IJERIP

ISSN: 3116-3475
International Journal
of Education, Research, and Innovation Perspectives
Volume 2 Issue 3 (March 2026)

Calibration of the Multiple-Choice Test in Genetics Using R Software

Ronna N. Anonas[1], Jhondel P. Baranggan[1], Lotis A. Baoc Daguisonan[1]
[1]*Mindanao State University*
anonasronna105@gmail.com

Date Submitted:
February 05, 2026

Date Accepted:
March 13, 2026

Date Published:
March 21, 2026

DOI:
10.5281/zenodo.19151184

ABSTRACT

The present research aims to calibrate the prelim multiple choice test in Genetics during the First Semester of 2016-2017 using ltm package in R Software. There was a total of 89 students enrolled in Biology 106 (Gensetics) of the Department of Biology, College of Natural Sciences and Mathematics which serves as the respondents of this study. The study made use of the descriptive research design to identify the test's item difficulty, test's item discrimination, test's item fitness, and students' ability and latent traits. An interview was also conducted among the respondents enrich and support the findings of the research. Findings showed that the prelim multiple choices in Genetics needed revision and replacement because of incorrect item keying as determined using the R Software. It was also found out that almost half of the students failed to pass the exam because they were confused in answering the test. Furthermore, the findings showed that the multiple item tests in Genetics are poorly constructed. In line with the research findings, the researchers highly recommended that teachers must construct a well-written test by applying the use of table of specification. They must also employ varied teaching strategies, techniques and models that would cater the learning needs of all their students. Furthermore, students must study their lesson and not to hesitate ask for clarification if they are confused with the lecture.

Keywords: *Item Response Theory, 2 PL (Parameter Logistics) Model, latent traits, R Software*

INTRODUCTION

Assessment plays a vital role to determine if the desired learning outcomes have been achieved. Over the century, different forms of assessment arise in the field of education. And today, the most demand in areas of education and psychology assessment is the item response theory (IRT). This theory assesses the appropriateness and effectiveness of the formulated test questionnaire. Aided by the IRT model, it helps evaluate student's subject matter proficiency and skill development. It measures the appropriateness of item keying, item difficulty and the latent traits.

With the help of improved mathematical models and computer technologies, new theories have been developed in the field of education and psychology assessment. Researchers in the field of educational assessment are continually developing new approaches to improve the efficiency of assessments. They were often concerned with the methodologies that can extract the most useful and accurate information from students' responses to test items (Wu and Adams, 2006).

Learning progressions are used to describe how students' understanding of a topic progress over time and to classify the progress of students into steps or levels. This study applies Item Response Theory (IRT) based methods to investigate how the design learning progression-based science assessment (Chen, 2012).

According to Geremew (2014), Item Response Theory (IRT) models are commonly used to model the intent associated with the items or survey. In education, testing is an inherent part of the curriculum as an assessment tool to evaluate student's subject matter proficiency and skill development. Apart from viewing the total score as an indicator of performance, one may wish to understand whether the testing instrument is adequately designed to measure particular aspects of knowledge and skills of the respondents.

IRT is a "modern" test theory utilizing a set proportion or mathematical models related to individual responses to items, providing a probabilistic way of linking observable data to theoretical constructs, with the ability to statistical adjust scores for properties test items such as difficulty, discriminating power and liability to guessing (Kō̈se, 2014).

In this sense, IRT has major benefits over Classical Test Theory (CTT). In the IRT framework, item characteristics are sample-independent and a person's latent scores are test-independent provided that the selected models fit the data well. Thus, scores that describe examinee's proficiency are not dependent on test difficulty. Their scores may be lower on more difficult test and higher on easier tests, but their ability scores remain constant over any test at the time of testing or surveying. IRT also permits calculation of the probability of a particular respondent selecting a category on a test item. Moreover, IRT can be used for scale refinement or development, as it is capable of the calculation of standard errors and therefore provides information on the quality of each item. This aids with making decisions in selecting items to exclude or include in a test or survey instrument. In addition, items are also selected based on their difficulty and discrimination indices, i.e., their capability of discriminating low and high trait groups (Le, 2013).

The original work on IRT began with tests of dichotomous items which had only two choices or multiple-choice items were scored as right and wrong. Item response theory (IRT) is a set of latent variable techniques especially designed to model the interaction between a subject's "ability" and the item level stimuli (difficulty, guessing, etc.). Moreover, the researchers investigated the proficiency, the ability and how difficult the multiple test questionnaire. Item response theory is different in this sense; it models the relationship between a respondent's trait level (ability, attitude) and the pattern of item responses. The estimation of individual latent traits differs even for two individuals with the same total scores (Le, Chalmers, and Liang, 2013).

In this study, the researchers conducted a survey to determine possible problems and difficulties for BS Biology major students which they have encountered all throughout their study so far. The results found out that students find difficulty in Genetics subject. Hence, this study was conceptualized to address the probable reasons behind.

