Psychometric Analysis of Mathematics Achievement Test Using Item Response Theory

Jolly S. Balila[1]

Norma G. Cajilig[2]

[1]Adventist University of the Philippines, Puting Kahoy, Silang, Cavite
[2]University of the Philippines, Diliman, Quezon City

Abstract

This study determined the psychometric properties of the Mathematics Achievement Test (MAT) using Item Response Theory (IRT). Three popular IRT models namely, 1PL, 2Pl, and 3PL IRT models were utilized following unidimensionality test. Data from 2448 second year high school students from selected public and private secondary schools in Region IVA were the subjects for analysis . Analyses of data were performed in R. Results indicated that the entire 50-item cognitive test did not meet the unidimensionality assumption. Items were grouped according to content strands and were subjected again to dimensionality test using Modified Parallel Analysis. The remaining items after some deletion process were subjected to item calibration. Data fit analysis were performed to each strand . The 1PL, 2PL, and 3PL IRT models fit the different strands of the MAT reasonably well. The researchers established an item pool that can be used in estimating students' mathematics ability.

Keywords: Item response theory, one –parameter logistics model, two-parameter logistic model, three-parameter logistic model.

I. INTRODUCTION

One of the important concerns in education is to develop a standard measure which can be used to estimate student achievement. Further, test developers are also concerned about the quality of test items and how examinees answer them. The measurement of cognitive ability has prominently featured the establishment of the psychology of science in general and the development of measuring instruments in particular. Problems are often encountered during the process of constructing instruments, problems such as lack of capacity to develop and process measures, and interpret them in a meaningful way. Thus, the development of standard measure for students is becoming more complicated (National Research Council, 2001).

According to Grigorenko and Sternberg as cited in De Beer (2004) who reviewed published empirical research on the reliability and validity of assessments, field of assessment has not yet lived up to its promise. The main practical and technical problem with assessment is finding suitable criterion measures to provide predictive validity evidence from learning potential measures. This has implications for establishing quality assessment for students.

Measurement is an important consideration in the construction of a quality student assessment even in the case of a classroom designed instrument. This is because measuring variables is a step in the research process (Eluwa, Eluwa & Abang, 2011). This concern can be addressed by the modern measurement approach called Item Response Theory (IRT), a new method for measuring the psychometric properties of a test instrument. It has been gaining ground, thereby becoming an important measurement framework. Developing a cognitive test that is psychometrically sound requires a thorough instrument development process. IRT models have a significant role in

questionnaire development and evaluation since they provide a clear information on the performance of each item in the scale and how the scale functions in measuring the construct of interest. IRT methods can lead to a short but reliable test for the population of interest. Procedures based on IRT have become increasingly more common in educational and occupational testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999)

Educational measurement and evaluation has been replete with studies focusing on IRT models despite the presence of numerous books, journals and internet resources that have been written exclusively on IRT to attest to the importance of these models in the development and analysis of test items. It is therefore the greatest aim of the researchers to use these models to address the measurement-related problems particularly on the analysis of a cognitive test.

Objectives of the study

This study determined the psychometric properties of Mathematics Achievement Test for High School (HSMAT) Students using IRT models. These models described the psychometric properties of High School Mathematics Achievement Test (HSMAT) strands with measurement precision. Specifically, this study sought to address the following:

a. To determine if the assumption of unidimensionality hold for the High School Mathematics Achievement Test.

b. To determine if the oneparameter logistic model (1PL), the two-parameter logistic model, (2PL), and the three-parameter logistic models (3PL) IRT models best fit the data from the mathematics achievement test.

Theoretical Foundations

The degree to which an instrument is valid and reliable, is an important indicator of an instrument's psychometric quality. The use of the modern psychometric methods can address this issue. This study was anchored on the modern psychometric method called Item Response Theory and Item Generation Theory. The principles of IRT are based on the two basic assumptions. First, a more able person could have greater probability of success on assessment of cognitive items such as mathematics achievement test items than a less able person. Secondly, any person should always be more likely to do better on an easier item than on a more difficult one. IRT assumes item difficulty and is characterized by influencing item difficulty estimates. This means that the items in a test should be written clearly and in a concise manner so that the items are not vulnerable to guessing (Linacre, as cited in Eluwa et al., 2011).

