Comparison of Item Difficulty Estimates in a Basic Statistics Test using ltm and CTT Software Packages in R

Jonald L. Pimentel[1*]
Department of Mathematics and Statistics
University of Southern Mindanao
Kabacan, Cotabato, Philippines

Marah Luriely A. Villaruz[2]
Ascend E-Commerce Phils., Inc.
Pasig City
Philippines

***Abstract*—Two free computer software packages "ltm" and "CTT" in the R software environment were tested to demonstrate its usefulness in an item test analysis. The calibration of the item difficulty parameters given the binary responses of two hundred five examinees for the fifteen items multiple choice test were analyzed using the Classical Test Theory (CTT) and Item Response Theory (IRT) methodologies. The software latent trait model "ltm" employed the IRT framework while the software classical test theory functions "CTT" operated under CTT. The IRT Rasch model was used to model the responses of the examinees. The conditional maximum likelihood estimation method was used to estimate the item difficulty parameters for all the items. On the other hand, all the item difficulty indices using the "CTT" software were also calculated. Both the statistical analyses of this study were done in the R software. Results showed that among the fifteen items, the estimates of their item difficulty parameters differed mostly on their values between the two methods. In an IRT framework, items showed extreme difficulty or easy cases as compared to CTT. However, when the estimated values were categorized into intervals and labelled according to its verbal difficulty description, both methodologies showed some similarities in their item difficulties.**

***Keywords*—Classical test theory; indices; item calibration; item difficulty; item response theory; R software**

I. Introduction

In the field of education particularly in test and measurement, it is important that any method that uses technology should be upgraded from time to time. This technology that performs computing and analysis requires speed and precision especially if the data is huge. Hand computation seems to be tedious and possible but it will take a longtime. In the case of a test or item test analyses, it is important that the item calibration for the estimates of its item parameters is accurate, fast and reliable. Statistical software packages that perform these calculations are available, either purchased commercially or as a free software in the internet.

Test item analysis is very important especially in the test construction. First, test can be classified with its degree of difficulty and second, for item banking that is, the calibrated items are stored traditionally in a box or electronically in a database. These items were given labels for its corresponding levels or index of difficulty of which it can be retrieved anytime for test construction. This method is useful for test makers in the composition of test items and the determination of the difficulty or easiness of the test instrument.

The primary objective of this paper is to demonstrate the usefulness of the two computer software programs, the latent trait model "ltm" [1] and the classical test theory functions "CTT" [2] in the R software environment in the calibration of item difficulty parameter estimates/indices for a multiple-choice test. Two methodologies, the item response theory and the classical test theory will be used. Specifically, this study will employ the Rasch model [3], an IRT probabilistic model which is part of the logistic model family, to model the responses of all the examinees for all the items. Estimation for all the item difficulty parameters will be carried using the conditional maximum likelihood estimation [4]. The calculated item difficulty estimates will be compared to the calculated difficulty indices of the same test examination that uses the scores of the examinees under the classical test theory methodology. One point of interest in this study is the comparison of the verbal description of the items in terms of its difficulty labels. Here we will know whether each item estimates are comparable for both methodologies or they both possess extreme differences.

II. Background of the Study and Related Studies

A. Item Calibration

In the calibration of item parameters, specifically the difficulty indices β of an examination test say in the case of a multiple-choice test in which the resulting data is a matrix of binary responses of the number of examinees who took the examination and the number of items being answered, Two methodologies are available at present in the literatures to handle such calibration. These methods are the Classical Test Theory (CTT) which is based on prediction of outcomes on a test that is, in particular an examinee's observed score which is composed of a true score and an error score and the Item Response Theory (IRT) which is based on a response probabilistic modeling [4]. CTT usually do the estimation of the reliability of a test and the item difficulty indices which comes from the score of the examinees. In practice, these indices are also known as the p-value and is valued from 0.0 to 1.0 for each item and it is based on the proportion of all the examinees who got the correct answers over the total

*Corresponding Author

examinees. The higher the proportion of getting correct answers, the easier is the item. CTT however, has many limitations as cited by [5]. On the other hand, test makers are also adopting model based IRT because it is powerful and can provide a framework for evaluating how good assessment do its job and how good its item do its job. For calculating item difficulty for example in a multiple-choice test, IRT traditionally applied based on a large number of historical correct or incorrect information gathered from the test [6] and in turn applies probabilistic models as mention by [4]. See also [7], [8], [9] and [10] for more discussion about these item response theory modeling.