Theoretical Framework

Item response theory is a measurement framework used in the design and analysis of educational and psychological assessments (achievement tests, rating scale, inventories, or other instruments) that measure mental traits. Item response theory, or IRT for short, is based on establishing a model that specifies the probability of observing each response option to an item as a function of the target trait being measured by the assessment, which is often a knowledge, skill, or ability (e.g., math ability). In testing situations where items are scored as correct or incorrect, IRT specifies the probability of a correct response to an item as a function of ability (Kolen and Brennan, 2004).

IRT gains popularity due to its advantage over the simpler measurement framework of Classical Test Theory (CTT). A primary advantage of IRT is that it offers a rigorous, yet flexible, framework for placing assessment of different items on a common scale. This holds substantial benefit when having to link the scores of multiple forms of an assessment onto a single reporting scale so that the scores have the same meaning across the different forms of the assessment. A related application of this advantage is computer adaptive testing, whereby each examinee is administered a set of items that is tailored to the examinee's level of ability, resulting in different examinees receiving different sets of items. Another advantage of IRT is its capability to specify reliability specific to each examinee. Whereas, reliability in CTT is summarized by a single index, i.e., applied equally to all examinees regardless of ability level. Item response theory has the flexibility to estimate reliability uniquely for each examinee. This information can be very useful when different individuals are administered different items (as in computer adaptive testing), or when building test forms with cut-scores or proficiency standards such that the forms can be built to maximize precision (minimize error) around those points on the scale (Dorans, Pomerich and Holland, 2007).

Conceptual Framework

This study focuses on the item discrimination like examinee's ability and item difficulty of Genetics course among BS-Biology, BSEd-Biology, BS-Zoology and BS- Nursing students who are currently enrolled in Genetics course in MSU-Main Campus, Marawi City. The study focuses on the pattern of responses rather than on composite or total variables and linear regression theory. The IRT framework emphasizes how responses can be thought of in probabilistic terms. In IRT, the item responses are considered the outcome (dependent) variables, and the examinee's ability and the item's characteristics are the latent (independent) variables.

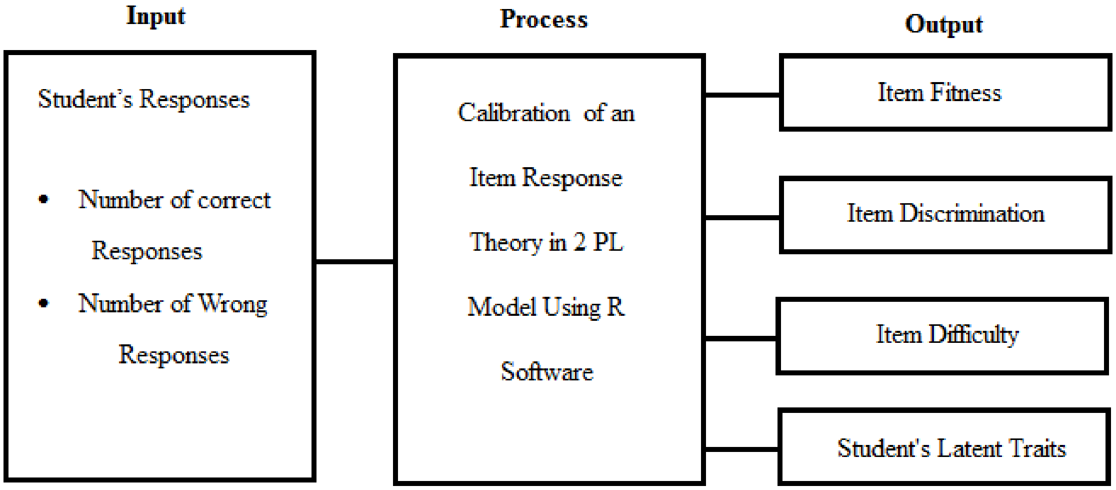

Input
Process
Output
Student's Responses
Number of correct Responses
Number of Wrong Responses
Calibration of an Item Response Theory in 2 PL Model Using R Software
Item Fitness
Item Discrimination
Item Difficulty
Student's Latent Traits

Figure 1.1**. *The Schematic Diagram of the Conceptual Framework of the Study***

Objectives

Generally, this study aims to calibrate the 30-item multiple choice test in Genetics course. The students under the Genetics course served as the respondents of this study and were composed of BS-Biology, BSEd-Biology, BS-Zoology and BS-Nursing who were enrolled in Genetics course in MSU-Main Campus, Marawi City during the First Semester, A.Y. 2016-2017. The researchers wished to answer the following questions:

1. What is the percentage of the correct and wrong answer of the responses of students in each item of the test?
2. What is the item fitness of the test items using 2 PL model?
3. What are the item difficulties of the multiple-choice test in Genetics using 2 PL Model?
4. What is the item discrimination of the multiple-choice test in Genetics using 2 PL model?
5. What are the latent traits of the test takers based on 2 PL model?

