e-ISSN: 2714-5506
p-ISSN: 2714-609X

SQUARE : Journal of Mathematics and Mathematics Education
Volume 4, No. 1, 2022, pp. 43-51
http://dx.doi.org/10.21580/square.2022.4.1.11297

Model-Data Fit using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and The Sample-Size-Adjusted BIC

Jimoh Kasali[1*], Adediwura Alaba Adeyemi[2]
Obafemi Awolowo University, Ile-Ife, Nigeria
jimoh.bukola@yahoo.com

ABSTRACT

The study determined if the 1PL, 2PL, 3PL and 4PL item response theory models best fit the data from the 2016 NECO Mathematics objective tests. Ex-post facto design was adopted for the study. The population for the study comprised 1,022,474 candidates who enrolled and sat for June/July SSCE 2016 NECO Mathematics Examination. The sample comprised 276,338 candidates who sat for the examination in three purposively Geo political zones in Nigeria (i.e., S/West, S/East and N/West). The research instruments used for the study were Optical Marks Record Sheets for the National Examination Council (NECO) June/July 2016 SSCE Mathematics objectives items. The responses of the tests were scored dichotomously. Data collected were analyzed using 2loglikelihood chi-square. The results of the likelihood ratio test revealed that 2PL fitted the data better than 1PL was statistically significant (χ^2 (59) = 820636.1, $p < 0.05$); the 2PL model fitted the data better than the 1PL model; 3PL model fitted the data better than the 2PL model and the result showed that the 4PL model fitted the data better than the 3PL model and the Likelihood ratio test that 4PL model fitted the data better than 3PL model was statistically significant, ($\chi^2(60)=216159.2$, $p<0.05$). The study concluded that four-parameter logistic model fitted the 2016 NECO Mathematics test items.

Keywords: *Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), one parameter model, two parameter model, three parameter model, four parameter model.*

1. INTRODUCTION

1.1. Background

Assessment is critical in the development of strong student performance when using a classroom-designed instrument. This is because measuring variables is one of the processes in the research process (Eluwa & Abang, 2011). Item response theory (IRT), a new approach of assessing a test instrument's psychometric properties, is a current measuring methodology. It has gained traction and, as a result, has established itself as a prominent measurement framework. To produce a psychometrically sound cognitive assessment, a rigorous instrument development method is required. According to Ojerinde (2013), a latent trait or ability attribute is a variable that is not readily evident but has a quantifiable impact on detectable attributes. Standard measures, which are the test items, can be used to make inferences about the presence or extent of specific traits based on the perception of these attributes. The objects are supposed to have a direct connection to the latent attribute, and the items are considered restrictively independent. The response to an object should be completely captured by this ability characteristic. Any correlation between the items is due to their regular reliance on the authorized latent characteristic, and there should be no covariance among item responses along any other latent measurement for the premise of objectivity to be met. The item response theory, which is probabilistic in technique, is a subsequent conceptualisation of the essential connection between the answer to an item and the ability controlled by a person.

The general of the IRT Model is

$$P\,(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-a(\theta-b)}}$$

Thus, *e* is the constant 2.718. while 1.70 is the scaling factor, *b* is the difficulty coefficient; *c* is the discrimination parameter; *L*= *a* (*θ*-*b*) is the logistic deviate (logit); *θ* is the ability level".