The theory of Item Development requires that the items should be constructed with a deliberate plan of action. It offers a conceptual framework for the task by giving structure, organization, and fluency. Rules, expressions, terminology, and many other aspects of item development can be consistently expressed when theory is followed. The significant aspect of this theory is that it commands and directs the item development activities so that constructs intended for appraisal are more likely to be accurately, fully, and appropriately addressed in the item. The rules for item generation were followed in order to arrive at the calibrated items ready for students' use (Osterlind, 2010).

Conceptual Foundations

Dimensionality. One important assumption of parametric IRT models is that a test which measures the construct is unidimensional, which means that the covariance among the items can be explained by a single underlying dimension. This has something to do with unidimensionality which refers to whether the instrument measures a single construct (Bond & Fox, 2001) or "the number of latent variables that account for the correlations among item responses in a particular data set" (Camilli, Wang, & Fesq, 1995, p. 80). To successfully measure a student's ability, confounding

variables should be removed from the instrument to ensure unidimensionality. The assumption of unidimensionality holds if a test measures a single construct; further, that the responses obey the principle of local independence, which states that item responses are independent conditioned on a particular level of ability (Nandakumar & Stout, 1993).

Item Response Theory. Item response theory is another branch of psychometric theory that may be regarded as roughly synonymous with latent trait theory. It is also referred to as the strong true score theory or modern mental test theory since IRT is the most recent body of theory with stronger assumption than classical theory. IRT involves a class of mathematical models used to predict examinee performance using item and person characteristics. These models have properties that offer many well-known advantages in testing applications. But the extent of which these properties are attained is dependent on the degree to which the IRT model itself is appropriate. IRT is a strong modeling method if assumptions are met. One important assumption of parametric IRT models is that the test that measures the construct is unidimensional, which means that the covariance among the items can be explained by a single underlying dimension (Kaplan & Saccuzzo, 1997; Magno, 2009).

There are several IRT models with potential application to educational research. The seven common models are the Rasch Model (1PL), the two-parameter logistic model (2PL), the three –parameter logistic model (3PL), graded model, nominal model, the partial credit model and the rating scale model. The item format for the first three models is dichotomous, for the last three models, it is polytomous (Embretson & Reise, 2000).

According to De Beer (2004), the first three general IRT models vary in terms of the item characteristics.

Each IRT model predicts the probability that a certain person will give a certain response to a certain item. The purpose of these models is to probably explain an examinee's responses to test items via a mathematical function based on his/her ability.

***The One-Parameter Logistic Model*.** The simplest Item Response Model for a dichotomous item has only one parameter. The 1PL (also known as the Rasch model) assumes that the difficulty parameter expresses the difficulty level of the item, the discrimination parameter equals one, and that there is no guessing parameter (Rizopoulus, 2006).

Difficulty is defined in both Classical Test Theory (CTT) and Item Response Theory (IRT) as the likelihood of a correct response, not in terms of the perceived difficulty or amount of effort required. Negative values in difficulty index indicate items that were easier to endorse, and positive values indicated items that were harder to endorse. In this model, it is possible to condition out or eliminate the student's abilities in order to estimate relative question difficulties; each response to each question must depend upon the ability and the question difficulty. When data fit the model, the relative difficulties of the questions are independent of the relative abilities of the students, and vice versa (Bhakta, Horton, & Andrich, 2005).

The Two-Parameter Logistic Model. The two-parameter logistic model allows for different discrimination parameters per item and assumes that the guessing parameter equals 0 (Rizopoulous, 2006). Item discrimination is a measure of how well an item is able to distinguish between examinees who answered the item correctly. When the discrimination index is high it means that the item differentiates (discriminates) between examinee.

