled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive)
Missing data allowed	Yes, depending on estimation method	Yes, depending on estimation method
Fixed (anchored) parameter values for persons and items	Yes, depending on software	Items: depending on software. Persons: only for distributional form.
Fit evaluation	Fit of the data to the model Local, one parameter at a time	Fit of the model to the data Global, accept or reject the model
Data-model mismatch	Defective data do not support parameter separability in an additive framework. Consider editing the data.	Defective model does not adequately describe the data. Consider adding discrimination (2-PL), lower asymptote (guessability, 3-PL) parameters.
Differential item functioning (DIF) detection	Yes, in secondary analysis	Yes, in secondary analysis
First conspicuous appearance	Rasch, Georg. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.	Birnbaum, Allan. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
First conspicuous advocate	Benjamin D. Wright, University of Chicago	Frederic M. Lord, Educational Testing Service
Widely-authoritative	David Andrich,Univ. of Western	Ronald Hambleton, University of

currently-active proponent	Australia, Perth, Australia	Massachusetts
Introductory textbook	Applying The Rasch Model.T.G. Bond and C.M. Fox	Fundamentals of Item Response Theory.R.K. Hambleton, H. Swaminathan, and H.J. Rogers.
Widely used software	Winsteps, RUMM, ConQuest	Logist, BILOG
Minimum reasonable sample size	30	200 (Downing 2003)

Linacre J.M. (2005). Rasch dichotomous model of Winsteps vs. One-parameter Logistic Model of Bilog MG3 (Rasch Measurement Transactions, 19:3, 1032)

Model appropriateness is determined by the type of test items and their scoring (Aliyu, 2015). The Bilog MG3 will adjust to adapt whatever type of data (includes invalid responses). The Winsteps however has tight standards in controlling the data. Unlike the Bilog MG 3 software, invalid responses such as guessing on item will not be accepted in Winsteps. It is described as unreliable person reliability. Critics of the winsteps software often regard the software as having strong assumptions that are difficult to handle and interpret.(Odili et aal, 2015) However, these are traits that make the software more appropriate in practice than the Bilog MG3.

To this end, the researchers want to assess the operation of Winsteps and Bilog MG3 in detecting item parameter in PAT. Researchers using standard errors of item parameter estimates need to know if their test statistics using item parameter SEEs calibrated from IRT computer programs (e.g., BILOG-MG 3) are accurate with that of Winsteps. Existing research indicates item parameter SEEs for the Rasch (1PL) model and 2PL model are accurate under short test lengths (e.g., 5, 10 and 20 items; Drasgow, 1986, p. 85) and small to moderate sample sizes (i.e., 100 … 2,000 examinees) when using JMLE as found in WINSTEPS. However, none of the aforementioned studies have examined the accuracy of item parameter SEEs produced in BILOG-MG 3 in comparison with Winsteps.

Research Questions

The following research questions were raised for the purpose of this study.

i. What is the difficulty index of each item in the constructed Physics Aptitude Test (PAT) items with winsteps software?
ii. What is the difficulty index of each item in the constructed Physics Aptitude Test (PAT) items with Bilog MG3 software?
iii. What is the reliability of the constructed Physics Aptitude Test (PAT) Items with Winsteps software?
iv. What is the reliability of the constructed Physics Aptitude Test (PAT) Items with Bilog MG3 software?

Research Methods & Design

This study focuses on the assessment of Winsteps and Bilog MG3 softwares in detecting item parameters of a multiple choice Physics Aptitude Test. Instrumentation research design was adopted because it aims at introducing new contents, procedures, technologies or instruments for educational practices. The target population for this study consists of all senior secondary school two students (SSII) in Oyo State. Ten (10) senior

secondary schools were sampled. The simple random sampling techniques of balloting were used for the selection of the ten (10) senior secondary schools. The sample size for the study was 755 respondents with 75 testees each from nine schools using non-proportionate stratified random sampling technique while 80 was taking from one out of the ten selected secondary schools.

Instrument of the study

The Physics Aptitude Test (PAT) developed by the researcher contained 100 items. The test content consists of three components. Test content was based on a well designed Test Blue Print convening the six levels of cognitive domain of learning. It consists of three components of aptitude test which include: Verbal Aptitude test with the highest number of thirty (30) items; Abstract Aptitude Test which contains twenty-seven (27) items and Numerical/Quantitative Aptitude Test with fourty-three (43 items). This shows how the 100 test items in the PAT were distributed among the content areas as well as the instructional objectives.