Item response theory (IRT) is a hypothesis that describes how The degree of separation between items is depicted by the **a-parameter**, also known as item discrimination. The slope of the item characteristic curve at the point of inflection is what it's called (Harris, 1989). The a-paremeter can be negative (-) or positive (+), with 2.0 being the most frequent number for multiple choice questions. In a wide range of scenarios, an item with a low a-parameter discriminates ineffectively (ability). With a higher a-value, the item differentiates well, but only over a narrow range of a-values (ability). Items with an a-parameter of 0.80 or less are considered bad. At the time of reflection, the bigger the a-parameter, the more clearly the items discriminate among examinees. The a-value can be computed numerically or with the aid of a computer program. The b-parameters are item difficulty, also known as item threshold parameter. This is the affectation point on the ability scale, or the point on the scale when examinees have a 50% chance of responding appropriately to an item. The range of difficulty value can be - 3 to +3 when a scale is scaled with a mean of 0 and standard deviation of 1.0. Items with high b estimates are tough, and even if their values are low, steep test participants have a moderate likelihood of answering effectively to an item (Harris, 1989 in Ojerinde, Popoola, Ojo & Onyeneho, 2012). The speculating record or lower asymptote respect is the c-parameter, often known as a pseudo-chance parameter. An examinee who has no understanding what the best option is in a multiple-choice question can succeed in reacting successfully by speculating sporadically. Theoretically, guessing can range from 0.00 to 1.0, although it is rarely exactly ≤0.3.

Item response theory, in practice, involves applying curve-fitting techniques to observed proportions of category replies in the hopes that the fit is good enough to encourage faith in the model being fitted, according to Baker (1992). However, Garcia-Pérez and Frary (1991) pointed out that this approach has the fundamental contradiction of measuring fit after parameters have been set in order to fit the model as closely as possible. In other words, there is no technique for independently verifying the appropriateness of modelled response functions, making it impossible to validate the fit of IRT models to data (Garcia-Pérez & Frary, 1991). Examining both the model's goodness of fit and the number of parameters modeled to obtain that fit is a more advanced approach of picking a model than depending solely on fit statistics (Sclove, 1987). This is commonly accomplished by employing a penalty term that increases in size as the number of parameters in the fitted models increases.

A significant problem of modern science is the choosing of the most appropriate model to describe the events under investigation. Because statisticians are inherently involved in this activity, it's not unexpected that several statistical techniques to dealing with this critical issue have been proposed throughout the years. Model selection has been thoroughly researched from both a frequentist and a Bayesian standpoint. In the literature, many approaches for determining the "best model" from a group of contenders have been proposed. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are two of the most extensively utilized model selection families of indicators (BIC). The AIC is a Kullback–Leibler Divergence-based information-theoretic indicator that essentially measures the information lost by a given model. As a result, the AIC criterion is based on the idea that the less information a model loses, the higher its quality is. The BIC criteria is based on Bayesian theory and is aimed to maximize the posterior probability of a model given the data.

The Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for selecting a model from a finite set of options in statistics. It is closely related to the Akaike information criteria and is based in part on the likelihood function (AIC). It is possible to increase the likelihood of fitting a model by adding parameters, however this may result in overfitting. By inserting a penalty term for the number of parameters in the model, the BIC overcomes this problem.

1.2. Mathematically

The BIC is an asymptotic finding based on the assumption that the data distribution is exponential in nature. Let x = the observed data; n = the number of data points in x, or the number of observations, or the sample size; and k = the number of free parameters to be estimated. p(x|k) = the probability of the observed data given the number of parameters; or, the likelihood of the parameters given the dataset; if the estimated model is a linear regression, k is the number of regressors, including the intercept; if the estimated model is a linear regression, k is the number of regressors, including the intercept; if the estimated model is a linear regression, k is the number of regressors, including the intercept; L is the maximum likelihood function value for the calculated model.

The formula for the BIC is

$$-2 \cdot \ln p(x|k) \approx \mathrm{BIC} = -2 \cdot \ln L + k \ln(n). \tag{1}$$

$$\mathrm{BIC} = n \cdot \ln(\widehat{\sigma_e^2}) + k \cdot \ln(n) \tag{2}$$

where $\widehat{\sigma_e^2}$ is the error variance.

The error variance in this case is defined as

$$\widehat{\sigma_e^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{3}$$

The Bayesian information criterion's characteristics are listed bellow.

1. It is unrelated to the past or the prior is "ambiguous" (a constant).
2. It can be used to assess the parameterized model's accuracy in forecasting data.
3. It penalizes the model's complexity, which is defined as the number of parameters in the model.
4. It's almost the same as the minimum description length criterion, but with a minus sign.
5. It can be used to determine the amount of clusters to utilize based on the inherent complexity of a dataset.
6. It shares a lot of similarities with other penalized likelihood criteria like RIC and the Akaike information criterion.