B. Available Statistical Software Programs

Statistical software packages are presently available for calibrating item parameters in which CTT and IRT models are used. These includes powerful commercial software such SAS [11], STATA [12], SPSS [13], M plus [14], the BILOG-MG software [15] and ConQuest software [16] for fitting item response latent regression models and many more. However, there are some commercial software packages that are not easy to learn as well, hence it is must to do an extensive training if you want learn it because some of the software corresponding documentations are difficult to comprehend and sometimes have program failures and limitation which can be frustrating.

There is also a software package developed by The National Institute for Educational Measurement of the Netherlands (CITO) called OPLM [17] which is free and can be obtained by request. Starting in the year 2000, a quite number of new IRT packages uploaded as a library were developed in the open source in R software environment [18]. These includes computer software programs called the latent trait model (ltm) which was intended for unidimensional item response theory [1] as mentioned earlier, the extended Rasch models called eRm [19], the software called mirt which is intended for multidimensional IRT [20], and the software called mlirt which is intended for multilevel and Bayesian estimation [21]. Also, a software in R that uses Bayesian methods is also available called R2WinBuGs [22]. Lastly the software called "CTT" is intended for the estimation of items parameters under the classical test theory methodology [2].

III. Material and Methods

A. The Dataset

The data used in this study were the responses of two hundred five (205) students who responded to a fifteen (15) items multiple choice test. The test was just part of the Basic Statistics Preliminary Examination of the Mathematics and Statistics Department of the College of Science and Mathematics, Mindanao State University –Iligan Institute of Technology during the second semester school year 2014-2015 [23]. The test questionnaire was made by the authors and was validated for its content. The responses of these students were tabulated in a 205 by 15 matrix of 1's, when student got correct answer to the given item and 0's, when student got a wrong answer to the given item (see Table I for the illustration).The data then was stored as a text file having file name. The data was processed using a personal computer.

TABLE. I. Illustrating the Data Matrix

examinee	Item 1	Item2	…	Item 15
1	1	0	.	0
2	0	0	.	1
.	.	.	.	.
.	.	.	.	.
205	1	0	.	1

B. Item Response Theory Models for Binary Response

In an IRT framework, one can specify the components affecting the probability that an examinee will respond in a particular way to a particular test item. We can choose a particular measurement model that will relate the responses of the examinees and the qualities of the items. In this study the Rasch model was the model used to obtain the estimates of the item difficulty applying the conditional maximum likelihood estimation method. This model is one of the simplest item response theory models [24]. In general, the model is characterized as a two parameter models with the ability parameter of the examinee and the other parameter is the characteristics of the item which is the difficulty parameter [25]. The model is given by.

$$P\left(X_{nj}=1\middle|\theta_n,\beta_j\right)=\frac{e^{(\theta_n-\beta_j)}}{1+e^{(\theta_n-\beta_j)}}$$

where X_{nj} refers to a response made by an examinee n to an item j, θ_n refer to the trait level or ability of an examinee n; and β_j refers to the difficulty characteristics of the item j and it may take values between -3 (Very easy) to 3 (very difficult). The expression, $P\left(X_{nj}=1\middle|\theta_n,\beta_j\right)$ is the chance or the likelihood that an examinee n will give an answer to an item correctly conditional to his ability (θ_n) and the difficulty of the item (β_j). It is common in IRT that all measurements in the ability and difficulty before being subjected to estimation under the Rasch model are transformed into standard normal so normal measurements will be used. In the Rasch modeling, an examinee's answer in a form of a dichotomous response (that is, in our data, 1 refers to an examinee who got a correct response to an item while the entry 0, means that the examinee got a wrong answer) can be explained by the examinee's ability and the difficulty characteristic of the item. Now, in order for the model to be generalized, assumptions are considered that will make the model hold. Please see [26] and [27] for more explanations. The majority of applications of item response theory models usually to categorical data as known in [3], [7], [8], [9] and [10] but they were also being applied to data where there are continuous responses. These can be seen in the literatures [28] and [29].