Significance of the Study

This study determines individual's responses to the items and its relations such as difficulty, discriminating power and liability to guessing of such items. It utilizes a theoretical based model known as IRT 2PL model and uses the ltm package in R software in calibrating the item. R software uses a range of types of data graphical manipulation, graphical presentation and statistical analysis. It has extensive and powerful graphics abilities that are tightly linked with its analytic abilities (Maindonald, 2013). This study was conducted to assess the performance of the students who were currently enrolled in Genetics course of Mindanao State University-Marawi Campus. This study is significant in the sense that it will help to resolve problems encountered by the students related not only to the course but also to test examinations. The study would be of great help to the following entities:

Students. They may become aware of how they should acknowledge their strengths and weaknesses and may give them an idea that makes them learn easily and could bring out the best in them. This study would also serve as a reflection on how students performed in Genetics, thus would enlighten them to strive more in obtaining a better performance.

Instructor. They may become aware and be guided on how to make appropriate test questionnaires and to become an effective teacher.

Parents. They may become aware of their children's performance, give time allocation for their children to study and fully support them so that their study becomes meaningful and have an essence all throughout.

Curriculum makers. They may be aware of making an appropriate curriculum that suits or in line with the capability of every student or learner.

Guidance Counsellors. They may become aware in advising or counselling the students to choose the appropriate course and always give an idea on how to pass the subject and give them pieces of advice to take studies seriously.

Future Researchers. This study will serve as basis for future studies which interested researchers may pursue. They will be able to confirm the findings on this study and even performed more complex and deeper research using the methods employed in this study.

Scope and Limitations

This study is confined to calibration of multiple choices in Genetics with only 30 items and will only focus on the three (3) variables. First, student's performance will be assessed whether the questionnaire could be easily answered by the students. Second, item fitness whether the questions are fitted to the 2 PL model or not. Third, item difficulty whether the item given by the Genetics Instructor is difficult or an easy one. Fourth, item discrimination of whether the test questionnaire given by the Instructor is discriminating or an ideal item. And finally, student's latent traits of whether the students can answer the difficult questions which would indicate that the students possess latent traits.

This study is achieved through the open source named R software which is an integrated collection of statistical and analysis tools that will give an accurate and precise result. The recently study conducted by Colvin (2015) that item parameter and ability estimates from R package ltm were estimated reasonably with accurate results similar to previous studies of established commercial software. And the interview questions support this study to enrich the results and findings. As a limitation, the researchers include all eighty-nine (89) respondents who were currently enrolled in Genetics course. This was conducted within the perimeter of MSU-Main Campus, Marawi City, for the period of one (1) whole semester, A.Y. 2016-2017. The study was not intended to create any conflicts among the students, instructors and the designed curriculum.

LITERATURE REVIEW

One of the most important improvements in the last century is Item Response Theory (IRT), also known as latent trait theory, in psychological measurement. IRT is a modern test theory which explains examinee's ability level by using responses to test items with strong assumptions. IRT is clearly a widely accepted tool in statewide assessment programs, with most states incorporating it in assessment practices (Ryan and Brockman, 2009).

According to Le (2013), IRT is an approach to modern educational and psychological measurement which addresses the measurement of a hypothetical latent construct such as ability or attitude. These latent traits cannot be measured directly on individuals and must be quantified via responses to items or questions in a test or survey. IRT methods are commonly used to obtain latent scores for individual respondents on qualities such as trait, ability, proficiency, or attitude in a test or survey.

Moreover, IRT is a psychometric tool to construct mathematical models using items on instruments, such as measuring mathematics ability using a multiple-choice assessment. The underlying premise of IRT is that every test taker has some level of knowledge, referred to as ability or proficiency, related to the test's content (de Ayala, 2009 and Kolen et al., 2011).

Besides, according to Tao (2008), Item Response Theory (IRT) is a contemporary measurement technique which has been used widely to model testing data and survey data. To apply IRT models, several assumptions have to be satisfied. Item Response Theory has experienced great popularity in practice because of its capability of addressing several problems encountered in Classical Test Theory or CTT. However, some strong assumptions have to be satisfied in order to generate valid results using IRT models.

Moreover, Yu (2013), provided that if the item is misfit, one reason is that the item is not well-written when low skilled students got the higher scores than high skilled students have given the right answer. By the same token, if a supposedly low skilled student answers many difficult items correctly in a block of questions; one explanation for this could be an instance of cheating.

In this sense, it is suggested to eliminate the item. Those items that misfit might consider rephrasing or deleting this item. The assumption can be tested through item analysis. Item-fit studies may assist in identifying faulty item construction, e.g., incorrect item keying (Mastura, 2016).

IRT models greatly facilitate test equating, permitting direct comparison of scores derived from different sets of items measuring the same construct (Embretson and Reise, 2000). These can support adaptive testing, which can shorten testing time while maintaining reliability by focusing testing on those items that target a given client's particular ability level (Bjorner, Chang, Thissen, and Reeve, 2007).