A total of 50 items that formed the PAT were drawn using the Classical Test Theory (CTT) procedure after the experimental try-out and revision of the test items. The difficulty and the discrimination indices found were used in selecting the total of fifty test items.

Reliability of the Instrument

Reliability of the PAT was established with the use of Kuder-Richardson formular 20 (KR-20). The calculated coefficient of reliability was 0.82 which indicated that the test items could be administered to the targeted audience. The research questions were analyzed using Winsteps and BILOG-MG3 statistical software to determine the: difficult level of PAT using the Rasch and 3-PL models of IRT. In WINSTEPS, the measures are determined through iterative calibration of item using the PAT. Research questions 1 & 3 were answered using winsteps while research question 2 & 4 were answered using Bilog-Mg3 software.

Analysis and Presentation of Result

The results obtained in this study are presented and discussed here. Winsteps 3.75.0 and Bilog-Mg3 were used to answer the research questions. The following are the stated research questions:

Research Question 1: What is the difficulty index of each item in the constructed Physics Aptitude Test (PAT) using the Rasch model?

Table 1: Difficulty indices of PAT using infit and outfit of MNSQ and ZSTD indices of Rasch

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	EXACT MATCH EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
42	68	755	1.28	.13	1.04	.4	1.39	2.8	.03	.17	91.2	91.1	I0042
19	94	755	.90	.11	.98	-.2	.93	-.6	.22	.19	87.8	87.7	I0019

6	96	755	.88	.11	.99	-.1	1.01	.2	.19	.19	87.5	87.4	I0006
9	114	755	.67	.10	1.03	.4	1.17	1.9	.12	.20	85.1	85.0	I0009
43	121	755	.59	.10	1.06	.9	1.19	2.2	.08	.20	83.8	84.1	I0043
37	130	755	.50	.10	1.04	.6	1.05	.6	.15	.21	82.9	83.0	I0037
35	134	755	.46	.10	1.03	.5	1.11	1.4	.14	.21	82.9	82.4	I0035
18	135	755	.45	.10	.93	-1.2	.90	-1.3	.32	.21	82.8	82.3	I0018
44	139	755	.42	.10	1.02	.3	1.04	.6	.17	.21	81.8	81.8	I0044
33	147	755	.34	.09	1.01	.2	1.01	.2	.20	.22	81.0	80.8	I0033
34	147	755	.34	.09	1.04	.8	1.05	.8	.15	.22	80.5	80.8	I0034
48	149	755	.33	.09	1.01	.2	1.05	.7	.19	.22	81.6	80.5	I0048
7	157	755	.26	.09	1.05	1.0	1.17	2.5	.10	.22	79.4	79.5	I0007
30	157	755	.26	.09	1.12	2.2	1.15	2.2	.03	.22	77.9	79.5	I0030
11	161	755	.22	.09	.99	-.3	.98	-.2	.24	.22	79.2	79.0	I0011
47	161	755	.22	.09	1.06	1.2	1.07	1.1	.13	.22	77.6	79.0	I0047
31	166	755	.18	.09	1.01	.2	1.03	.4	.21	.22	78.8	78.4	I0031
49	166	755	.18	.09	1.03	.7	1.04	.6	.17	.22	79.0	78.4	I0049
16	167	755	.17	.09	.98	-.3	1.04	.6	.23	.22	78.4	78.2	I0016
41	167	755	.17	.09	1.01	.2	1.01	.1	.21	.22	78.6	78.2	I0041
14	168	755	.16	.09	.93	-1.4	.88	-2.0	.34	.22	78.2	78.1	I0014
50	168	755	.16	.09	.93	-1.4	.88	-2.0	.34	.22	78.2	78.1	I0050
29	178	755	.09	.09	1.02	.4	1.01	.1	.20	.23	77.1	76.8	I0029
46	183	755	.05	.09	.98	-.5	1.00	.0	.25	.23	76.9	76.2	I0046
20	187	755	.02	.09	1.03	.7	1.10	1.7	.16	.23	75.6	75.7	I0020
38	187	755	.02	.09	1.00	-.1	.98	-.3	.24	.23	76.1	75.7	I0038
45	191	755	-.01	.09	1.02	.5	1.14	2.5	.18	.23	75.9	75.2	I0045
15	197	755	-.06	.09	.91	-2.3	.87	-2.6	.38	.23	75.6	74.5	I0015
26	198	755	-.07	.09	1.01	.4	1.02	.4	.21	.23	74.7	74.4	I0026
22	200	755	-.08	.09	1.05	1.2	1.03	.6	.17	.23	72.3	74.1	I0022
17	203	755	-.10	.08	.96	-1.1	.93	-1.3	.31	.23	74.7	73.7	I0017
1	211	755	-.16	.08	1.01	.2	.97	-.6	.24	.24	71.0	72.7	I0001
12	211	755	-.16	.08	.94	-1.7	.94	-1.3	.33	.24	73.9	72.7	I0012
40	211	755	-.16	.08	1.01	.4	1.04	.7	.21	.24	71.8	72.7	I0040
36	213	755	-.17	.08	1.03	.9	1.03	.7	.19	.24	72.5	72.5	I0036
21	215	755	-.19	.08	.99	-.2	.97	-.6	.26	.24	71.0	72.3	I0021
10	217	755	-.20	.08	1.03	.9	1.04	.9	.18	.24	70.7	72.0	I0010
32	218	755	-.21	.08	.95	-1.4	.92	-1.8	.33	.24	72.9	71.9	I0032
13	219	755	-.21	.08	1.00	.0	.99	-.2	.24	.24	72.0	71.8	I0013
23	224	755	-.25	.08	.98	-.5	1.02	.4	.26	.24	70.0	71.1	I0023
5	226	755	-.26	.08	.96	-1.0	.95	-1.1	.30	.24	71.5	70.9	I0005
39	248	755	-.41	.08	.99	-.2	.98	-.6	.26	.25	68.0	68.3	I0039
4	284	755	-.63	.08	1.00	-.1	.98	-.7	.26	.25	63.7	64.7	I0004
2	290	755	-.66	.08	1.01	.4	.99	-.4	.25	.25	61.0	64.1	I0002
24	303	755	-.74	.08	.96	-1.9	.94	-2.0	.32	.25	67.2	63.1	I0024
25	322	755	-.85	.08	.99	-.3	1.00	.0	.26	.26	64.3	61.8	I0025
27	323	755	-.86	.08	.99	-.6	.98	-.7	.28	.26	61.8	61.8	I0027
28	323	755	-.86	.08	.96	-2.1	.96	-1.5	.32	.26	64.2	61.8	I0028
3	341	755	-.96	.08	.92	-4.5	.91	-3.8	.39	.26	71.8	60.9	I0003
8	359	755	-1.07	.08	.96	-2.0	.97	-1.4	.31	.26	65.0	60.3	I0008
MEAN	195.9	755.0	.00	.09	1.00	-.2	1.02	.0			75.5	75.1	
S.D.	65.4	.0	.49	.01	.04	1.1	.09	1.4			6.8	7.3	