The two-parameter logistic model (2PL) allows the slope or discrimination parameter (*a*) to vary across items instead of being

constrained to be equal as in the oneparameter logistic or Rasch model.

This means that both item difficulty (b) and item discrimination (*a*) are included in the exponential form of a logistic model. The relative importance of the difference between a person's trait level and item threshold is determined by the magnitude of the discriminating power of the item (Embretson & Reise, 2000. The constant, 1.7, is added to the model as an adjustment so that the logistic model approximates the

normal ogive model (Thissen, Steinberg & Wainer, 1993).

***The Three-Parameter Logistic IRT Model*.** This model is called three-parameter logistic model (3PL), and *a, b,* and *c* which are often called by their practical interpretations: discrimination, difficulty, and guessing, respectively. Including cparameter in the model was to allow for statistical adjustment for Item response function for the nonzero performance of low-proficiency examinees on multiple choice items. The *c*-parameter is sometimes called the guessing parameter because examinees with very low ability would be expected to get the item correct only by guessing (Han, 2012).

Model Fit. Model-data fit issues are a major concern when applying item response theory (IRT) models to real test data. Model fit is defined as how well the model as a whole explained the data. When a model is over identified, it is expected that model fit will not be perfect; it is therefore necessary to determine the actual degree of model fit, and whether the model fit is statistically acceptable. Ideally, indicators should load only on the specific latent variable identified in the measurement model (Kline, 2010).

One of the basic assumptions of the application of parametric IRT models is that the model is appropriate for the data. This involves choosing the right model and the evaluating model fit (Edelen & Reeve, 2007). The first consideration when choosing the right model is the number of item response categories. The 1, 2, and 3 IRT models can be used for dichotomous data.

II. METHODOLOGY

Research design

This test-validity study further aimed to demonstrate the process of item calibration using three major IRT models, namely: the one-, two-, and three–parameter logistic model. However, prior to item calibration, it was necessary to check the unidimensionality of the test as well as its subtests.

Traditionally, IRT models have been based on the assumption that the item pool being analyzed is effectively unidimensional. This study focused solely on unidimensional parametric IRT models.

Instrumentation

The HSMAT was a researcherconstructed instrument which consisted 80 items. These were reduced to 50 after the instrument was subjected to content validation. The test was found to have a Cronbach alpha reliability of .79. The test covered topics in Elementary Algebra for first year high school and Intermediate Algebra for second year high school. It includes concepts in Exponents and Radicals, Algebraic Equations and Functions, Special Products and Factoring, Quadratic Functions, Variations and Arithmetic Sequences which were based on the 2002 Basic Curriculum of the
Department of Education.

Respondents of the Study

Data were collected from 2,448 second year high school students enrolled in three public and two private secondary schools in Region IV-A. The sample students were predominantly females (1,462 or 61.56%) and came from public schools (1,763 or 72%).

Data Gathering Procedures

The school's mathematics teachers administered the test to the second year high school students, 2 weeks before they took the National Achievement Test (NAT). The students were given 1 hr 30 min. to answer the test which was under the supervision of their math teachers. Their mathematics teachers personally administered the test in order to minimize the monitoring effect of the proctor on the actual scores of the students. The test materials were retrieved right after the test.

Ethical Considerations

The test papers were treated with utmost care and confidentiality. The school heads were assured that the data will be used for research purposes only. The principals were also assured that the scores of the respondents would not be compared across the five schools.

Analysis of Data

The dimensionality using Modified Parallel Analysis (MPA) was investigated in this study in order to report evidence of validity, which ensures that the items are assigned to the same dimension. It tests whether the items were measuring one underlying dimension or several separate dimensions. The method of MPA compares the eigenvalues from the created data to those estimated from real data (Hambleton, Swaminathan, & Rogers, 1991).

