AN INVESTIGATION OF INVARIANCE PROPERTIES OF ONE, TWO AND THREE PARAMETER LOGISTIC ITEM RESPONSE THEORY MODELS

O. A. AWOPEJU, E. R. I. AFOLABI, O. A. OPESEMOWO

Obafemi Awolowo University, NIGERIA

Abstract. The study investigated the invariance properties of one, two and three parameter logistic item response theory models. It examined the best fit among one parameter logistic (1PL), two-parameter logistic (2PL) and three-parameter logistic (3PL) IRT models for SSCE, 2008 in Mathematics. It also investigated the degree of invariance of the IRT models based item difficulty parameter estimates in SSCE in Mathematics across different samples of examinees and examined the degree of invariance of the IRT models based item discrimination estimates in SSCE in Mathematics across different samples of examinees. In order to achieve the set objectives, 6000 students (3000 males and 3000 females) were drawn from the population of 35262 who wrote the 2008 paper 1 Senior Secondary Certificate Examination (SSCE) in Mathematics organized by National Examination Council (NECO). The item difficulty and item discrimination parameter estimates from CTT and IRT were tested for invariance using BLOG MG 3 and correlation analysis was achieved using SPSS version 20. The research findings were that two parameter model IRT item difficulty and discrimination parameter estimates exhibited invariance property consistently across different samples and that 2-parameter model was suitable for all samples of examinees unlike one-parameter model and 3-parameter model.

Keywords: classical test theory, item response theory, item difficulty & invariance

Introduction

In educational measurement invariance is the bedrock of objectivity and the lack of it tends to raise a lot of questions about the scientific nature of the measurement (Adedoyin et al., 2008). A measuring theory must not be seriously affected in its measuring function by the object of measurement. In other words, our measuring theory should be independent of what it is measuring. If this is true of the theory, then it possesses the property of invariance. Measurement theory that changes in results or findings when used across different objects or group of items cannot contribute significantly to the growth of science or to the growth of objective knowledge in any area. In other words, it is not dependable. A test can be studied and the items in the test can be evaluated according to different theories. Two such theories are the Classical Test Theory (CTT) and the Item Response Theory (IRT). These two theories are based on different assumptions and use different statistical approaches. CTT is regarded as the "true score theory." The theory starts from the assumption that systematic effects between responses of examinees are due only to variation in ability of interest. The central model of the CTT is that observed test scores (X) are composed of a true score (T) and an error score (E) where the true and the error scores are independent. The variables are established by Spearman (1910) and Novick (1966) and best illustrated in the formula: X = T + E

Based on the premise that observed scores are a function of only factors – true scores and measurement error – the theoretical basis for CTT resides in the following formula: X = T + E. This equation represents the three components as discussed above, with T being the hypothetical indicator, X the observed indicator, and E the amount of disagreement between T and X. IRT is generally regarded as an improvement over CTT. For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information.

For test items that are dichotomously scored, there are three IRT models, known as three-, two- and one- parameter IRT models. A primary distinction among the models is the number of parameters used to describe items. The equation of the Item Characteristics Curve (ICC) for one parameter logistic model is given as:

$$P_i(\theta) = \frac{1}{1+e^{-\ (\theta-b_i)}} \tag{1}$$

The two-parameter model equation is:

$$P_i(\theta) = \frac{1}{1+e^{-D_{ai}(\theta-b_i)}} \tag{2}$$

The three parameter IRT model equation is:

$$P_i(\theta) = c_i + (1-c_i)\frac{1}{1+e^{-D_{ai}(\theta-b_i)}} \tag{3}$$

where: $P_i(\theta)$ is the probability of a current response for the ith item; b_i is the difficulty parameter for the ith item; a_i is the discrimination parameter for the ith item; c_i is the guessing parameter for the ith item; θ is the ability level; D represents a scaling factor.

These theories enable the studying of tests by identifying parameters of item difficulty, item discrimination and the ability of test takers. CTT and IRT analyse items qualitatively, in terms of their content and form, which includes content validity, as well as item-writing procedures and quantitatively, in terms of their statistical properties, which includes the measurement of item difficulty and discrimination. Both the validity and the reliability of any test depend ultimately on the item difficulty and discrimination.

These theories are concerned not only to determine the reliability and validity of tests but also to holistically improve the quality of test items.