C. Information Characteristics Curve (ICC) under the Item Reponse Theory

The Information Characteristics Curve (ICC) represents the item response function (IRF) which is the likelihood or chance of getting a positive response to each item which is represents the function of the proficiency or ability θ of the examinees. One can represent in the same graph the observed and the expected ICCs to get the fit of each item [30].

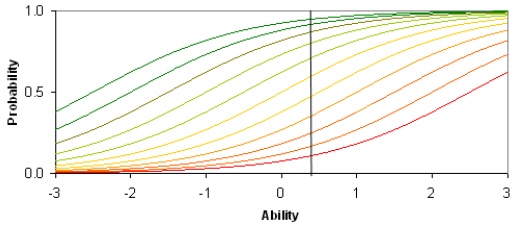

Fig. 1. Information Characteristics Curves (ICCs).

Fig. 1 illustrates ICCs for a number of items. ICCs highlight the change in the chances or likelihood of a successful response for an examinee with its ability location at the vertical line. The examinee will likely respond correctly to the easiest items (with locations to the left and higher curves) and unlikely to respond correctly to difficult items (locations to the right and lower curves) that is, the x-axis is the theoretical ability or proficiency level, ranging from -3 to +3. This graph only represents theoretical modeling rather than empirical data. To be specific, there may not be examinees that can reach a proficiency or ability θ level of +3 or fail so miserably as to be in the -3 group. Nonetheless, to study the characteristics of an item, we are interested in knowing, given a person whose θ is +3, what the probability of giving the rights answer to an item. The ICC indicates that when θ is zero, the examinee is on an average ability or proficiency hence, the chances of the examinee of answering the item correctly is approximately 0.5 or 50%. When the ability level θ is -3, the probability is almost zero to correctly get the item. When θ is +3, the probability to correctly answer the item increases to 0.99 or 99%.

D. Maximum Likelihood Estimation for the Item Difficulty with the Rasch Model

In order to calculate the values of the estimates of the item difficulty parameters, the computer software program "ltm" which means Latent Trait Models under the IRT framework [1] will be used in the calibration of item difficulty estimates and is based on the environment of the software R. Further, the "ltm" adapted both the conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation methods that handles the calculation in estimating item parameters and person parameters mathematically. In our study we employed the conditional maximum likelihood to calculate the item difficulty parameter. For more details, please see [6], [27], [31] and [32]. Although, maximum likelihood methods are the common estimation methods for years in the calibration of examinee's proficiency or ability and item parameters particularly the discrimination, difficulty and the guessing parameters another alternative estimation method emerged. The development of the Bayesian framework as an alternative but very powerful sampling-based estimation techniques have encouraged the application of Bayesian methods. The Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling and Metropolis-Hastings (M-H), were used to simultaneously estimate all model parameters. An MCMC implementation are introduced for the sampling of all model parameters that combines various advantages of different MCMC schemes for sampling IRT parameters [33]. For example [34] estimated the ability and item parameters of the IRT model for the observed data using MCMC.

The Bayesian inference requires the computation of the posterior distribution for a collection of random variables (parameters or unknown observables). At present, numerous simulation-based methods emerged. On sampling by [35] and [36], Other Bayesian works see [37] and [38]. Statistical software packages like WinBUGS (Bayesian inference Using Gibbs Sampling) [39] is a popular software for analyzing complex statistical models using MCMC methods.

E. Classical Test Theory

The classical test theory (CTT) is a theory of measurement error. The classical test theory has an assumption that each examinee's actual observed score X is the sum of the examinee's true score T and the error score E that is X=T+E. The key concepts of this theory involved the determination of test's reliability and validity for which test can be assessed mathematically [40]. The study and application for the classical test theory has been continuing which can be seen in the literatures [41]. Further, major applications of this theory are also on the test and item analysis and observed score equating. An article published in [42] looks for working on the classical test theory in combination with the concept of the item response theory. Their paper, emphasized that since the classical test theory was built in the assumption of exchangeability and the item response theory was based on conditional independence then they concluded that item response theory can be considered as an extension of the classical test theory where the concepts for both theories are related with each other. What is interesting in their work is the capability of IRT to provide the classical test theory statistical values where it can provide.