The assessment of data-model fit is important because the application of an IRT model can be justified only when data fit the model. The most general approach for assessing model-data fit of IRT models is to compare an observed score distribution with an expected score response distribution across discrete ability levels for each item (Yang, 2007).

IRT models in the dichotomous case is a wide range of classical item discrimination indices may suggest the need for the discriminating parameter in an IRT model, otherwise considerable information would be lost and model fit would be poorer. The level of difficulty of multiple-choose items provides an indication of the need of a "guessing parameter" in the IRT model. If items are easy, the guessing parameter may not be necessary (Zao, 2008).

For example, in unidimensional IRT models, the examinee's responses to items are assumed to be related to a general underlying dimension, representing proficiency, ability, achievement, preference, etc. Item response probabilities are assumed to be independent conditional on the latent variable (Thissen & Steinberg, 2009).

In lieu with this, multiple choice items are commonly scored as either right or wrong (dichotonomous response) in which models from item response theory or IRT seems especially useful. In IRT modeling, when model fit is good, it leads to item and examinee parameter in variance and other attractive features that are valued by test developers (Bastari, 2000).

According to Baker (2001), items with negative discrimination occur in two ways. First, the incorrect response to a two-choice item will always have a negative discrimination parameter if the correct response has a positive value. Second, sometimes the correct response to an item will yield a negative discrimination index. This tells you that something is wrong with the item: Either it is poorly written or there is some misinformation prevalent among the high-ability students. In any case, it is a warning that the item needs some attention".

Based on Foley (2010), item response theory (IRT) models have been growing in popularity in recent decades. IRT is now a well-known and accepted method that is widely used across a variety of assessment programs. These models provide a way to model the probability of giving a correct answer on an item based on the underlying ability of the examinee.

Yu (2006) defines item response theory as a collection of modeling techniques for the analysis of examinee responses. Many possible IRT models exist which are all based on the assumption that an examinee's probability of answering an item correctly depends on the examinee's latent traits and item characteristics. Item Response Theory (IRT) which is a contemporary measurement technique can be used to model testing data and survey data. Because of this, it has experienced great popularity in practice because of its capability of addressing several problems encountered in Classical Test Theory (CTT). However, some strong assumptions have to be satisfied in order to generate valid results using IRT models (Tao, 2008).

Additionally, the estimated item parameters are invariant with regard to who is sample from the population, the estimated proficiency level remains constant regardless of which items are administrated and also IRT can estimate examinee ability with more precision of measurement and less measurement errors (Lee, 2007). Also, An and Yung (2014) provided that item response theory (IRT) is concerned with accurate test scoring and development of test items. Test items are properly designed to measure various kinds of abilities such as math ability, traits like extroversion, or behavioral characteristics. Responses to test items can be binary (such as correct or incorrect responses in ability tests) or ordinal. Traditionally, IRT models have been used to analyze these types of data in psychological assessments and educational testing.

With the foregoing studies mentioned above, these will greatly help this research study to have a strong foundation based solely on these claims, evidences and results.

The following studies were conducted by other researchers and their studies are disclosed in this part since these are relevant to the present study.

Several studies have been conducted regarding Item Response Theory using the implementation of the open source statistical tool known as R Software. To name a few, Brzezińska (2016) conducted a study entitled "Polytomous Item Response Theory Models Using R" which provides a summary of some of the latest developments in polytomous item response theory (IRT) and helps realize that psychometric tools can now be used for theory testing in addition to the traditional role of improving construct measurement. Application of general polytomous item response theory models as well as practical considerations have been presented where the gpcm and grm package available in R were used that covers the most important features of the analysis for poltomous IRT models. The main functions of these packages have been presented and a comparison over the presented polytomous IRT models was discussed. Coefficients for several poltytomous IRT models were calculated and conducted in R from CRAN, which is a widely-used and well-known environment for statistical computing and graphics, as well as the information criteria for testing the goodness of fit. For model selection, the likelihood ratio test, AIC and BIC were applied to compare the fit of alternative models to these data. In testing the goodness of fit, the model with the lowest information criterion indicating the best fitting model was chosen.

Moreover, Sahin and Colvin (2015), on their study entitled "Evaluation of R Package ltm with IRT Dichotomous evaluated the accuracy of the item parameter and ability estimates generated by the open-source R package ltm. In this simulation study, item and ability estimates were compared to the true parameters under six conditions that differed in the numbers of items and examinees. After looking at the resulting bias, mean absolute deviation, and root mean square error, it was found out that item parameter and ability estimates from ltm were estimated reasonably accurately with results similar to previous studies of established commercial software.

In addition, Molenaar, Tuerlinckx, and van der Maas (2015) conducted a study entitled "Fitting Diffusion Item Response Theory Models for Responses and Response Times Using the R-Package diffIRT" where diffIRT, an R-package that can be used to fit item response theory models that are based on a diffusion process was presented. Parameter estimation and model fit assessment was discussed; viability of the package in a simulation study, and the use of the package with two datasets pertaining to extraversion and mental rotation was shown and illustrated. It was also illustrated how the package can be used to fit the traditional diffusion model (as it has been originally developed in experimental psychology) to data.