In answering the **RQ 1**, Winsteps software programme was used to calibrate the responses of the 755 testees to the 50 PAT items. The table 2 shows the difficulty indices in the fourth column, item **42** is the most difficult item in the test. The difficulty of this item is estimated to be **1.28logits** with the standard error of **0.13** while item **8** is the easiest with **-1.07logits** and standard error of **0.08**.

Research Question 2: What is the difficulty index of each item in the constructed Physics Aptitude Test (PAT) using Bilog MG3?

Table 3: Estimates of b, a and c parameter of PAT

ITEM	INTERCEPT S.E.	**SLOPE(a)** S.E	**THRESHOLD(b)** S.E.	LOADING S.E.	**ASYMPTOTE(c)** S.E.	CHISQ (PROB)	DF
ITEM0001	-1.093	0.410	2.663	0.380	0.159	118.8	6.0
	0.220*	0.098*	0.490*	0.091*	0.041*	(0.0000)	
ITEM0002	-0.977	0.444	2.200	0.406	0.254	106.7	7.0
	0.278*	0.126*	0.465*	0.115*	0.054*	(0.0000)	
ITEM0003	-0.624	1.193	0.523	0.766	0.167	144.0	5.0
	0.209*	0.213*	0.121*	0.137*	0.046*	(0.0000)	
ITEM0004	-3.972	2.418	1.643	0.924	0.339	36.3	6.0
	1.422*	0.900*	0.117*	0.344*	0.021*	(0.0000)	
ITEM0005	-4.097	2.810	1.458	0.942	0.243	23.4	6.0
	1.250*	0.941*	0.090*	0.316*	0.019*	(0.0007)	
ITEM0006	-3.175	0.927	3.426	0.680	0.126	16.0	7.0
	0.851*	0.365*	0.890*	0.268*	0.015*	(0.0255)	
ITEM0007	-2.131	0.507	4.206	0.452	0.188	5.4	7.0
	0.507*	0.165*	1.047*	0.147*	0.025*	(0.6132)	
ITEM0008	-0.715	0.820	0.872	0.634	0.267	61.2	7.0
	0.267*	0.182*	0.202*	0.141*	0.058*	(0.0000)	
ITEM0009	-3.807	1.144	3.327	0.753	0.153	16.1	7.0
	1.236*	0.501*	0.888*	0.329*	0.014*	(0.0239)	
ITEM0010	-3.303	1.913	1.726	0.886	0.247	26.1	7.0
	0.878*	0.522*	0.115*	0.242*	0.020*	(0.0005)	
ITEM0011	-2.907	1.582	1.838	0.845	0.172	26.1	7.0
	0.810*	0.509*	0.145*	0.272*	0.019*	(0.0005)	
ITEM0012	-1.755	1.243	1.412	0.779	0.172	27.2	7.0
	0.416*	0.295*	0.116*	0.185*	0.029*	(0.0003)	
ITEM0013	-1.792	0.575	3.116	0.499	0.248	36.7	7.0
	0.459*	0.201*	0.734*	0.175*	0.032*	(0.0000)	
ITEM0014	-4.861	3.638	1.336	0.964	0.148	26.5	6.0