In our study, the software package "CTT" will be used to calculate the item difficulty indices of the test which is based on the proportion of the total number of examinees who got a correct answer on the given item and the total number of examinees. The closer the value of the index of difficulty of the given item to 1, the easier is the item and the closer it is to 0 the item will be very difficult. An index of 0.5 means that the item is average in its difficulty. We will also categorize the different intervals so that item difficulty indices can be given a verbal description. Moreover, in this study for reasons of simplicity and completeness of the estimation of item difficulty parameter for each of the 15 items, we assume that the item response theory's Rasch model fits the data, that is in every responses of examinees on each item fits in the model. Although, we will check the goodness of fit of the items to the model by statistical means as done by [43] in the process. For the sake of comparison, we do not discard items in the estimation that do not fit the given item response model, that is we need the complete 15 item difficulty estimates so we can compare it to the values in the classical test theory.

IV. Results and Discussions

This section will present three results. First is the presentation and discussions of the information characteristic curve (ICCs) of the fifteen items under the item response theory methods. Second and third is the simultaneous tabular

presentation of the calculations and analysis of the results for the estimation of the item difficulty under the item response theory (IRT) and the classical test theory (CTT) methods.

A. The Information Characteristics Curves of the Items

Fig. 2 are the information characteristic curves of the fifteen items, the ICCs of the fifteen items above can be converted into an Item Characteristic Curves (ICC) which are graphical functions that shows the examinees proficiency or ability as a function of the likelihood or chances of answering correctly the item. We can see through the curve that most of the items (9 out of 15 items) are difficult because higher abilities are needed to get higher probabilities of getting correct answer. We can see in the figure that items 8 and 12 are the very difficult items. To further support these observations, we will calculate mathematically using the Rasch model under the item response theory methods the values of the estimates of the difficulty of the fifteen items. Then we will incorporate the results of fifteen difficulty indices of the items under the classical test theory.

B. Fit of the Item to the Rasch Model

Checking the fit of the data to the Rasch model, the results show that some items are a "misfit", a terminology in modeling for those items that do not fit the model. As we mentioned above those items that do not fit in the Rasch model are supposedly discarded but for the purpose of comparison with difficulty indices under the classical test theory we will retain it. Table II shows those items that fit and also do not fit the Rasch model. To test the fit of the data responses of the item to the model, we use the Chi-square test. If the p-value of that item is less than 0.05 or 5%, we do not reject the hypothesis that the item fits the model. Based on the results in Table II, the following items fit the model in particular items 2, 3, 7, 10, 11, and 12. On the other hand, items 1, 4, 5, 6, 8, 9, 13, 14, and 15 in the test did not fit the model. Items that misfit the Rasch model means that the Rasch model is not a good model for these items, hence for model fitting purposes, another type of item response theory should be considered as a recommendation.

C. Comparison of Item Difficulty Estimates Between IRT and CTT

The item difficulty estimates of the data using the IRT and CTT methodology are presented in Table III.

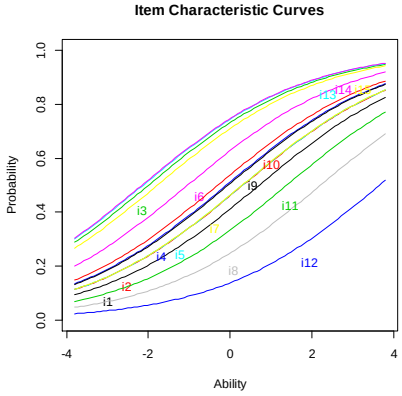

Fig. 2. ICCs of the 15 items using IRT Rasch Model.