Several studies also found in various literatures wherein ITR was employed using different statistical tool or software other than the R Software. The study of Kō̄se (2014) was conducted in Abant Izzet Baysal University which was based on a simulated data set. Two major steps in the simulation are employed and these are data generation and data calibration. A computer program WINGEN was used to simulate the item response data. Responses of 1000 examinees to a dichotomously scoring 20 item tests were simulated with 25 replications. And based on this study, 2-PL Model fits significantly better than the 3-PL models.

Moreover, Johns, Mahaven and Woolf (2006) on their study entitled "Estimating student proficiency using an item response theory model" used the dichotomous IRT models to estimate a student's proficiency in answering multiple choice questions. The purpose of these models is to probabilistically explain an examinee's responses to test items via a mathematical function based on his/her ability. Multiple experiments were conducted to find the most appropriate modeling assumptions given our data set. Data exist for 401 high school students and 70 multiple choice problems. And the best results, which predicted a student's response with 72% accuracy, were achieved using the 2- parameter logistic equation.

Haberman and Sinharay (2013) conducted a study based on the assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. Item parameters for the test were estimated using the computer program PARSCALE. The examinees are scored on 3-category polytomous items in addition to 39 dichotomous items in the part considered here. The 2 PL model is operationally used for the test for the dichotomous items. Furthermore, 2 parameter logistic models demonstrated that it can be utilized for both descriptive and explanatory purposes to analyze test data.

In a different study conducted by Atar and Aktan (2013) entitled "Person Explanatory Item Response Theory Analysis: Latent Regression Two Parameter Logistic Model" purposely illustrated the application of latent regression two-parameter logistic (2-PL) model as explanatory item response model (EIRM) on TIMSS 2007 Science data for Turkish students. The goal of this study is to investigate the effects of the person properties on students' achievement and data were then analyzed under latent regression 2-PL model by PROC NLMIXED program.

In addition, the study of Edelen and Reeve (2007) on applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement and the Methods utilizes example data coming from 6,504 adolescent respondents in the National Longitudinal Study of Adolescent Health public use data set who completed to the 19- item Feelings Scale for depression. The results showed that 19 items varied in their discrimination (slope parameter range: .86–2.66), and item location parameters reflected a considerable range of depression (–.72–3.39). Based on the conclusions of this study, when used appropriately, IRT can be a powerful tool for questionnaire development, evaluation, and refinement, resulting in precise, valid, and relatively brief instruments that minimize response burden.

As per the study of Gonzalez (2008) entitled "Principles of multiple matrix booklet designs and parameter recovery in largescale", it utilized simulated responses to the items under four different conditions: a) All examinees were administered the 56 items; b) Items were randomly assigned to one of seven blocks, labeled A, B, C, D, E, F, and G. Based on the form booklet design, every examinee was administered three blocks (24 items in total) in the assessment pool. Based on the simulated data with specific characteristics, and therefore in some senses are relatively clean compared to what can be expected from real data stemming from operational data collection. It is likely, therefore, that different results would be obtained if real assessment data were used. For example, in the simulations, a well-targeted set of items, with difficulties well within the range of the abilities of the respondents was used. These items, moreover, were administered to random samples of respondents, with abilities normally distributed around the area where items provide more information.

Additionally, Thorpe and Favia, (2012) conducted a study entitled "Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications" incorporated changes in the probability distribution shape among the grouped items by using a two-parameter logistic (2PL) model, with a discrimination parameter that measures the steepness of the Item Characteristic Curve (ICC). It was shown that participants with higher ability levels have more than a 50% chance of responding with the correct answer. A high value for discrimination parameter corresponds to high item discrimination, which indicates a strong division between higher and lower-achieving participants for the item. They further noted that the items themselves should have enough common variance to give reasonably 21 unbiased estimates of item difficulty. In general, an unbiased analysis for dichotomously-scored items (those with two possible response codes, e.g., 0 or 1) may have as few as 100 participants, whereas 5-point response formats require a sample size of at least 500 participants.

Furthermore, Hula, Feradiotis, and Martin (2012) conducted a study to identify the most appropriate item response theory (IRT) measurement model for aphasia tests requiring 2-choice responses. The empirical data that were used to develop the simulation parameter was collected from a sample of 70 persons with aphasia who were given the PPT for research purposes. The descriptive statistics for the empirical sample of participant the age, months post onset, and education are given sample for large sample of N=111 that previous reported and the current sample of n= 70 was drawn. The study used WinGen 3.01 to generate distributions of item difficulty, item discrimination, and person ability that served as known true values. These results demonstrated the expected finding that sample size.

From the related studies presented above, item response theory is a modern test theory which explains examinee's ability level by using responses to test items. It focuses on how specific test items function in assessing constructs. The basic concept of item response theory rests upon the individual items of a test rather that upon some aggregate of the item responses such as a test score. It is also concerned with accurate test scoring and development of test items.