Analysis based on CTT has been used over the years and is still useful nowadays in test constructions but varies from sample to sample, and, because item parameter indices are sample dependent, it lacks invariance across groups of examinees (Hambleton et al., 1991). One great advantage of IRT is the item parameter invariance. The property of invariance is the cornerstone of IRT, and it is the major distinction between IRT and CTT (Hambleton, 1994). The property of IRT item parameter estimates to remain unchanged across various groups of examinees and ability estimates to remain invariant across groups of items makes IRT applicable and useful over CTT. Group invariance of the item parameters says that the values of the item parameters are a property of the item, not of the group that responded to the item (Mallikarjuna & Natarajana, 2012).

Many researchers assume that the invariance characteristics of IRT parameter estimates make it superior to CTT in educational measurements (Ojerinde, 2013; Awopeju & Afolabi, 2016). Researches examining their properties have revealed consistent, demonstrable differences, but, the empirical studies examining the degree of invariance characteristics in the IRT models are very scarce. The aim of this study is to: (1) examine the best fit among one parameter logistic (1PL), two parameter logistic (2PL) and three parameter logistic (3PL) IRT models for SSCE, 2008 in Mathematics; (2) investigate the degree of invariance of the IRT models based item difficulty parameter estimates in SSCE in Mathematics across different samples of examinees; and (3) examine the degree of invariance of the IRT models based item discrimination estimates in SSCE in Mathematics across different samples of examinees.

Research questions

In order to carry out this study, the following research questions were raised: (1) which of the IRT models is the best fit to evaluate Senior Secondary School Certificate Examination in Mathematics; (2) what is the invariance of CTT and IRT models based item difficulty estimates across different samples of examinees; (3) how invariance is the CTT and IRT models based item discrimination estimates across different samples of examinees.

Methodology

Research design

The research design used was ex-post-facto. Ex-post-facto design is relevant to this study because it allows analysis to be performed on existing data. In this case, the responses of students to multiple-choice items in Senior School Certificate Mathematics Examination, May/June 2008, of Osun State constituted the data for the study. Also, in ex-post-facto design, manipulation becomes impossible and data collected are near perfection since they are collected in a controlled environment.

Population and sample

The population comprised all the students that sat for NECO senior school certificate Mathematics Examination Paper 1 (May/June, 2008) in Osun State. A computer-based simple random sample of responses of six thousand students (6000 students), 3000 males and 3000 females, from a total population of 35, 262 students who took the examination were selected.

Three sampling plans were employed to estimate item difficulty and item discrimination of the test scores under the CTT and IRT measurement frameworks. The sampling plans were random samples, gender group sampling and truncated group sampling. The sampling plans allow for the comparability of each framework across progressively less comparable samples.

Two different sample size conditions were employed to investigate the functionality of CTT and IRT estimates. In large scale measurement situations, one set of samples was randomly selected with n=1000. And clinical situations were often constructed with small sample sizes; a second set of sample was randomly selected, n=100. The second set of random sample was drawn to look at the effect of small sample.

One set of random samples consisting of 1000 examinees, were drawn from the 6000 examinees. The second set of random samples, consisting of 100 examinees, was also drawn from the 6000 examinees. 1000 random samples of each gender group were drawn. The same process was employed to generate the small sample replicates. 100 samples were randomly drawn from both the female and the male group. As Fan (1998) noted, because the gender samples are subpopulations of the total population, theoretically, disparity between statistics calculated from different samples will be larger than that found in random sampling plan.

A third sampling involved truncated high-ability and low ability group samples. For this sampling plan, 1000 samples were randomly drawn from both the low-ability and high-ability groups. For small samples, 100 samples were randomly drawn from both the low and high-ability groups. The low-ability sample was comprised of students whose total test score fell in the 0 to 21 mark out of 60 while the high-ability group fell in the 39 to 60 mark out of 60. One-hundred samples were randomly drawn from both the low and high ability group. These truncated high-ability and low-ability group samples should theoretically display the greatest dissimilarity between the CTT and IRT statistics, because “these two groups were defined in terms of test performance, not in terms of a demographic variable” (Fan, 1998).

Research instrument

The instrument for this study was the May/June 2008 NECO Senior School Certificate Examination Mathematics Paper 1. It was a dichotomous multiple choice examination consisting 60 items and based on the Senior Secondary School Mathematics Curriculum. The Nigeria Senior School Certificate Examination was administered at the end of the third year of senior school certificate course to measure the achievement level of candidates at that point. The examination was used as a tool to qualify students who were to proceed to the next level of education, which is tertiary institutions and also as an assessment mechanism that measures the extent to which basic competencies and skills have been acquired. The instrument was assumed to have been moderated and validated by NECO before it was administered on the students. The 60 multiple-choice Mathematics questions covered a wide range of topics in the Senior Secondary School (SSS) syllabus, showing that it had content validity. The reliability coefficients of the students' responses to the 60 multiple-choice Mathematics questions using Cronbach's Alpha coefficient was found to be 0.853, (n = 6000).