TABLE. II. FITTING THE DATA USING RM (MODEL FIT)

Items	Test	
	X^2	p-value
1	18.49	0.01*
2	5.66	0.46
3	11.51	0.07
4	14.41	0.03*
5	22.23	<0.01*
6	24.78	<0.01*
7	9.97	0.13
8	23.08	<0.01*
9	15.16	0.02*
10	11.84	0.07
11	7.70	0.26
12	7.49	0.28
13	14.01	0.03*
14	34.71	<0.01*
15	22.62	<0.01*

Legend: * Significant at 5% (using chi-square test)

TABLE. III. ITEM DIFFICULTY ESTIMATES/ INDICES

Items	β (IRT)	description[a]	β (CTT)	description[b]
1	0.44	D	0.40	A
2	0.21	D	0.45	A
3	-1.01	VE	0.71	E
4	0.01	A	0.50	A
5	0.23	D	0.45	A
6	-0.59	E	0.62	A
7	0.21	D	0.45	A
8	1.23	VD	0.25	D
9	0.03	A	0.49	A
10	-0.10	E	0.52	A
11	0.75	D	0.34	D
12	1.91	VD	0.16	D
13	-1.09	VE	0.72	E
14	-1.20	VE	0.74	E
15	-0.96	VE	0.70	E

Legend: β=item difficulty values IRT $\in$ [-3,3], CTT $\in$ [0,1]
[a]Description: VE = Very Easy, E = Easy, A = Average, D = Difficult, VD = Very Difficult
[b]Description: VD=0-0.125, D=0.126-0.375, A=0.376-0.625 E=0.626-0.875, VE=0.876-1.0

Includes corresponding verbal descriptions of all the items. Discussing the item difficulty estimates of the fifteen items that was included in the test under the IRT framework that employed the Rasch model, results showed that the level of difficulties, the test in particular items 1, 2, 5, 7, and 11 are difficult items because they are above 0. Note that an item whose difficulty is zero is considered an average item. Items 8

and 12 can be considered very difficult items because they are almost near at the upper right extreme.

Further, items 4 and 9 are items on an average difficulty. On the other hand, items 6 and 10 can be considered easy items while items 3, 13, 14, and 15 are very easy items because they are in extreme left near the value -3 considered the easiest item. In the case with CTT methods, results of the analysis show that items 8,11 and 12 are difficult items and items 1, 2, 4, 5, 6, 9 and 10 are items with average difficulty while the rest of the items, items 3, 13, 14 and 15 are easy items.

For the point of comparisons in accordance to the item difficulty estimates calculated under the two methodologies in particular, the items verbal descriptions, with regards to classical test theory (CTT), results revealed that there were no very difficult items in the test. Examination and further there were also no very easy items. This can be explained maybe due to the choice of the categorized interval from 0 to 1. Three items out of the total fifteen items namely items 4, 9 and 11 showed similar descriptions in their item difficulty for both item response (IRT) and classical test theory (CTT) methodologies. However, there are also some items that do not have the same common description (about 9 out of 15 items or 60% of the items). These items are items 1, 2, 3, 5, 6, 7, 8, 10 and 12. These results are expected since the two methodologies have different assumptions in their formulations.

As we mentioned above, the assignment of the degree of difficulty depends on the kind of interval that was made. As we observed, some intervals are narrow and some are wide. In the case of item 1, it is difficult under the IRT formulation but is an average item in CTT. It is also the same result with items 2, 5, 7 while item 3 is very easy in IRT but is an easy item in CTT. Items 6 and 10 are both easy in IRT but were average items in CTT while items 8 and 12 are very difficult items in IRT but only difficult items in CTT and lastly items 13,14,15 are very easy items in IRT but are easy items in CTT. A study by [44] compared CTT and IRT for the examinee change assessment. According to them a lot of investigators were eager to know of how IRT can be used in greater advantage as compared to CTT in change assessment but available results showed that they did not differ when compared based the examinee change assessment. However, when compared in term of their type 1 errors and detection percentages, their results showed that IRT is better than CTT in the examinee ‘s change detection with the condition that the test must consists twenty (20) items or more. For shorter tests, however they further mentioned that CTT has the advantage of correctly knowing change in the examinees. In our study, however there was also some variations in the results between IRT and CTT among the item difficulties when they were compared but the objective of this study was achieved. The two free computer software programs the “ltm” and “CTT” were very useful in doing the statistical analysis using the R software for the item test calibrations.