The evidences and claims presented in the related studies would be of great help to support the findings and results for the next chapter and it will give reliability and validity of our studies. This further proved that the IRT 2 PL Model is an effective tool in assessing the test item and student's ability and traits in answering certain set of questionnaires.

METHODOLOGY

Research Design

The study utilized the descriptive design which aims in finding out the results of item discrimination such as the student's ability and item difficulty. Descriptive design is usually concerned with describing the population with respect to important variables. The method used in this is a qualitative method to generate data and item calibration with adequate and accurate interpretation of the findings. An interview method was used with the aid of interview-guide questions to conduct the study. The researchers used in- depth interview techniques to gather varied and divergent answers from the research questions.

Research Locale

This study was conducted at Mindanao State University (MSU) in Marawi City, which is situated on hills overlooking Lanao Lake, the second largest lake in the Philippines. It covers one (1) thousands of hectares located about four (4) kilometers from downtown Marawi, the Islamic City of the South, right in the province of Lanao del Sur.

Mindanao State University was established on September 1, 1961, through RA 1387, as amended, by the late Senator Domocao A. Alonto, as one of the government's responses to the so-called "Mindanao Problem" (Student Handbook, 2008 Edition).

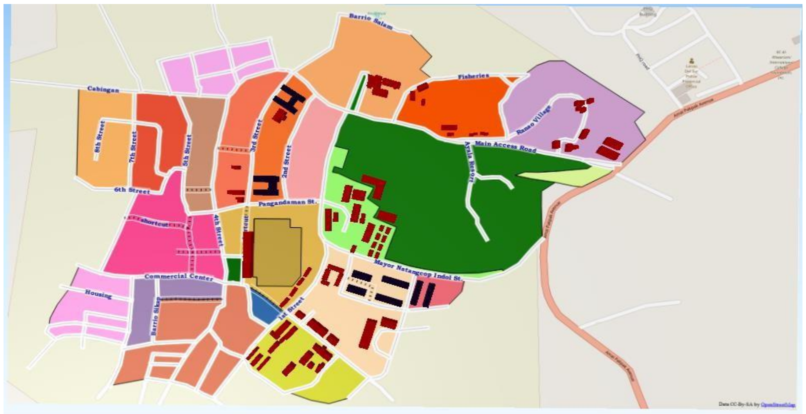


Figure 3.1 ***Map of Mindanao State University, Marawi City***

Research Respondents

The respondents of the study were composed of all eighty-nine (89) students of BS-Biology, BSEd-Biology, BS-Zoology and BS-Nursing students in Biology Department of College of Natural Sciences and Mathematics, who are currently enrolled in Genetics in Mindanao State University, Main-Campus Marawi

City. The participants of the interview were coded as B1, B2, B3, B4, B5, B6, B7 and B8 respectively in each theme.

Research Instrument

There are many software packages commercially available for psychometric analyses and specifically for the use with item response theory (IRT) models. The software packages available in the open-source R software program are gaining attention, which even led to a special volume on use of R packages in psychometric in the Journal of Statistical Software (JSS, 2007). Despite the abundant availability of such packages, a critical aspect of these software packages is the accuracy of estimations. Simulation studies are effective for evaluating the accuracy of estimations since such studies allow researchers to compare estimates with the true values.

Data Gathering Procedures

The researchers asked permission from the authority of the Genetics Instructor for the conduct of this study. After obtaining the permission to conduct, then it proceeded to data generation. Firstly, the researchers gathered the examinee's results of the Prelim Exam from the Genetics Instructor. Secondly, the researchers tabulated the results of each item whether correct or incorrect with a binary coding of 1 and 0 respectively. Thirdly, input the binary numbers 1 for correct item and 0 for incorrect item in to the R software ltm package. Lastly, the program analyzed the results of the item fitness, item difficulty, item discrimination, examinee's ability and to determine the latent traits.

The researchers selected the participants to be interviewed, two (2) from BS- Biology, two (2) from BSEd-Biology, two (2) from BS-Nursing and two (2) from BS- Zoology. Upon consent with the participants, the researchers conducted the interview with them.

Statistical Tools Used

R software is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. R software can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks. The most convenient way to use R is at a graphics workstation running a windowing system. This guide is aimed at users who have this facility. In particular, occasionally this refers to the use of R on an X window system although the vast bulk of what is said applies generally to any implementation of the R environment (Smith and Venables, 2016).