In this vein, the Akaike information criterion (AIC) (Akaike, 1977) provides a model selection method that has been used in IRT on occasion (e.g., Wilson, 1992). According to one study (Sclove, 1987), the AIC is not asymptotically consistent, and it is questionable if the AIC is suitable for comparing models with a variety of parameters, such as local category models and cumulative border models. The ideal observer index (IOI) is an AIC extension that uses the probability ratio of two comparison models instead of a single baseline model (Levine, Drasgow, Williams, McCusker, & Thomasson, 1992). This index is more appropriate than the AIC when comparing models with different sorts of parameters. Wainer and Wright (1980) noted, "It appears that the Rasch model gives fairly reasonable assessments of ability and difficulty even when its assumption of equal slopes is very lightly approximated." Furthermore, "it appears that if the number of items is extremely large, then inferences about an examinee's ability based on his total test score will be very much the same whether" the Rasch model or the 3PL model is used, according to Lord and Novick (1968).

The application context, as well as one's philosophy of whether data should match the model or vice versa, should be addressed while deciding between the one parameter, two parameter, and three parameter models (e.g., sample size, instrument characteristics and considerations, assumption tenability, political realities, etc.). Given that the one parameter (1PL) model is the most restricting of the three, there has been a lot of research concerning how to use it when it mismatches. Forsyth, Saisangjan, and Gilmer (1981), for example, investigated the Rasch model's robustness when the dimensionality and constant assumptions were violated. It was assumed that some examinees would guess because their empirical data came from a multiple-choice question type test. "Even when the model's assumptions are not met, the Rasch model yields reasonably invariant item parameter and ability estimates," Forsyth et al. concluded. Simulated data was utilized by Dinero and Haertel (1977) to achieve similar conclusions. Wainer and Wright (1980) noted, "It appears that the Rasch model gives extremely excellent assessments of ability and difficulty."

The evaluation of model-data fit is divided into two stages: 1. The data test is fitted to the four IRT models, and the models' fitness to the data set is compared. The model that fits the data the best is named the model that fits the data. To achieve this purpose, a variety of strategies might be used. According to Oguoma, Metibemu, and Okoye (2016), the Chi-square difference test and information indices are relevant measures (Finch & French, 2015).

Model complexity is penalized in information indices, which are simply measurements of variation that a model does not explain. Among the most well-known of these indices are the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), and the sample-size-adjusted BIC (SBIC; Enders & Tofihi, 2008). These information indices are derived using the chi-square value of -2loglikelihood and interpreted in such a way that the model with the lower value fits the data better. The chi-square and likelihood ratio goodness of fit tests are also used to test the null hypothesis that two nested models provide the same fit to a set of data.

The models under investigation differ, as shown by a statistically significant probability. The model must be appropriate for the data, which is one of the most basic requirements in the use of parametric IRT models. This requires picking the right model and evaluating its fit (Edelen & Reeve, 2007). When choosing the right model, the number of item answer categories is the first thing to think about. The 1, 2, 3, and 4 IRT models can be used with dichotomous data.

1.3. Objective

Determine if the 1PL, 2PL, 3PL and 4PL item response theory models best fit the data from the 2016 NECO Mathematics Objective tests.

1.4. Research Questions

1. To what extent does 2PL fit data better than 1PL in 2016 NECO Mathematics test items?
2. Does 3PL fitted the data better than 2PL in 2016 NECO Mathematics test items?
3. Does 4PL fitted the data better than 3PL in 2016 NECO Mathematics test items?

2. METHOD

The study employed ex-post facto design. The population for the study comprised candidates who sat for June/July SSCE 2016 NECO Mathematics Examination. The sample comprised 276,338 candidates who sat for the examination in three purposively Geo political zones in Nigeria (i.e South-West, South-East and North-West). The research instruments used for the study were Optical Marks Record Sheets for the National Examination Council (NECO) June/July 2016 SSCE Mathematics objectives items. The responses of the tests were scored dichotomously. Data collected were analyzed using 2loglikelihood chi-square.