V. Conclusions and Recommendations

This paper demonstrated the usefulness of the free computer software programs, the “ltm” and “CTT” in the R software environment for the calibration of the item difficulty parameter estimates/indices of the multiple-choice test examination using both the item response theory (IRT) and classical test theory (CTT) methodologies. We also demonstrated the usefulness of the Rasch model, an IRT probabilistic logistic model used to estimate the values of the item difficulty parameters of the test examination which were compared to the estimated values under the CTT method. Further, we also demonstrated that it was possible to plot the item characteristics of different items, so the proficiency or ability of the examinee can be estimated so that we will be able to know the higher chance of getting the item correctly. The Item characteristic curves (ICC) also gave us a glimpse of the difficulty characteristic of the item. The study also found some differences and similarities in the interpretation with the labeling of the item difficulty in the form of a verbal description for the items. The study concluded that these differences are due to the assumptions of the different methods in the item analysis.

The study further concluded that for the possibility of convenience for teachers in all levels and test constructors, they can do an item analysis for their test electronically using either ltm or CTT software packages in R for free in which first, they can do item calibration and assigned description for the item level of difficulty or indices and second, for the purpose of item banking especially in the test construction where items are stored in a database and labeled with their corresponding item level difficulty or indices.

This study further recommends that other item characteristics namely, the item discrimination and the item guessing parameters shall also be investigated to complete the test item analysis using both the classical test theory functions and other appropriate item response theory models that involves item calibration for discrimination and guessing.

Acknowledgment

Our heartfelt thanks to all the students who participated and have tried their best to answer the different items in the test. The first author also expressed his gratitude to his employer, the University of Southern Mindanao and his current President Dr. Francisco Gil N. Garcia for the encouragement and support.

References

[1] D. Rizopoulos, “ltm: An R Package for latent variable modeling and item response analysis,” Journal of Statistical Software, vol.17, no.5, 2006.

[2] J.T. Willse, Classical Test Theory Functions, 2018. https://cran.rproject.org/web/packages/CTT.

[3] G. Rasch, Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

[4] J.L.Pimentel, Item Response Theory Modeling with Nonignorable Missing Data. University of Twente, The Netherlands ISBN:90-365-2295-1, 2005.

[5] R.K. Hambleton, H. Swaminathan and H.J. Rogers, Fundamentals of item response theory, Newbury Park, CA: Sage, 1991.

[6] F.B. Baker, Item response theory: Parameter estimation techniques. New York, NJ: Dekker, 1992.

[7] F. Samejima, “Estimation of latent proficiency using a pattern of graded scores,” Psychometrika, Monograph Supplement, no. 17, 1969.

[8] R.D. Bock, "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," Psychometrika, vol. 37, pp. 29 – 51, 1972.

[9] F.M.Lord, Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

[10] G.N.Masters, "A Rasch model for partial credit scoring," Psychometrika, Vol. 47, pp.149 – 174,1982.

[11] R.Codey and J.K. Smith, Test Scoring and Analysis Using SAS. SAS Institute Inc., Cary, North Carolina, USA, 2014.

[12] J.S.Yang, X. Zheng, "Item Response Data analysis using Stata Item Theory Package," Journal of Educational and Behavioral Statistics, vol. 43, no. 1, pp. 116–129, 2018. Available; DOI: 10.3102/1076998617749186 © 2017AERA.

[13] P.J. Pascale and J.S. Pascale, "Item Analysis Using the Statistical Package for the Social Sciences (SPSS),"Educational and Psychological Measurement, vol.40 no.1, pp.163-164,1980.

[14] L.K. Muthen and B.O. Muthen, MPLUS: The comprehensive modeling program for applied researcher, users guide. Los Angeles, CA: Muthen & Muthen, 1998.

[15] M.F. Zimowski, E. Muraki, R.J. Mislevy and R.D. Bock, Bilog-MG. Lincolnwood, IL, Scientic Software International Inc.,2002.

[16] M.L.Wu, R.J. Adams, and M.R. Wilson. ConQuest: Multi-Aspect Test Software, Camberwell: Australian Council for Educational Research, 1997.

[17] N.D. Verhelst, C.A.W. Glas and H.H.F.M.Verstralen, OPLM: computer program and manual, Arnhem: Cito, the National Institute for Educational Measurement, the Netherlands, 1995.

[18] R Core Team, R: A language and environment for statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[19] P. Mair and R. Hatzinger, "Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R," Journal of Statistical Software, vol. 20 no.9, 2007.