Data collection

The data used in this study were responses of candidates who wrote May/June 2008 NECO SSCE Mathematics in Osun State. These responses were on marked optical recorder mark (OMR) sheets and OMR sheets containing the responses of these candidates were collected from NECO head office, Minna. NECO is an examination body in Nigeria mandated to conduct Senior School Certificate Examinations and award certificates to candidates based on the individual candidate results. Senior School Certificate examination in May/June is typically taken by school-bound students in SSS 3. The NECO senior school certificate examination is given via easy written and pencil and paper objective tests.

Sixty multiple-choice Mathematics questions were administered on SSS 3 students in their respective schools under the supervision of the representatives of NECO appointed supervisors and school invigilators in each school. The demographic data of each of the students such as name, Centre number, candidate number and sex were printed on the OMR sheet to ensure proper coding for computer analysis.

Analysis of data: item response theory

The three known IRT models for binary response were used; one parameter (1PL), two parameter (2PL) and three parameter (3PL) logistic IRT models. Unidimensionality of the subject which is the major assumption of IRT models was investigated using SPSS version 20 through the eigenvalues in a factor analysis.

The BILOG-MG 3 was used to estimate the item parameters. Outputs phase 2 of BILOG-MG contains the IRT calibration results. The beginning of this output contains information about the execution; the maximum number of EM cycles, the convergence criterion, the assumption of a Gaussian person prior and the quadrature point and corresponding weights.

The -2 LOG likelihood values showed the expected progressively decreasing pattern of a well-behaved solution. The marginal maximum log likelihood function value (-2 LOG LIKELIHOOD) after the last cycle was used for comparing model fit. The columns labelled SLOPE and THRESHOLD contain the IRT-based item discrimination parameter estimates and item parameter (item location) estimate respectively. While the column labelled ASYMPTON contains the guessing parameter estimates.

Comparability of IRT and CTT statistics: two item statistics

The comparability of item characteristics for both methods was obtained by correlating: (a) the item difficulty, and (b) the item discrimination parameters. For each sampling plan, both the CTT- and IRT- based (one-, two- and three-parameter) item difficulty and discrimination estimates were obtained using BILOG-MG's marginal-maximum likelihood method.

The CTT-based item difficulty estimates were correlated with the 1PL, 2PL and 3PL IRT-based item difficulty parameter estimates, denoted by p in

IRT models but referred to threshold parameter in BILOG-MG. Also, the CTT-based item discrimination parameter, both the item-test point-biserial and the transformed item-test point-biserial correlation, were correlated with the 2PL and 3PL IRT-based item discrimination parameter estimates. 1PL IRT-based item discrimination parameter estimates were not available. All the correlation analysis was achieved using SPSS version 20.

Degree of invariance between CTT and IRT

The three sample techniques employed in this study generated progressively dissimilar samples across the two sample techniques. The three sampling frames used to evaluate invariance were; (a) random sample (b) gender group sampling and (c) truncated ability group samples. The item characteristics parameters from different samples, within the same sampling plan, within the same measurement framework (i.e., IRT to IRT, CTT to CTT), were correlated to evaluate the degree of invariance.

The Bias in Sample Correlation Coefficients were corrected by the use of *both* the Fisher and the Olkin and Pratt corrections.

Fig. 1 is the scree plot for the 60 multiple-choice SSC Mathematics Examination items. The factor analysis that was performed on the items using extraction method of principal component analysis showed that the first factor having the initial eigenvalue of 10.81 which clearly exceeded that of the second factor of 5.265, as also revealed in Figure one. From Figure one, the Scree plot showed a visual of the total variance associated with each factor. The steep slope showed the large factors associated with the loading greater than the eigenvalue of 1. The gradual trailing off (scree) showed the rest of the factors lower than an eigenvalue of 1. There are thirteen factors whose values are greater than eigenvalue of l and one extracted communality factor distinctly higher than others, showing that the test is unidimensional in nature.