	2.495*	2.068*	0.066*	0.548*	0.016*	(0.0002)	
ITEM0015	-3.541	2.561	1.382	0.932	0.189	18.4	6.0
	1.450*	1.175*	0.088*	0.427*	0.020*	(0.0054)	
ITEM0016	-3.394	2.303	1.474	0.917	0.156	4.3	6.0
	0.857*	0.625*	0.079*	0.249*	0.017*	(0.6411)	
ITEM0017	-1.412	0.659	2.144	0.550	0.178	42.0	7.0
	0.332*	0.191*	0.310*	0.159*	0.036*	(0.0000)	
ITEM0018	-3.206	2.180	1.470	0.909	0.108	8.7	6.0
	0.730*	0.536*	0.075*	0.223*	0.016*	(0.1928)	
ITEM0019	-2.410	0.834	2.891	0.640	0.102	31.7	7.0
	0.516*	0.294*	0.566*	0.226*	0.018*	(0.0000)	
ITEM0020	-2.660	0.845	3.149	0.645	0.235	15.2	7.0
	0.739*	0.337*	0.777*	0.257*	0.021*	(0.0341)	
ITEM0021	-1.231	0.487	2.527	0.438	0.183	93.5	7.0
	0.278*	0.140*	0.482*	0.126*	0.041*	(0.0000)	
ITEM0022	-1.682	0.380	4.422	0.355	0.224	59.5	7.0
	0.414*	0.120*	1.266*	0.113*	0.034*	(0.0000)	
ITEM0023	-1.899	1.204	1.578	0.769	0.213	14.9	7.0
	0.457*	0.296*	0.139*	0.189*	0.028*	(0.0369)	
ITEM0024	-1.898	1.512	1.255	0.834	0.300	4.4	6.0
	0.514*	0.380*	0.120*	0.210*	0.030*	(0.6265)	
ITEM0025	-0.981	0.715	1.373	0.581	0.278	36.8	7.0
	0.318*	0.190*	0.230*	0.155*	0.053*	(0.0000)	
ITEM0026	-2.913	0.920	3.165	0.677	0.254	9.2	7.0
	0.865*	0.383*	0.816*	0.281*	0.020*	(0.2393)	
ITEM0027	-0.760	0.672	1.130	0.558	0.231	64.1	6.0
	0.243*	0.151*	0.220*	0.125*	0.056*	(0.0000)	
ITEM0028	-1.417	1.155	1.227	0.756	0.307	21.6	6.0
	0.456*	0.341*	0.147*	0.224*	0.040*	(0.0014)	
ITEM0029	-2.887	0.782	3.692	0.616	0.232	32.4	7.0
	0.872*	0.300*	1.128*	0.236*	0.020*	(0.0000)	
ITEM0031	-2.845	0.992	2.868	0.704	0.208	15.7	7.0
	0.815*	0.413*	0.609*	0.293*	0.019*	(0.0280)	
ITEM0032	-3.588	2.393	1.500	0.923	0.232	6.6	6.0
	1.006*	0.713*	0.090*	0.275*	0.020*	(0.3569)	
ITEM0033	-3.055	0.965	3.167	0.694	0.189	15.5	7.0
	0.874*	0.398*	0.776*	0.286*	0.018*	(0.0301)	