[20] R.P. Chalmers, "mirt: A multidimensional item response theory package for the R environment," Journal of Statistical Software, vol. 48 (6), 2012.

[21] G.J.A Fox, "Multilevel IRT modeling in practice with the package MLIRT,"Journal of statistical software, vol. 20, no.5, 2007.

[22] S. Sturtz, U. Ligges and A. Gelman, A."R2WinBUGS: A Package for Running WinBUGS from R," Journal of Statistical Software, vol.12, no.3 , pp.1–16, 2005.

[23] M.L.A.Villaruz, Test Item Calibration for Multiple Choice Test Using IRT, Unpublished Undergraduate Thesis. MSU-IIT, Iligan City, Philippines, 2015.

[24] I.W. Molenaar, "Estimation of item parameters", In G.H. Fischer,and I.W. Molenaar (Eds.), Rasch models: foundations, recent developments and applications, New York, NJ: Springer, 1995.

[25] M.G.H. Jansen, "A Model for the Latent Traits in Rasch's Speed Tests," Applied Psychological Measurement, vol. 27, no. 2. pp.128-151, 2003. DOI: 10.1177/0146621602250536.

[26] G.H. Fischer, "Derivations of the Rasch model". In G.H. Fischer & I.W. Molenaar (Eds.), Rasch models: foundations, recent developments and applications, pp.39-52, New York, NJ: Springer,1995.

[27] G.H.Fischer and I.W. Molenaar, Rasch models. Their foundation, recent developments and applications. New York, NJ: Springer. pp.44-49, pp. 219-224, 1995.

[28] G.J. Mellenbergh, "A Unidimensional Latent Trait Model for Continuous Item Responses," Multivariate Behavioral Research, vol. 2, no.3, pp. 223-236, 1994.

[29] A. Skrondal and S.Rabe-Hesketh, Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models, Chapman and Hall/CRC,2004.

[30] F.B.Baker, The Basics of Item Response Theory. Second Edition. 2001.

[31] S.E. Embretson and S.P. Reise,. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum. pp.210-218, 2000.

[32] R.K. Hambleton and H. Swaminatan, Item response theory: Principles and applications (2nd edition). Boston MA: Kluwer Academic Publishers, 1985.

[33] J. Fox, J.Pimentel and C.Glas, "Fixed Effects IRT Model", Behaviormetrika 33, pp. 27-42, 2006.

[34] A.E. Gelfand and A.F.M. Smith, "Sampling-based approaches to calculating marginal densities," Journal of the American Statistical Association, vol. 85, pp. 398-409, 1990.

[35] M..H. Chen, Q.M.Shao and J.G. Ibrahim, Monte Carlo methods in Bayesian computation. New-York: Springer-Verlag, 2000.

[36] B.D. Ripley, Stochastic simulation, New York: Wiley,1987.

[37] C.P. Robert and G. Casella, Monte Carlo statistical methods, New York, NY: Springer,1999.

[38] A. Gelman, J.B. Carlin, H.S. Stern and D.B.Rubin, Bayesian data analysis, London: Chapman and Hall, 1995.

[39] D.Spiegelhalter, A. Thomas, N.Best, and D. Lunn, WinBUGS User Manual, MRC Biostatistics Unit, Cambridge, 2003.

[40] M.R. Novick "The axioms and principal results of classical test theory," Journal of Mathematical Psychology, vol. ,3, no.1, pp. 1-18,1966.

[41] R. Traub, "Classical Test Theory in Historical Perspective," Educational Measurement: Issues and Practice, vol. 16 no.4, pp.8–14, 1997. doi:10.1111/j.1745-3992.1997. tb 00603.x.

[42] T.M. Bechger,G. Maris, H.F.M Verstralen, and A.A. Beguin, "Using Classical Test Theory in combination with Item Response Theory", Applied Psychological Measurement, vol. 27 No.5. pp. 319-334, 2003. doi: 10.1177/0146621603257518.

[43] R.M. Smith, "Theory and practice of fit". Rasch Measurement Transactions, vol.3 (4) pp. 78, 1990.

[44] R. Jabrayilov, W.H.M. Emons, and K. Sijtsma, "Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment,"Applied Psychological Measurement vol.40,no.8,pp.559-572, 2016.