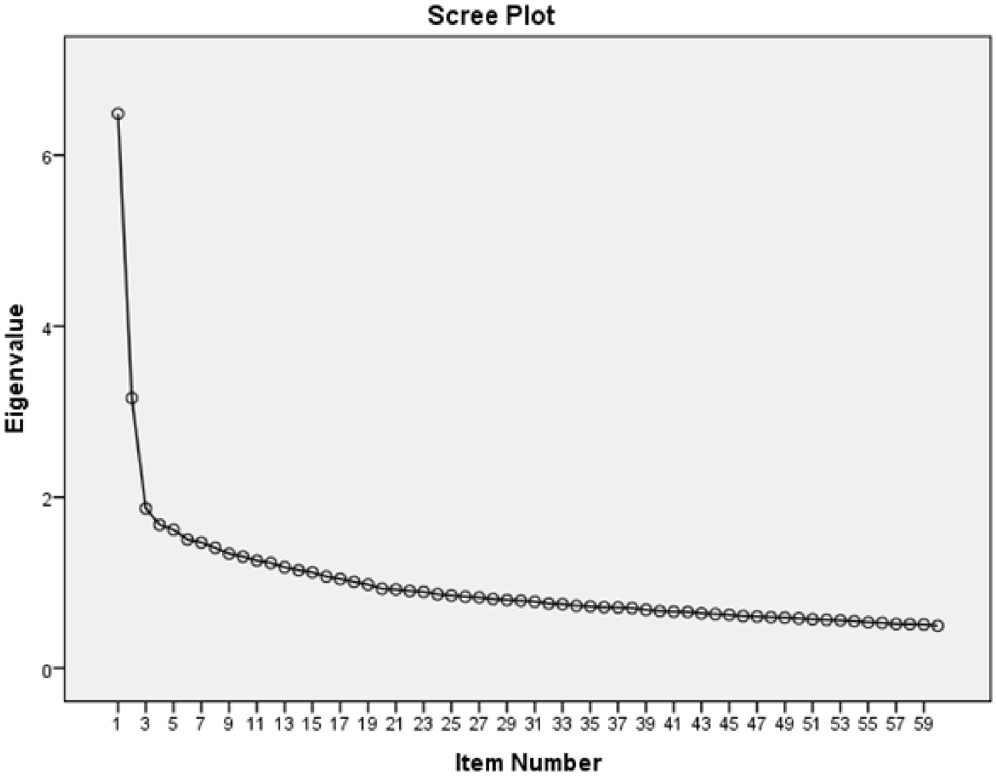

Figure 1. Scree Plot for 60 dichotomous items

Results

Research question 1: which of the IRT models is best fit to evaluate Senior Secondary School Certificate Mathematics examination?

Table 1. Model fit statistics: comparison of 1Pl, 2PL and 3PL models

Model	**-2 in L**	**Relative change**	**Number of parameters**	**AIC**	**BIC**
1PL	428400.862		60	428400.862	428922.833
2PL	420312.241	0.0189	120	420432.241	421356.183
3PL	419550.952	0.0018	180	419,730.952	421116.865

Table 1 showed that the 2PL model relative change, R^2_Δ =0.0189 (1.89%), this indicated that the 2PL model results in a 1.89% improvement fit over the 1PL model. The 3PL model relative change, R^2_Δ = 0.0018 (0.18%), this showed that 3PL model results in an improvement of fit of 0.18% over the 2PL model. The Bayesian information criterion (BIC) in the one-parameter model is high (BIC = 428,922.833) compare to other two models. In the two-parameter model, the Bayesian information criterion is lower than in the one-parameter model (BIC = 421,356.183) and the three-parameter model has the lowest Bayesian information criterion (BIC = 421,116.865).

The assessment of the IRT model fit indicates that the 2PL model comparing to the 1PL model fits the data significantly better (the difference in -2log likelihoods) and also that the 3PL model fits to the data better than the 2PL for SSCE in mathematics.

*Research question 2***:** what is the invariance of CTT-based and IRT-based item difficulty estimates when compared across different samples?

Table 2 and 3 present the results addressing the fourth research questions by analyzing the comparability of correlations between item difficulty estimates from two different sample sizes derived from the same measurement framework (i.e., CTT vs CTT, or IRT vs IRT). Table 2 presents the n=1000 data while Table 3 presents the n=100 data.

To obtain the entries in Table 2 and 3, the following two steps were taken: (a) for each of the 1000 and 100 samples, the IRT one parameter, two parameter and three parameter models estimates and CTT estimates were obtained; (b) for each sample the CTT- and IRT-based item difficulty estimates were correlated with opposing estimates within the sampling plan (e.g., males vs. females, high-ability vs low-ability). Each of the 1000 females sample was correlated with the corresponding male sample. Likewise, each of the 1000 high-ability samples was correlated with the corresponding low-ability sample.