Data Analysis Procedure

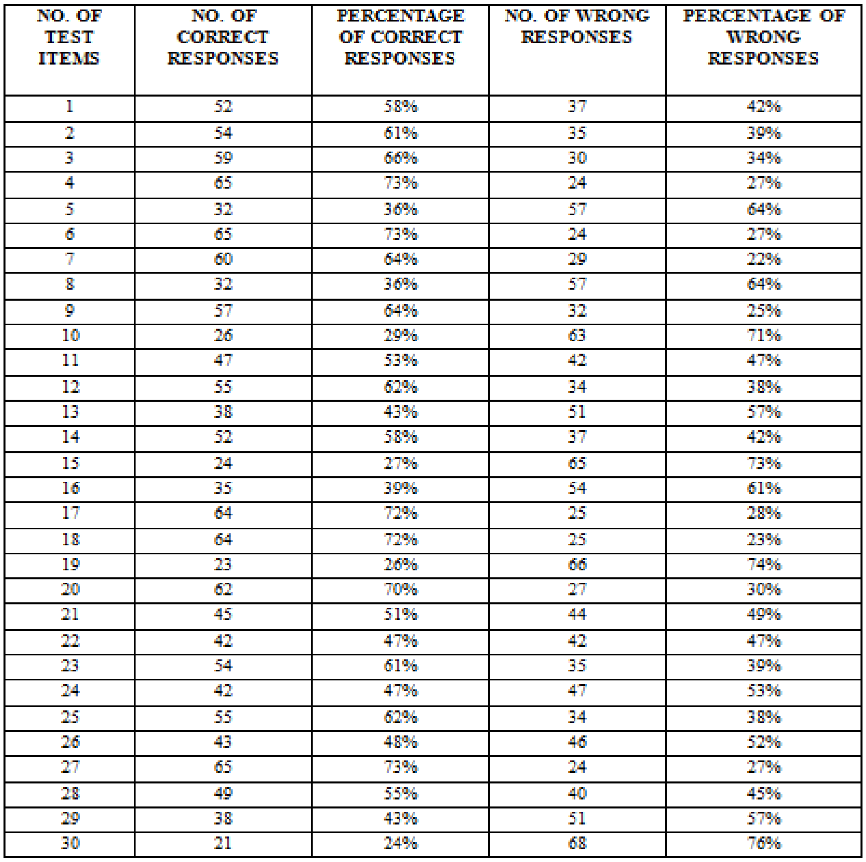
The ability of an examinee can be considered on a scale from negative to positive infinity, in practice, ability is often quantified between -3 and +3. Since ability and item difficulty are on the same scale, item difficulty can take both negative and positive values. In practice, *b* values often range between -2 and 2 where smaller values indicate easier items. Although, theoretically, the discrimination parameter *a* can range from negative to positive infinity, in practice only items with positive *values* are used. A negatively differentiating item means that for an examinee with lower ability there is a higher probability of providing a correct response to that item; therefore, items with negative *values* are considered problematic and eliminated from tests. Moreover, it is also not usual to obtain *a* values great than two. Therefore, in practice *a* values range between 0 and 2. Third, *c* represents the probability of an examinee with infinitely low ability correctly answering the item. Since *c* represents a probability, it ranges from 0 to 1, where larger *c* values indicate not well-written items (Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991).

The interviewed responses of the participants were quoted in the discussion of the findings of this study and they were coded as B1, B2, B3, B4, B5, B6, B7 and B8 in each theme.

RESULTS AND DISCUSSION

I. Percentage of the correct and wrong answer of the responses of students in each item of the test

Table 4.1 ***Percentage of the correct and wrong answer of the responses of students***



NO. OF TEST ITEMS	NO. OF CORRECT RESPONSES	PERCENTAGE OF CORRECT RESPONSES	NO. OF WRONG RESPONSES	PERCENTAGE OF WRONG RESPONSES
1	52	58%	37	42%
2	54	61%	35	39%
3	59	66%	30	34%
4	65	73%	24	27%
5	32	36%	57	64%
6	65	73%	24	27%
7	60	64%	29	22%
8	32	36%	57	64%
9	57	64%	32	25%
10	26	29%	63	71%
11	47	53%	42	47%
12	55	62%	34	38%
13	38	43%	51	57%
14	52	58%	37	42%
15	24	27%	65	73%
16	35	39%	54	61%
17	64	72%	25	28%
18	64	72%	25	23%
19	23	26%	66	74%
20	62	70%	27	30%
21	45	51%	44	49%
22	42	47%	42	47%
23	54	61%	35	39%
24	42	47%	47	53%
25	55	62%	34	38%
26	43	48%	46	52%
27	65	73%	24	27%
28	49	55%	40	45%
29	38	43%	51	57%
30	21	24%	68	76%