3. RESULTS

3.1. Result

Table 1 shows the model-data fit assessment of 2016 NECO Mathematics test items. The table shows that when the fitness of 1PL and 2PL models' data were compared, the result showed that the 2PL had AIC =17289402, SABIC =17290282, BIC =17290664 values that were less than the AIC =18109920, SABIC =18110368, BIC =18110562 values of the 1PL. In addition, the Likelihood ratio test revealed that 2PL fitted the data better than 1PL was statistically significant (χ^2 (59) = 820636.1, p < 0.05). These results showed that the 2PL model fitted the data better than the 1PL model.

Table 1. Model-Data Fit Assessment of 2016 NECO Mathematics Test Items

PL model	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	*df*	*p*
Comparing 1 and 2 PL model									
1	18109920	18109920	18110368	18110107	18110562	-9054899	820636.1	59	0
2	17289402	17289402	17290282	17289769	17290664	-8644581			

Table 2. Model-Data Fit Assessment of 2016 NECO Mathematics Test Items

PL model	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	*df*	*p*
Comparing 2 and 3PL models									
2	17289402	17289402	17290282	17289769	17290664	-8644581	169158.5	60	0
3	17120304	17120305	17121624	17120854	17122196	-8559972			

The results showed that the 3PL model fitted the data better than the 2PL model (3PL model's AIC = 17120304, SABIC = 17121624, BIC = 17122196 values were respectively less than the 2PL model's AIC =17289402, SABIC =17290282, BIC =17290664; the Likelihood ratio test that 3PL model fitted the data better than 2PL model was statistically significant, (χ^2 (60) = 169158.5, p < 0.05)). Furthermore, in search for a better model for the test data, the fitness of 3PL model to the 2016 NECO Mathematics test items were in turn compared to the fitness of 4PL model to the test data.

Table 1.3. Model-Data Fit Assessment of 2016 NECO Mathematics Test Items

PL model	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	*df*	*p*
Comparing 3 and 4PL models									
3	17120304	17120305	17121624	17120854	17122196	-8559972	216159.2	60	0
4	16904265	16904266	16906025	16904998	16906788	-8451893			

The results showed that the 4PL model fitted the data better than the 3PL model (4PL model's AIC = 16904265, SABIC = 17121624, BIC = 17122196 values were respectively less than the 3PL model's AIC = 17120304, SABIC = 16906025, BIC = 16906788; the Likelihood ratio test that 4PL model fitted the data better than 3PL model was statistically significant, (χ^2 (60) = 216159.2, p < 0.05)). The result revealed that unidimensional four-parameter logistic model fitted the 2016 NECO Mathematics test items. Thus, the test was calibrated using four-parameter logistic model.

3.2. Discussion of Findings

The 2PL model suited the data better than the 1PL model, according to the study's findings. Yen (1981) used simulation research to create several data sets based on various models and assess the fit of the 1PL, 2PL, and 3PL models to these data. When she utilized the 3PL model to generate data, she discovered that the 2PL model suited the data almost as well as the 3PL model, despite the fact that the item parameter estimates were not the same. It was difficult for the 2PL to simulate a nonzero lower asymptote when an item was challenging and had a moderate to high discrimination, she said. She concluded that, while the 2PL model performed almost as well as the 3PL model in modeling answer vectors, sample dependency may be observed when discrimination parameters for difficult items are calculated with low-proficiency examinees.

Mokobi and Adedoyin (2014) employed MULTILOG to examine item level and model fit statistics in a three-parameter logistic model with 2010 Botswana Junior Certificate Examination Mathematics paper one in a similar study. In order to test item fit to 1PL, 2PL, and 3PL models, the researchers used X2 goodness of fit statistics. The findings revealed that 10 things suit the 1PL model, 11 items fit the 2PL model, and 24 items fit the 3PL model.