Table 2. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT item difficulty indexes (n = 1000)

CTT MODEL			**IRT MODEL**		
Sampling Frame	**p values**	**Transformed p values**	**1PL**	**2PL**	**3PL**
Random Samples	0.992	0.992	0.991	0.923	0.954
Female - male samples	0.974	0.974	0.969	0.884	0.933
High - low ability samples	0.363	0.363	0.794	0.775	NC

Results in Table 2 showed that both the CTT p and transformed CTT p were strong invariance for the random sampling plan (r = 0.992). The IRT-based item difficulty estimates for one-parameter also indicated strong signs of invariance (r= 0.991). The two-parameter IRT-based item difficulty estimates were lower, but still strong (r = 0.923). A better strength of the correlation was found in the three-parameter IRT-based item difficulty estimates (r = 0.954).

For the gender sample plan (female-male) both the CTT p and transformed CTT p are similar and showed signs of strong invariance (r = 0.974). In the same sample plan, the IRT-based item difficulty estimates for the one-parameter model also indicated strong sign invariance (r = 0.969). The two-parameter IRT-based item difficulty estimates indicated lower but still strong invariance (r = 0.884). An improvement in the strength of the invariance was found in the three-parameter IRT-based item difficulty estimates (r = 0.993).

For the ability sample plan (high-low ability), both the CTT p and the transformed CTT p yielded results that ran contrary to the other sampling plans. They both showed signs of weak invariance (r = 0.363). The IRT-based item difficulty estimates for one-parameter model indicated strong sign of invariance although showing a decrease in invariance from the previous sampling plans. It showed a higher degree of invariance than that of the CTT-based item difficulty estimates (r = 0.794). In the same ability group, the two-parameter IRT-based item difficulty estimates were lower but still showed sign of strong invariance (r = 0.775).

Table 3. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT item difficulty indexes (n = 100)

CTT MODEL			IRT MODEL		
Sampling Frame	**p values**	**Normalised p values**	**1PL**	**2PL**	**3PL**
Random Sample	0.917	0.917	0.912	0.869	0.902
Female - male samples	0.839	0.839	0.838	0.765	0.894
High - low ability samples	0.296	0.296	0.764	0.759	NC

Table 3 (n=100) indicated that, for the random sample plan, both CTT p and normalized CTT p item difficulty estimates showed strong correlations (r = 0.917), indicating that invariance held for the CTT-based estimates. For the one-parameter IRT item difficulty estimates, the results indicated strong correlation (r = 0.912) which indicated that invariance held for IRT one-parameter model. The two-parameter model demonstrated weaker correlation but still strong (r = 0.869). And the three-parameter model showed strong correlation (r = 0.902), indicating strong invariance.

For the gender sample plan, as was shown, a continued degeneration of the correlations was found in both CTT p and normalized CTT p item difficulty (r = 0.839). For the one-parameter IRT item difficulty, the correlation was similar to what was found in CTT-based item difficulty (r = 0.838), showing sign of invariance. The two-parameter IRT-based item difficulty indicated weaker correlation (r = 0.765). And the three-parameter model showed strong correlation (r = 0.894) which demonstrated better invariance to what was found in the two-parameter IRT model.

For the ability sample plan, the result showed that both the CTT p and the normalised CTT p item difficulty indicated sign of very weak invariance (r

= 296). For the same sample plan, the IRT-based item difficulty estimates for the one-parameter model indicated good sign of invariance (r = 0.764). A further drop in the strength of invariance was found in the two-parameter IRT-based item difficulty estimates (r = 0.759)

Table 4. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT item difficulty indexes with Fisher transformed correction for bias (n = 1000)

CTT MODEL			**IRT MODEL**		
	Fisher Transformed		**Fisher Transformed**		
Sampling Frame	**p values**	**Normalized p values**	**1PL**	**2PL**	**3PL**
Random Sample	0.964	0.964	0.959	0.876	0.935
Female - male samples	0.950	0.950	0.905	0.836	0.901
High - low ability samples	0.362	0.362	0.764	0.752	NC

Table 4 showed the results of table (n=1000) except the sample correlations from table 2 have been corrected for bias using Fisher transformed correction.