ITEM0034	-3.317	1.078	3.077	0.733	0.194	35.0 7.0
	1.059*	0.473*	0.763*	0.322*	0.017*	(0.0000)
ITEM0035	-3.210	1.030	3.117	0.717	0.173	14.4 7.0
	0.931*	0.434*	0.742*	0.302*	0.017*	(0.0449)
ITEM0036	-2.014	0.584	3.452	0.504	0.254	19.6 7.0
	0.520*	0.209*	0.904*	0.180*	0.028*	(0.0065)
ITEM0037	-3.497	0.988	3.539	0.703	0.174	18.2 7.0
	1.108*	0.400*	1.034*	0.285*	0.016*	(0.0112)
ITEM0038	-1.758	0.535	3.288	0.472	0.203	42.2 7.0
	0.416*	0.181*	0.784*	0.160*	0.032*	(0.0000)
ITEM0039	-2.136	0.731	2.923	0.590	0.301	11.5 7.0
	0.593*	0.280*	0.673*	0.226*	0.027*	(0.1195)
ITEM0040	-1.390	0.642	2.166	0.540	0.188	26.4 7.0
	0.297*	0.134*	0.278*	0.113*	0.036*	(0.0004)
ITEM0041	-3.220	0.978	3.291	0.699	0.218	8.0 7.0
	0.990*	0.412*	0.893*	0.294*	0.018*	(0.3333)
ITEM0042	-6.178	1.527	4.046	0.837	0.110	14.4 7.0
	2.502*	0.659*	0.717*	0.361*	0.012*	(0.0439)
ITEM0043	-5.321	0.945	5.629	0.687	0.182	7.9 7.0
	1.577*	0.382*	1.951*	0.278*	0.014*	(0.3421)
ITEM0044	-3.135	0.904	3.469	0.671	0.182	9.4 7.0
	0.916*	0.361*	0.960*	0.268*	0.017*	(0.2249)
ITEM0045	-2.362	0.712	3.319	0.580	0.235	14.3 7.0
	0.615*	0.269*	0.842*	0.219*	0.023*	(0.0456)
ITEM0046	-5.611	3.543	1.584	0.962	0.203	15.1 6.0
	1.559*	1.035*	0.088*	0.281*	0.017*	(0.0193)
ITEM0047	-2.283	0.695	3.287	0.571	0.192	22.7 7.0
	0.536*	0.196*	0.498*	0.161*	0.022*	(0.0019)
ITEM0048	-3.587	1.281	2.801	0.788	0.193	29.1 7.0
	1.215*	0.608*	0.574*	0.374*	0.016*	(0.0001)
ITEM0049	-4.806	1.248	3.850	0.780	0.239	9.6 7.0
	1.899*	0.532*	1.118*	0.333*	0.016*	(0.2094)
ITEM0050	-4.861	3.638	1.336	0.964	0.148	26.5 6.0
	2.495*	2.068*	0.066*	0.548*	0.016*	(0.0002)

In order to answer this research question, BILOG MG-3 software programme was used to calibrate the responses of 755 testees to the 50-items of Physics Aptitude Test. The table 3, column 4 shows the item difficulty parameter estimates obtained using Bilog MG3 of

the three- parameter model (3-PL model) which ranges from **.523** to **5.629** for item **3** and **43** respectively.

Research Question 3: What is the reliability of the constructed Physics Aptitude Test (PAT) items with Winsteps software?

Table 4 - Reliability table of 50 MAT items in logit

MEAN	195.9	755.0	.00	.09	1.00	-.2	1.02	.0		75.5	75.1
S.D.	65.4	.0	.49	.01	.04	1.1	.09	1.4		6.8	7.3

SUMMARY OF 50 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MODEL MEASURE	ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
MEAN	195.9	755.0	.00	.09	1.00	-.2	1.02	.0
S.D.	65.4	.0	.49	.01	.04	1.1	.09	1.4
MAX.	359.0	755.0	1.28	.13	1.12	2.2	1.39	2.8
MIN.	68.0	755.0	-1.07	.08	.91	-4.5	.87	-3.8

REAL RMSE	.09	TRUE SD	.48	SEPARATION	5.31	Item	RELIABILITY .97
MODEL RMSE	.09	TRUE SD	.48	SEPARATION	5.35	Item	RELIABILITY .97

S.E. OF Item MEAN = .07

Reliability of the PAT items Using the Winsteps software

Reliability of item difficulty measures was **.97** with Winsteps software. This showed that the reliability for the items was very good with .97. That is, the chances that the difficulty ordering of the items is repeated if the test were given to another group is extremely high. This is because there is a wide spread of difficulty in the items as the separation index is 5.31. The separation index of **5.31** with reliability of **.97** translates to a item strata index of **2.8**. Item strata index indicates the number of distinct category levels which can be identified by the test. The minimum item strata index is 2, which means that the test is capable of distinguishing at least 2 strata group namely, highly-ability and low-ability items.