The calibrations of items were performed using the three IRT models for dichotomous items, i.e., the one- parameter (1PL), the two- parameter (2 PL) and the three – parameter (3 PL) logistic models.

Model fit was explored under the three popular IRT models for dichotomous data. The fit to the IRT model is achieved when a summary chi-square interaction statistics turn out to be non-significant, showing no deviation from model expectation. The item and person summary fit statistics show a mean of zero and a standard deviation of 1, where individual items show non-significant chi-square fit statistics (Latimer, Covic, Cumming & Tennant, 2009).

Akaike's Information Criterion (AIC) and Bayesian Information Criteria (BIC) can be additional information that may compliment the Likelihood Ratio Test. The AIC is an indicator of comparative fit across nested models with an adjustment for model complexity. The AIC is not an indicator of fit for a specific model, but instead the model with the lowest AIC from among the set of nested models is considered to have the best fit. When comparing fitted items, the smaller the AIC or BIC, the better the fit (Acquah, 2010).

The Likelihood Ratio was used to determine the fit of the IRT models to HSMAT and strands data, which were previously found to provide good fits to several cognitive ability tests. The hypotheses on unidimensionality assumption and best model fit among IRT models were tested at .05 and .01 levels of significance.

Item responses were scored and transformed into binary data. An item coded "0" indicated an incorrect response while an item coded "1" indicated correct response.
The dimensionality and model fit tests were done in R, a free software programming language and a software environment for statistical computing and graphics.

III. RESULTS

A test of unidimensionality using Modified Parallel Analysis (MPA) was performed to the 50-item High School Mathematics Achievement Test (HSMAT).

The results revealed that the 50– item HSMAT deviated significantly from the unidimensional model ($p < .05$), implying that the mathematics test is multidimensional. This could be due to the fact that the test was developed with several strands. The hypothesis that states that the cognitive test in mathematics considering all items is unidimensional is thereby rejected.

Since multidimensionality was evident for the entire HSMAT, the items in each content strand were subjected to unidimensionality test. Table 1 displays both the second eigenvalues and the average of the second eigenvalues in each strand. Based on results, the eigenvalues in the observed data and Monte Carlo samples are all less than 1($p > .05$) therefore the null hypothesis of unidimensionality could not be rejected at the .05 level of significance for each of the strand

.

Table 1
Dimensionality Analysis of the HSMAT Strands

Content Strands	Second Eigenvalues in observed data	Average of Second Eigenvalues in Monte Carlo Samples	p-value
1.Exponents	0.38	0.44	0.17
2.Radicals	0.18	0.35	0.15
3**. Algebraic Equations and linear functions**			
Linear equations –one unknown	0.46	0.45	0.36
Linear equations-two unknown	0.40	0.35	0.09
Special product			
Special product A	0.52	0.37	0.12
Special product B	0.32	0.32	0.23
Quadratic functions	0.32	0.31	0.46
Variations	0.06	0.87	0.75
Arithmetic sequence	0.33	0.30	0.24

However, when analyzed by strand, Exponents, Linear Equations in Two Unknown, Special Products A and B, Variations, and Arithmetic Sequences turned out to be unidimensional.

Content Strands	Original Items
1.Exponents	9,10,28,46
2.Radicals	12,17,21,31,(37)
3**. Algebraic Equations and linear functions**	
Linear equations – one unknown	1,4,14,18,(22),26, 27,29,30,32,33,42
Linear equations-two unknowns	16,19,20,36,40,42
Special product	
Special product A	3,24,39,44,45,50
Special product B	2,25,38,49
Quadratic functions	6,23,35,43,(34)
Variations	5,7,11
Arithmetic sequence	8,13,15,47
TOTAL NUMBER OF ITEMS RETAINED	47

Note: Deleted items are in parenthesis

Items that were removed during the first analysis under Special Products were again subjected to dimensionality test and formed a secondary dimension called Special Product B.