The random sample plan, both transformed CTT p and normalized CTT p item difficulty indicated signs of strong invariance (r = 0.964). The results from the one-parameter IRT item difficulty estimates had strong invariance (r = 0.959), while a weaker sign of invariance was found in the two-parameter IRT item difficulty estimates (r = 0.876). However, the three-parameter IRT item difficulty demonstrated stronger invariance in the same sample plan (r = 0.935). The gender sample plan, Table 4 indicated that both the transformed CTT p and normalized CCT p had strong invariance (r = 0.950). The IRT-based item difficulty estimates for the one-parameter also indicated strong sign of invariance (r = 0.905). The two-parameter IRT-based item difficulty estimates were lower, but still strong (r = 0.896). A better strength of the correlation was found in the three-parameter IRT-based item difficulty estimates (r = 0.901), indicating a better invariance.

For the ability sampling plan, both the transformed CTT p and CTT p show signs of weak invariance (r = 0.362). The IRT-based item difficulty estimates for the one-parameter and the two-parameter item difficulty estimates demonstrated strong correlations (r = 0.764 and r = 0.752 respectively).

Table 5. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT item difficulty indexes with Fisher transformed correction for bias n = 100

CTT MODEL			**IRT MODEL**		
	Fisher Transformed		**Fisher Transformed**		
Sampling Frame	**p values**	**Normalized p values**	**1PL**	**2PL**	**3PL**
Random Sample	0.911	0.911	0.897	0.831	0.899
Female - male samples	0.889	0.889	0.881	0.775	0.851
High - low ability samples	0.253	0.253	0.760	0.748	NC

Table 5 showed the results of Table 3 (n=100) except that the sample correlations from Table 15 have been corrected for bias using Fisher transformed correction. None of the correlations found in Table 5 matched those found in Table 3.

For the random sample plan, both transformed CTT p and normalized CTT p item difficulty indicated signs of strong invariance (r = 0.911). The results from the one-parameter IRT item difficulty estimates had strong invariance (r = 0.897) while a weaker sign of invariance was found in the two-parameter IRT item difficulty estimates (r = 0.831). However, the three-parameter IRT item difficulty demonstrated stronger invariance in the same sample plan (r = 0.899).

For the gender sample plan, Table 5 indicated that both the transformed CTT p and normalized CCT p had strong invariance (r = 0.889). The IRT-based item difficulty estimates for the one-parameter indicated strong sign of invariance (r = 0.881). The two-parameter IRT-based item difficulty estimates showed lower, but still strong sign of invariance (r =0.775). An increase in the strength of the invariance was found in the three-parameter IRT-based item difficulty estimates (r = 0.851).

For the ability sampling plan, both the transformed CTT p and normalized CTT p showed identical and signs of weak invariance (r = 0.253). The IRT-based item difficulty estimates for the one-parameter and the two-parameter item difficulty estimates demonstrated high invariance (r = 0.760 and r = 0.748 respectively).

Research question 3: how invariant are the CTT-based and IRT-based item discrimination estimates when compared across different samples of examinees.

Tables 6 and 7 present the results addressing the fifth research question, “How invariant are the CTT and IRT item discrimination estimates when compared across different samples of examinees?” by analyzing the comparability of correlations between item discrimination estimates from two different samples derived from the same measurement framework. Table 6 presents the n=1000 data while Table 7 presents the n=100 data. No correlations could be produced for one-parameter model because this model assumes fixed item discrimination for all items. Therefore, the one-parameter IRT estimates are listed as N/A in the following tables.

Table 6 (n=1000) indicated that, for the random sample plan, both point-biserial CTT and normalized point-biserial CTT item discrimination indicated sign of strong invariance (r = 0.903). The one-parameter IRT item discrimination estimate was not available. A sign of invariance was found in the two-parameter IRT item discrimination estimates (r = 0.871). However, the three-parameter IRT item discrimination demonstrated stronger invariance in the same sample plan (r = 0.887).

Table 6. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT item discrimination indexes (n=1000)

CTT MODEL			**IRT MODEL**		
Sampling Frame	**Point-biserial**	**Normalized Point-biserial**	**1PL**	**2PL**	**3PL**
Random Sample	0.903	0.903	N/A	0.871	0.887
Female - male samples	0.866	0.866	N/A	0.762	0.793
High - low ability samples	0.278	0.278	N/A	0.730	NC

For the gender sampling plan, the correlation of CTT-based item discrimination estimates was identical in point bi-serial and normalized point biserial (r = 0.866), and showed strong invariance. For the same sampling plan, the IRT-based item discrimination estimates for the two-parameter model showed strong invariance (r = 0.762). However, the three-parameter model correlation demonstrated an increase in strong invariance (r = 0.793).