The 3PL model suited the data better than the 2PL model, according to the findings. Using the more sophisticated 3PL model, on the other hand, resulted in a better fit than the 2PL model. However, in light of the increased model complexity, this is not regarded a significant improvement in fit. According to De Ayala (2009), the 3PL model fits much better than the 2PL and 1PL models, although it does not result in a significant fit improvement over either model. Additionally, Orlando and Thissen (2000) analyzed model-data fit from fixed format tests, the results showed that the three parameter logistic models combined with the generalized partial credit model among various were the best match. The logistic model with only one parameter has the most misfit elements. Three fit statistics are compared. Finally, the study's findings revealed that the 4PL model better fit the data than the 3PL model. Because models with more parameters tend to fit a data set better in general than models with fewer parameters, it's a good idea to factor in the extra parameters when evaluating model–data fit. One can decrease the tendency toward model over parameterization by considering the number of parameters necessary to attain a specific degree of fit (De Ayala, 2009).

4. CONCLUSION

Based on the findings, the study concluded that by examining the invariance of item parameter estimates across random calibration subsamples, the likelihood ratio and the AIC and BIC statistics were best approaches for assessing model–level fit and 4-PL was found to fit the test data. The following recommendations were made based on the findings of this study.

1. That the selection of best item response theory model should depend on assessing item fit statistics as the first step to apply item response theory with confidence.
2. That the use of more than one item response programs will provide the choice of the best program that provide more useful information about the real data set.
3. That the use of unidimensionality tests such as stout’s t-statistics, factor analysis and conditional item covariance should be adopted in model data fit.
4. That item response theory be further used by test developers, researchers and stakeholders to better understand how to develop psychometrically sound measures.

5. REFERENCES

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.*), Applications of statistics* (pp. 27−41). Amsterdam: North Holland.

Baker, F. B. (1992). Item response theory: *Parameter estimation techniques.* New York: Marcel Dekker.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res 16*(1) 5–18.

Eluwa, I., Eluwa, A., & Abang, B. (2011). *Evaluation of Mathematics Achievement Test: A comparison between classical test theory (CTT) and Item Response Theory (IRT).* Proceedings of the 2011 International Conference on Teaching, Learning, and Chance. International Association for Teaching and Learning (IATEL).

Finch. W. H., & French, B. F. (2015). The impact of group pseudo-guessing parameter differences on the detection of uniform and non-uniform DIF. *Psychological Test and Assessment Modeling, 56*(3), 25-44.

Garcia-Pérez, M. A., & Frary, R. B. (1991). Finite state polynomic item characteristic curves. British Journal of Mathematical and Statistical Psychology, 44,45−73.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

Dinero, T. E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 1,* 581–592.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5,* 175–186.

Harris, D. (1989). *Comparison of 1, 2, 3 parameter IRT models America college testing.* Program, NCME Instructional Model, POB 168, IOWA City, IA 52243.

Levine, M. V., Drasgow, F., Williams, B., Mccusker, C., & Thomasson, G. L. (1992). Measuring the difference between two models. *Applied Psychological Measurement, 16*,261−278.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: AddisonWesley.

Mokobi, T., & Adedoyin, O. O., (2014). Identifying location biased items in the 2010 Botswana junior certificate examination Mathematics paper one using the item response characteristics curves.

Oguoma, C. C., Metibemu, M. A., & Okoye, R. (2016). An assessment of the dimensionality of 2014 West African Secondary School Certificate Examination mathematics objective test scores in Imo State, Nigeria. *African Journal of theory and Practice of Educational Assessment, 4*(2), 43-45.

Ojerinde, D. (2013). *Introduction to item response theory, parameter models, estimation and application.* Abuja, Nigeria: Marvellous Press.

Ojerinde, D., Popoola, O., & Oyenneho, P. (2012). *Introduction to item response theory: parameter models, estimation and application.* Goshen Print Media Ltd.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45,* 373–391.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16,* 309.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied *Psychological Measurement, 5,* 245–262.