The ten items on Special Products created two unidimensional strands labeled as Special Products A and B. The items retained and deleted are shown in Table 2, where one item was removed from each of the following strands: Radicals, Linear Equations in One Unknown and Quadratic Functions.

Best-Fitting IRT Models Following the unidimensionality test was the item calibration process. Three IRT models for dichotomous data were tested to determine which model significantly best fit each of the strands in the HSMAT. These IRT models included the Rasch model (one-parameter logistic model constrained to one), the one-parameter logistic model (not constrained), the two-parameter logistic

model (2PL), and the three-parameter logistic model. The main model fit statistics used was the Likelihood Ratio and was complimented by the BIC and AIC. Based on the model fit test, only unconstrained 1PL, the 2PL and the 3PL appeared to be the most appropriate models for the HSMAT strands ($p < .01$ and .05). The unconstrained 1PL best fit the strand Radicals; the 2PL best fits Special Products B and Arithmetic Sequence. Meanwhile, the 3PL best fit the following strands: Exponents, Linear Equations in One and Two Unknowns, Special Products A, Quadratic Functions and Variations. The constrained 1PL model was proven to be not fitted statistically to any of the HSMAT strands. Table 3 presents the summary of the bestfitting IRT models for the HSMAT strands.

Table 3

Summary of Best Fit IRT Models for HSMAT Strands

Content Strands	Best-Fitting IRT Model	Significance
1.Exponents	3 PL	P < .001
2.Radicals	1 PL(U)	P < .001
Algebraic Equations and linear functions 3. Linear equations – one unknown	3 PL	P < .001
4. Linear equations-two unknown	3 PL	P < .001
Special product		
Special product A	3 PL	P < .05
Special product B	2 PL	P < .001
Quadratic functions	3 PL	P < .001
Variations	3 PL	P < .05
Arithmetic sequence	2 PL	P < .001

In the model fit and parameter estimation process, the discrimination (*a)* difficulty (*b*), and guessing (*c*) of the IRT parameters were considered in the decision process for selecting a psychometrically sound items. It is also the basis of what items to be deleted for the final version of HSMAT.

A total of 47 items were subjected to parameter estimation process after dimensionality test per strand. From this process, 17 items were found to be problematic for the following reasons: the items were either very difficult or very easy, nondiscriminating, and had *c*- parameter higher than .30.

IV. DISCUSSION

Dimensionality

Unidimensionality is the most important assumption common for all IRT models. It assumes whether a dominant factor exists among all the items in the test. In this study the whole 50–item HSMAT is multidimensional based on Modified Parallel Analysis (MPA). Because of this result, MPA for each mathematics strand was performed.

The assumption for unidimensionality for each strand of the HSMAT should be established before IRT applications. In this analysis when the p-value is greater than .05 or 5%, the test is unidimensional. After subjecting to MPA, the theory was confirmed. The items were statistically loaded to the same mathematics strands the way these items were originally constructed by the researchers, except that, one item each was removed to Radicals, Linear equation in one unknown, and Quadratic functions to meet the assumption. However, the items under Special product formed two sub strands and were labeled A and B. According to Child and Opler (1999), it is possible that other subsets of items may form distinct dimensions for the purpose of IRT calibration.

All in all, there were three items removed from the 50-item cognitive test in mathematics as reflected in Table 3. Item 37 was expressed in *number sentence*. Probably the student can easily understand the problem if it is expressed in mathematical symbols like the symbol for square root ($\sqrt{}$). Item 22 was the only item constructed with *inequality symbol* that may attribute to the departure from unidimensionality of the items under linear equation in one unknown. One item in quadratic function (Item 43) was removed to meet the unidimensionality assumption of this particular strand. Probably, students were not familiar of the different figures as one of the options on the test.

The result of the dimensionality test is consistent with the study conducted to assess the dimensional structure of mathematics achievement test among grades

3-8. Mathematics achievement test are complex and exhibit multidimensionality (Burg, 2008). Jang and Roussos (2007) investigated the dimensions of two forms of test in English as Foreign Language and the results also revealed that the two studies have a strong evidence of multidimensionality. With the use of confirmatory factor analysis the dimensionality structures of the test were identified.