For the ability sample plan, the CTT-based item discrimination estimates were appreciably lower (r = 0.278) than the other sampling plan. This showed that CTT had low invariance and degrading degree of invariance across different sample of examinees. For the same sampling plan, the IRT-based item discrimination estimates for the two-parameter model (r = 0.730) indicated strong invariance of item discrimination.

Table 7. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT item discrimination indexes (n=100)

CTT MODEL			**IRT MODEL**		
Sampling Frame	**Point-biserial**	**Normalized Point-biserial**	**1PL**	**2PL**	**3PL**
Random Sample	0.766	0.766	N/A	0.760	0.781
Female - male samples	0.749	0.749	N/A	0.748	0.765
High - low ability samples	0.082	0.082	N/A	0.689	NC

Table 7 (n=100) indicated that, for the random sample plan, both point-biserial CTT and normalized point-biserial CTT item discrimination indicated sign of strong invariance (r = 0.766). The one-parameter IRT item discrimination estimate was not available. A sign of invariance was found in the two-parameter IRT item discrimination estimates (r = 0.760). However, the three-parameter IRT item discrimination demonstrated stronger invariance in the same sample plan (r = 0.781).

For the gender sampling plan, the correlation of CTT-based item discrimination estimates in point-biserial and normalized point-biserial indicated strong invariance (r = 0.749). For the same sampling plan, the IRT-based item discrimination estimates for the two-parameter model indicated sign of strong invariance (r = 0.748). However, the IRT item discrimination estimates for the three-parameter model indicated stronger invariance (r = 0.765).

For the ability sampling plan, a near collapse of invariance was indicated in the CTT –based item discrimination (r = 0.082). For the same sampling plan, the IRT-based estimates for the two-parameter model indicated moderate invariance of item discrimination (r = 0.689).

Table 8. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT Item Discrimination indexes (n=1000)

CTT MODEL			**IRT MODEL**		
	Fisher Transformed		**Fisher Transformed**		
Sampling Frame	**Point-Biserial**	**Normalized Point-Biserial**	**1PL**	**2PL**	**3PL**
Random Sample	0.741	0.741	N/A	0.738	0.740
Female - male samples	0.732	0.732	N/A	0.724	0.733
High - low ability samples	0.178	0.178	N/A	0.673	NC

Table 8 showed the results of Table 6 (n=1000) except that the sample correlations from Table 8 have been corrected for bias using Fisher transformed correction.

For the random sample plan, both the fisher transformed point-biserial CTT and normalized point-biserial CTT item discrimination indicated sign of strong invariance (r = 0.741). The one-parameter IRT item discrimination estimate was not available. A sign of invariance was found in the two-parameter IRT item discrimination estimates (r = 0.738). However, the three-parameter IRT item discrimination demonstrated stronger invariance in the same sample plan (r = 0.740).

For the gender sample plan, the correlation in both the fisher transformed CTT point-biserial and CTT normalized point-biserial showed strong invariance of item discrimination (r = 0.732). In the same sample plan, the IRT-based estimates for two-parameter model demonstrated strong invariance (r = 0.724) while three-parameter model showed stronger invariance (r = 0.733).

The ability sample plan indicated degeneration of the correlations found in the gender sample plan, corrected CTT p and Normalized CTT p correlation was poor (r = 0.178). For the same sampling plan, the IRT-based estimate for the two-parameter model had moderate invariance of item discrimination estimates (r = 0.673).

Table 9. Invariance of item statistics from the two measurement frameworks: correlations of CTT and IRT Item Discrimination indexes (n=100)

CTT MODEL			**IRT MODEL**		
	Fisher Transformed		**Fisher Transformed**		
Sampling Frame	**p values**	**Normalized p values**	**1PL**	**2PL**	**3PL**
Random Sample	0.703	0.703	N/A	0.687	0.689
Female - male samples	0.667	0.667	N/A	0.663	0.671
High - low ability samples	0.161	0.161	N/A	0.651	NC

Table 9 showed the results of Table 7 (n=100) except that the sample correlations from Table 9 have been corrected for bias using Fisher transformed correction. None of the correlations found in Table 9 matched those found in Table 7.

For the random sample plan, both fisher transformed point-biserial CTT and normalized point-biserial CTT item discrimination indicated sign of strong invariance (r = 0.703). The two-parameter IRT item discrimination estimates demonstrated moderate invariance (r = 0.687). However, the three-parameter IRT item discrimination demonstrated a little stronger invariance in the same sample plan (r = 0.689).