Research Question 4: What is the reliability of the constructed Physics Aptitude Test (PAT) items with Bilog MG3 software?

Summary Statistics for Score Estimates

Empirical

Reliability: 0.5933

Marginal Latent Distribution(S)

MEAN = -0.026

S.D. = 0.988

Parameter	Mean	Stn Dev
ASYMPTOTE	0.206	0.052
SLOPE	1.270	0.849
LOG(SLOPE)	0.061	0.583
THRESHOLD	2.537	1.09

Reliability of the PAT items Using the Bilog MG 3 software

Reliability of item difficulty measures was **.59** with Bilog MG3 software, however this suggested that the ordering of item difficulty was to be reconsidered with another software or comparable sample of testees. This was in support of one of the recommendations of Aliyu, 2015. The standard error of measurement (SEM) associated with the b-parameter of each of the PAT item is used to estimate its reliability. All items had SE within the range of 0.066 and 1.951 and the mean of the marginal latent distribution of PAT was -0.026 with SD of 0.988. This accounted for the moderate item reliability of 0.59. Thus, it can be concluded that the TEST according to the Bilog MG3 software was not too adequate in measuring the PAT. With larger sample sizes, separation tends to increase and error decrease. Often time the standard error of measurement (SEM) associated with b-parameter of each item is used to estimate reliability, however this was not absolutely observed in the Bilog MG 3 software.

Discussion of Findings

Item Parameters of the PAT items using the Winsteps software

The means of the infit and outfit MNSQ was **1.00** and **1.02** respectively and the means of the infit and outfit ZSTD of **-.2** and **0.0** respectively, were very close to the expected value by the model (1.00 for MNSQ and .0 for ZSTD; Linacre, 2012; Aliyu, 2015). The most difficulty item of this test is item **42** which is estimated to be **1.28logits** with standard error of **0.13** while item **8** is the easiest with **-1.07logits** with standard error of **.08**.The standard deviation of both the infit and outfit MNSQ and ZSTD (**.04 & .09 and 1.1 & 1.4**,) respectively were insignificant compared with the expected value, these difference discrepancies were not too many and showed that most data demonstrated fit from the Rasch Model expectation with the Winsteps software, the seven (7) items that were not fit showed overfit to the Rasch model expectation in the Winsteps software.

Item difficulty measures spread in approximately .00*logits* (from **-1.07logit** to **1.28logit**). The mean for item difficulty was **.00logit** (standard error = **.01logit**), while the standard deviation is **0.49**. The main difference in mean measures of the testees and the items indicated that the PAT targeted the testees well (Aliyu, 2015 ; Odili et aal, 2015). Therefore, the items distribution on the scale showed that the items were adequate in accessing important features of the constructed PAT.

Item Parameter of the PAT items using the Bilog MG3 software

The difficulty index (b) ranged from .523 to 5.629. This shows that generally the items are difficult for the respondents. By implication, **thirty-nine (39) items** were scientifically and statistically significant and do not fit into the 3-PL model of IRT and by interpretation 11 items did fit into the 3-PL model. This is in agreement with Aliyu & Akinoso (2017). All item fit/misfit were determined at 0.05 level of significance in the Bilog

MG3 software. Among the items that fit into the 3-PL model with Bilo MG3 were observed not to fit into the Rasch model in the Winsteps software.

Conclusions

From the data analysed and described in the study, the 50 Items constructed showed that only few of the items scaled through the 3-PL model using Bilog MG3 while a large number scaled through the Rasch model objectively using Winsteps software. It was noted that few of the 43 items that fit into the model through Winsteps were not recognized by the Bilog MG3 (3-PL model) since only 11 items were recognized by the model. These 11 items were not significant and fit into 3-PL model of IRT. This implies that the Winsteps software and the Bilog MG3 software have functioned differently on the constructed PAT items. This actually shows the disparity between the two models which may be as a result of sample size and the software handling them. According to Bergan (2010) and Aliyu & Akinoso (2017). "In the Rasch approach, data that do not fit the theory expressed in the mathematical model are ignored or discarded. In the scientific [IRT] approach, theory is discarded or modified if it is not supported by data. Bergan admits that "Adherence to a scientific [IRT] approach does not imply that there are no bad items. Indeed, measurement conducted in accordance with the scientific approach facilitates effective item evaluation and selection. Generally, an important aspect of the IRT approach is the selection of an IRT model to represent the data". The researcher's conclusion "is that for this assessment, the Rasch model is preferred over the 3-PL models because the model offers a significant improvement in the fit of the data to the model over the alternative models. In other words, the additional parameters estimated in the Rasch model are justified because they help provide a better fit to the data." This could be as the result of the objectivity of the Rasch in item selection of fitness using Winsteps.