Model Fit Test

As seen in Table 3, the threeparameter IRT model best fit most of the strands of HSMAT. This means that the difficulty, discrimination, and guessing parameter should be considered in developing a psychometrically sound test. It cannot be denied that some students guess answers to test items if they don't know the answer. Therefore, the result suggests that during the validation process, the guessing parameter should be part of the process.

In this study it was suggested based on the result that the items under the strand *Radic*al that only difficulty or the one-parameter be considered, and for Special products B and Arithmetic sequence, the difficulty plus the discrimination or the two-parameter was suggested.

Assessing goodness of fit of item response theory models typically involves evaluating differences between observed and expected score response distributions using a chisquare test statistic. When these methods are applied to assessments that are shorter in length, uncertainty with which ability is estimated greatly affects the approximation to null chisquare distribution. (Stone & Hansen, 2000).

V. CONCLUSIONS

The use of IRT provides this research a powerful tool for evaluating the psychometric properties of the High School Mathematics Achievement Test. The HSMAT was multidimensional based on the Modified Parallel Analysis (MPA). This multidimensionality was explained by the several content strands comprising the test.

This study established an item pool that can be used in estimating students' cognitive ability in Mathematics based in IRT methodologies. Of the 50 items in the High School Mathematics Achievement Test, six were recommended for deletion and 20 items for revision. Twenty one (21) items were considered psychometrically *good* items because the psychometric properties had been carefully established using IRT. Item Response Theory is an important tool in the development of standard metrics for measuring cognitive test.

VI. RECOMMENDATIONS

Based on the findings several recommendations were made. The recommendations such as: use of other unidimensionality tests like Stout's T statistics, Factor Analysis and Conditional Item Covariance to verify if the result is consistent with MPA; that IRT be further used by researchers to better understand how to develop psychometrically sound measures; use of other IRT applications that can be further explored and can be used in the measurement process like the analysis of polytomous items using Graded Response Model, Nominal Model and Partial Credit Model and Rating Scale model; that calibrated items of the MAT test be further analyzed, particularly examining the option characteristics of the test; and IRT methodologies should be introduced to Higher Education Institutions especially test validation to improve teachers' assessment practices.

REFERENCES

Acquah, H.D. (2010). Comparison of Akaike information critierion (AIC) and Bayesian information criterion (BIC) in selection of an Asymmetric price relationship. Journal of Development and *Agricultural Economics, Vol. 2(1,* 001-006. Available online at http://www.academicjournals .org/JDAE

American Educational Research Association, American Psychological Association, & National Council on

Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in udergraduate medical education. *BMC Medical Education* 5-9, doi:10.1186/1472-6920/5/9.

Bond, T. G., Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ"Earlbaum.

Burg, S.S. (2008). *An investigation of dimensionality across grade levels and effects on vertical linking for elementary grade mathematics achievement tests.* National Council on Measurement and Evaluation. Retrieved from https://cdn.lexile.com/m/uploads/whitepapers/BurgNCME2008_MetaMetricsWhitepaper.pdf

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the laws chool admission test. *Journal of Educational Measurement*, *32*(1), 79–96.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. SA *Journal of Industrial Psychology*, 30(4), 5258.

Childs R. A. & Oppler S. H. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a High Stakes Testing Program. *Educational and Psychological Measurement*. Sage: Vol. No. sage. 6, (939-955).

Edelen, M.O., & Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res* (2007) 16:5–18 DOI 10.1007/s11136-007-9198-0

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).

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists*. London: Laurence Erlbaum Associates, Inc.

Hambleton, R.K., & Swaminathan, H. (2001). *Item response theory: Principles and applications.* Boston/Dordrecht/Lancaster. John Wiley and Sons, Inc.