In the female-male sample plan, the correlation in the Fisher Transformed CTT point-biserial and CTT normalized point-biserial showed moderate invariance of item discrimination (r = 0.667). In the same sample plan, the IRT-based estimates for two-parameter model demonstrated a moderate invariance (r = 0.663), while three-parameter model showed a higher invariance (correlation coefficient = 0.671).

For the ability sample plan, results indicated degeneration of the correlations found in the gender sampling plan, fisher transformed CTT point biserial and Normalized CTT point-biserial correlation were poor (r = 0.161), indicating poor invariance. For the same sampling plan, the IRT-based estimate for the two-parameter model has moderate invariance of item discrimination estimates (r = 0.671).

Summary

In the theory of measurement, there are two common competing measurement frameworks, Classical Test Theory and Item Response Theory. The present study empirically examined how the item characteristics behaved under the two competing measurement frameworks. The study compared CTT- and IRT-based item characteristics, replicated the work done by Fan (1998). This study focused on three objectives: (1) how comparable are the CTT and IRT-based item difficulty and item discrimination; (2) which of the IRT models best fit the SSCE Mathematics and (3) how invariant are the CTT and IRT-based item difficulty and item discrimination across different samples of examinees.

The data used in this study were from National Examination Council, Minna, Niger State. The instrument for this study was marked Optical Mark Recorder (OMR) sheets containing the responses of candidates who took May/June 2008 NECO senior school certificate Mathematics examination paper 1 in Osun state. The NECO SSC Mathematics paper 1 which was a multiple choice examination paper, consisted of sixty items (60 items) based on the three years' senior secondary school examination in mathematics curriculum. A sample of 6,000 examinees, were randomly drawn from an examinee population of 32,460. The sample of 6,000 was composed of 3,000 males and 3,000 females.

To replicate the functionality of the two measurement theories in large scale measurement situations, one set of samples were randomly selected to equal with an n = 1,000. Conversely, to replicate clinical situations where tests are often constructed with small sample sizes, a second set of samples were randomly selected with an n = 100. Each of the samples were drawn under the three sampling plans, each progressively dissimilar, thus enabling theoretically greater disparity between the statistics conducted from the different samples.

Findings

(1) The three-parameter IRT model best fit the data used in this study.

(2) The IRT-based item characteristics estimates exhibited the invariance property consistently across different samples of examinees. That is, differences across samples of examinees have 5no significant influence on the item difficulty and discrimination estimates based on IRT.

(3) The IRT-based item estimates in the three-parameter model were more invariant than the one-parameter and two-parameter models.

(4) The three-parameter model had no convergence at low-high ability samples. It may be that it is not suitable for all samples of examinees unlike the two-parameter model that converged in all samples.

(5) All the statistics indicated a progressive decay in the correlations as the sampling frameworks became more dissimilar.

(6) Both CTT and IRT models can be used together in estimating item characteristics and in test development.

Conclusion

The three-parameter IRT model best fit the data used in this study although this model may not be suitable for all samples of examinees. Furthermore, two-parameter model IRT-based item parameter estimates exhibited invariance property consistently across different samples. This feature portrays IRT two-parameter model as a better option in giving adequate information concerning the behaviour of an item as well as the examinees irrespective of the samples.

Recommendations

The following are the recommendations: (1) IRT two-parameter model will be suitable for use irrespective of the samples; (2) For institutions and researchers that wish to use IRT in solving measurement problems should make efforts to use an appropriate model.

REFERENCES

Adedoyin, O.O, Nenty, H.J. & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educ. Res. & Rev., 3*(2), 83-93.

Awopeju, O.A. & Afolabi, E.R.I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *Eur. Sci. J., 12*(28), 263-284.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educ. & Psych. Measurement, 58,* 357-381.

Hambleton, R.K. (1994). Item response theory: a broad psychometric framework for measurement advances. *Psicothema, 6,* 535-556.

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

Mallikarjuna, G. & Natarajana, V. (2012). Investigate into invariance properties of item response theory (IRT) by two and three parameter models. *Int. J. Inform. Tech. & Business Management, 1*(1), 28-34.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *J. Math. Psych., 3,* 1-18.

Ojerinde, D. (2013). *Classical test theory (CTT) vs. item response theory (IRT): an evaluation of comparability of item analysis results*. Ibadan: Institute of Education.

Spearman, C. (1910). Correlation calculated from faulty data. *British J. Psych., 3,* 271-295.

✉ Dr. O. A. Awopeju (corresponding author)
Department of Educational Foundations and Counselling,
Obafemi Awolowo University,
Ile-Ife, Nigeria.
E-Mail: josyemus66@gmail.com