The most under-fitting item is item 42 (highest difficulty value of Rasch) with an outfit mean-square **1.39.** The most over-fitting item is item 43 (with the highest 3-PL difficulty) with an outfit mean-square value of **1.19** in Winsteps software. Both items (42 & 43) are underfit in Winsteps software showing item redundancy in the test. This shows that the item does not adequately differentiate between the high and low ability examinees. The most difficult item should be able to differentiate between high and low ability examinees, with a high discrimination value whereas item 3 with difficulty index of **.523** has a higher discrimination value of **1.193** than item 43 in 3-PL. Bilog MG3 did not show the true picture of item 43 **(b= 5.629, a=.945)** and 3 **(b=.532, a= 1.193)** in the model. Therefore, the most appropriate model (i.e. the model involving the least number of estimated parameters with objectivity measure) is preferred to represent the data" and this would motivate the selection of Winsteps software (Rasch) over the Bilog MG3 (3-PL)!

Recommendations

This paper therefore recommends the use of Winsteps (Rasch model) software over Bilog MG3 (3-PL model) software for item selection since items fit show unidimensionality of the test in Rasch. Also, item measure order in Winsteps reduces any bias of any form according to literatures. It also does not discriminate between samples and also, shows high content and construct validity. Further, none of the studies reviewed have considered the effect of test length, sample size, number of quadrature points, underlying item parameter(s) distribution(s), and underlying θ distribution(s) on the accuracy of item parameter SEEs for the three IRT models found in BILOG-MG 3 in respect to 4-parameters found in Winsteps.

References

Ahmad, Z.K. & Nordin, A. (2012). Advance in Educational Measurement: A Rasch Model Analysis of Mathematics Proficiency Test. *International Journal of Social Science and Humanity, 2(3).*

Aliyu, R. T. & Akinoso, S. (2017). Development and validation of Mathematics Aptitude Test using the Rasch and 2-PL Models of IRT. *Ibadan Journal of Educational Research 16 (2), 1-15.*

Aliyu, R.T. (2015). Construct Validity of Mathematics Test Items using the Rasch Model. *An International Journal of Social Science and Humanities Research. 3(2), 22-28*

Aliyu, R. T. (2015). *Development and validation of Mathematics Achievement Test using the Rasch Model.* An Unpublished Ph.d thesis in Delta State University, Abraka.

Aliyu, R. T. & Ocheli, O.E. (2013). Development and Validation of College Mathematics with Item response Theory (IRT) Models in Attaining Quality Education in Nigeria. *A paper published in the Delta Journal of Educational Research and Development (DJERD), 12(1), 130-140*

Andrich, D. (1992). "The application of an unfolding model of the PIRT type to measurement of attitude". *Applied Psychological Measurement, 12, 33-35.*

Baghaei, P. & Amrahi, V. (2011). "Rasch Model as a construct validation tool" in *Rasch Measurement Transaction*, 22(1),*1145-1146*

Baker, F. B., & Kim, S.H. (2004). Item Response Theory*:* Parameter Estimation Techniques (2nd ed) *New York: Marcel Dekker, Inc*

Bergan J.R. (2010) Assessing the Relative Fit of Alternative Item Response Theory Models to the Data. *Tucson AZ: Assessment Technology Inc.* http://ati online.com/pdfs/researchK12/AlternativeIRTModels.pdf

Bock, R.D. & Aitken, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: An Application of an EM Alogarithm. *Psychometrika, 46, 443-459*

Bond, T. G. & Fox, C. M. (2001). Applying the Rasch model: Fundamental Measurement in Human Sciences 1st Ed. *Mahwah, NJ: Lawrence Erlbaum*

Chen, S.Y., Ankenmann, R.D. & Chang, H.H. (2000). A comparison of item selection Rules at the early stage of computerized adaptive testing. *Apply Psychological Measurement, 24, 241-255*