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

Han, K.T .(2012, January). Fixing the c Parameter in the ThreeParameter Logistic Model. *Practical Assessment, Research and Evaluation*. Vol. 17. No1.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariancebased nonparametric approach. *Journal of Educational Measurement*, 44, 1-22 Kaplan, R. M. & Saccuzo, D.P. (1997). *Psychological Testing: Principles, applications and issues.* Pacific Grove: Brooks Cole Pub. Company.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Latimer, S., Covic, T., Cumming S., & Tennant, A. (2009). Psychometric analysis of the Self-Harm Inventory using Rasch Modelling. *BMC Psychiatry*. 9:53, doi:10.1186/1471-244X-9-53.

Magno, C., (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological* Assessmen.t Vol. 1, Issue 1, pp. 1-11

Rizopoulos, D. (2006). Ltm: an R package for latent variable modeling and item response Theory Analyses. *Journal of Statistical Software*. Vol. 17, Issue 5. Retrieved at http://www.jstatsoft.org/.

Nandakumar, R., & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*(1), 41–68.

National Research Coucil. (2001). *Adding it up: Helping children learn mathematics*. In J. Kilpatrick, J. Swafford, & B. Findell (Eds.) Mathematics learning study committee, center for education, division of behavioral and social sciences and education (p. 375). Washington, DC: National Academy Press.`

Osterlind, S.J. (2010). *Modern measurement: Theory, principles, and application of mental appraisal*. (2nd Ed.). NY:

Pearson Education, Inc.

Stone, C.A. & Hansen, M.A. (December 2000). The effect of errors in estimating ability on goodness-offit tests for IRT models. *Educational and Psychological Measurement*. Vol. 60. No.6. pp. 974 – 991.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W.

Holland & H.

APPENDIX A.
Sample Output for IRT Best Fitting Model

Model Comparisons for Linear Equations in One Unknown

MODEL	AIC	BIC	Log.lik	LRT	df	p-value
1PL©	36008.05	36077.69	-7992.03			
1pl(U)	35420.57	35496.01	-17697.28	589.48(1PL©vs. 1PL(U)	1	<.001
2PL	34973.65	35112.93	-17462.80	468.91(1PL(U)vs. 2PL)	11	0.084
3PL	34898.08	35043.15	-17424,04	77.58(2PL vs. 3PL)	1	<0.001

APPENDIX B
IRT MODELS FORMULAS
Three – parameter logistic function

$$P(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}$$

where P(y) indicates the probability of correct response given y and the item parameters (more fully expressed as P(x¼1|y, a, b,c)). The subscript i indicates the item, i. The e in the function is a mathematical constant, the exponential function, approximately 2.718. Its counterpart is the natural log function; the natural log of e¼1.4 The 1.7 is a scaling parameter; it is not necessary, but omitting it would change the scale of the a-parameter.

Two-parameter logistic function

$$P(\theta) = \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}.$$

For the 2PL model, the lowerasymptote's value is fixed to zero.

One –parameter logistic Function

$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$ The 2PL

and 1PL IRT models are special cases, or constrained versions, of the 3PL model. To constrain a model means to fix the value of one or more of the parameters.

APPENDIX C.
Sample items in HSMAT

38. The area of a square is represented by $x^2 \square \square 8\ 16x$. What is the length of each side?

A. $x \square 8$ C. $x \square 4$

B. $x \square 4$ D. $x \square 8$

39. The area of a rectangle is represented by $ac\ ad\ bc\ bd \square \square \square$, what represents one of its sides?

A. $a\ c \square$ C. $a\ d \square$

B. $a\ b \square$ D. $b\ c \square$

Authors:
*Jolly S.Balila, Ph.D. Research and Evaluation, University of the Philippines, Diliman; Director, University Research Center, Adventist University of the Philippines, Puting Kahoy, Silang, Cavite

Norma G. Cajilig, Ph.D, Retired Professor, University of the Philippines. Diliman, Quezon City.