Crocker, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. *Orlando, FL: Holt, Rinehart and Winston Inc.*

Downing, S. M. (2013) "Item response theory: Applications of Modern Test Theory", *Medical Education, 37, 739-745.*

Du Toit, M. (Ed), (2003). *IRT from SSI, BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnnwood, IL: Scientific Software International

Embretson, S.E. & Reise, S. P.(2000). Item Response Theory for Psychologists. *Mahwah, NJ: Lawrence-Erlbaum.*

Green, K. E. & Frantom, C. G. (2002). Survey Development and Validation with the Rasch model. *A paper presented at the international conference on questionnaire, development, evaluation and testing, Charleston, SC, November 14- 17, pp 3-8*

Hambleton, R.K. & Swaminathan,H. (1985). Item Response Theory: Principles and Applications. *Boston: Kluwer.Nijhoff.*

Harwell, M.R. & Baker, F.B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15, 375-389*

Harwell, M.R., Baker, F.B. & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm : A didactic. *Journal of Educational Statistics, 13, 279-291*

Havens, A. (2002). Examinations and learning: An Activity – Theoretical Analysis of the Relationship between Assessment and Learning. *Retrieved December 03, 2010 from http://www.leeds.ac.uk/educol/documents/00002238.htm*

Linacre, J. M. (2012). A user's guide to WINSTEPS-MINISTEP: Rasch model computer programs. Chicago, IL: winsteps.com.

Linn, R. L. (2000). Assessment and Accountability. *Educational Researcher,29,(2) 4-6* Linacre, J.

M. (2012). A user's guide to Winsteps

Nenty, H. J. (2005). The application of Item Response Theory in strengthening Assessment Role in the implementation of National Education Policy.

Nitko,A. J. (1996). Educational Assessment of Students. *The wright map. 2nd. ed. Englewood Cliffs, NJ: Merrill.*

Odili, J. N., Osadebe, P. U. & Aliyu, R. T. (2015). Assessment of Stability of Item Parameter in a Mathematics Achievement Test Under The Rasch Model. *Journal of Association of Educational Researcher and Evaluators of Nigeria (ASSEREN), 1(1), 1-8*

Olaleye, O. O. & Aliyu, R. T. (2013). Development and Validation of Mathematics Achievement Test Items Using Item Response Theory (IRT) Models in Attaining Quality Education for National Development. *A paper presented and published in the Proceedings of Mathematics Association of Nigeria (MAN) at the 50th Anniversary of the Annual National conference of MAN, 82-95*

Opasina, O. C. (2009). *Development and validation of alternative to practical Physics test using item response theory model.* An unpublished Ph.D thesis, University of Ibadan.

Osadebe, P. U. (2010). Construction and validation of Test Items. *An unpublished lecture note, Delta state university.*

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. *Chicago: University of Chicago Press.*

Rehmani, A. (2003). Impact of Public Examination System on Teaching and Learning in Pakistan *Retrieved December 24, 2010 from http://www.aku.edu/AKUEB/pdfs/pubexam.pdf*

Reza, P. ,Baghaei, P & Ahmadi, H. S. (2011). Development and validation of English Language Teacher competency Test using Item Response Theory. *The international Journal of Education and psychological assessment, 8(2),54-68.*

Shah, J. H. (2002). *Validity and credibility of public examinations in Pakistan.* An unpublished Ph.D. in the Department of Education, Islamia University Bahawalpur, Pakistan.

Swamminathan, H. & Gifford, J. A. (1985). Bayesian Estimation in the two-parameter logistic model. *Psychometrika, 50, 349-364*

Thissen, D., & Orlando, M. (2001) Test Scoring. Mahwah, NJ: Lawrence ErlbaumAssociates. 3PL, Rasch, Quality-Control and Science. J.M. Linacre. *Rasch Measurement Transactions, 2014, 27(4) 1441-4*

Wang, T & Vispoel, W. (1998). Properties of Abilities Estimation Method in Computer Adaptive Testing. *Journal of Educational Measurement, 35, 109-135*

Wiberg M. (2004). Classical Test Theory Vs Item Test Theory. An Evaluation of the Theory Test in the Swedish driving-license Test. http:www.eedusci.umn.se/digital Assets159/5929-em-no-50p.d.f.cited 4/1/2016 1.43p.m.

Zimowski, M.F, Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG for windows: Multiple-group IRT analysis and test maintainance for the binary items (version 3.0) Computer software. Chicago, IL: Scientific Software International