Figure 4.1 ***Percentage of Correct Answer of Responses of the Students***

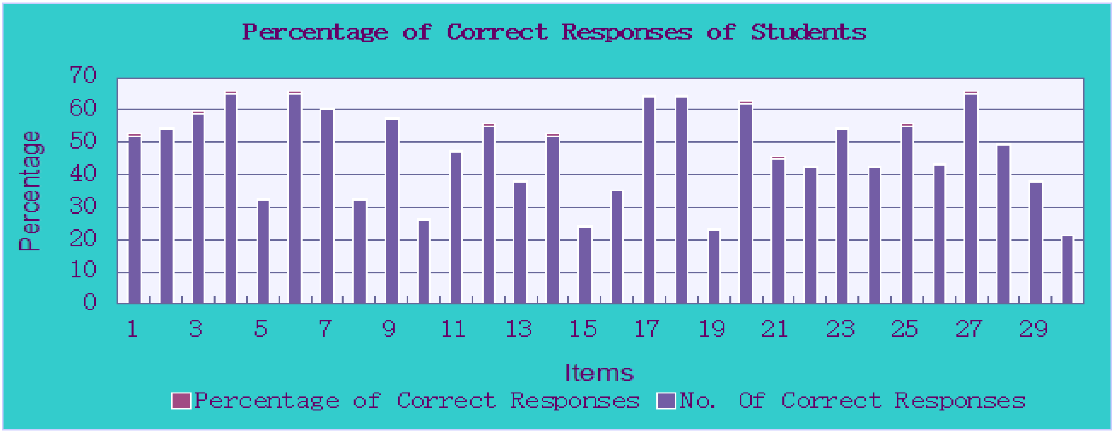


Figure 4.2 ***Percentage of Wrong Answer of Responses of the Students***

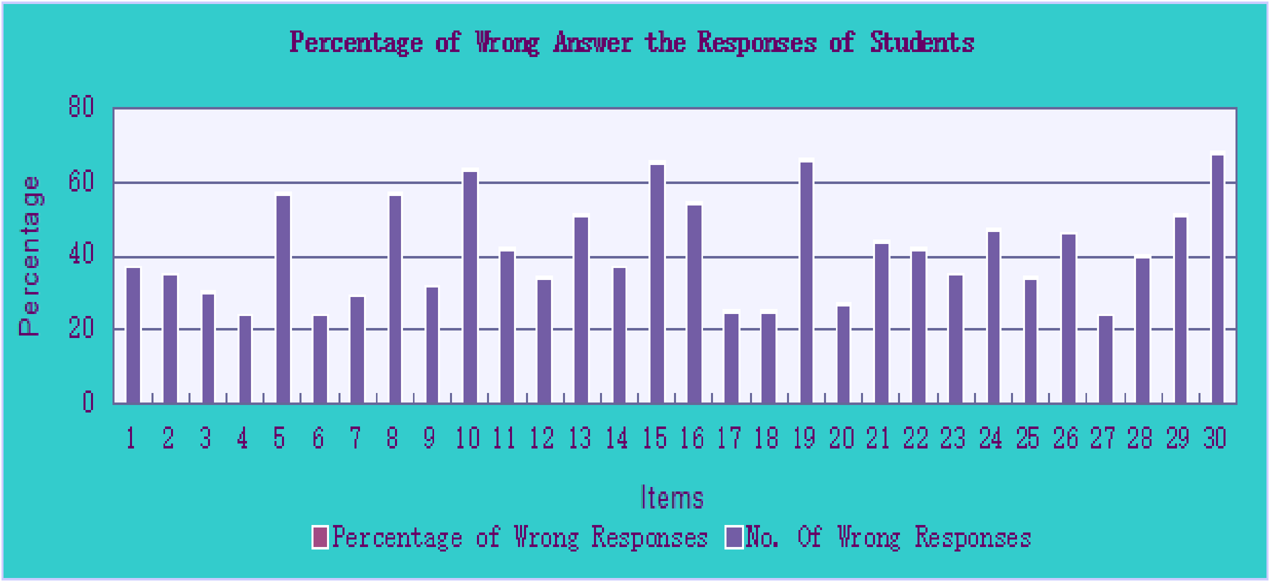


Table 4.1 shows the percentage distribution of the correct and wrong answer of the responses of students on each test item of the biology 106 (Genetics) examination results. Figure 4.1 shows the graphical representation of the percentage of correct responses of students in each item of the test. The data indicates that there are eighteen (18) items namely 1, 2, 3, 4, 6, 7, 9, 11, 12, 14, 17, 18, 20, 21, 23, 25, 27 and 28 where the students got the correct responses. While Figure 4.2 shows the graphical display of the percentage of wrong responses of students in each item and the data indicates that there are twelve (12) items namely 5, 8, 10, 13, 15, 16, 19, 24, 26, 29 and 30 where the students got the wrong responses. The finding shows that more than half of the examinee passed the said exam and almost half failed. This could clearly show and indicates that perhaps Genetics subject is one of the most difficult subjects in Biology Courses because it does not merely delve with theories but also involves mathematical computations and analyses.

This was confirmed with the follow-up interview to those selected students in Biology 106 after taking their exam. Some students revealed that Genetics subject is a difficult subject to them. They reported that:

B1: Mahirap siya at ito'ng subject na nahirapan ako sa lahat ng biology.

(It's difficult and I find this [Genetics] subject as the most difficult in all of biology courses.)

B2: Genetics is quite a difficult subject and it studies about genes.

B4: For me it's hard.

B5: Difficult!

B8: Genetics subject is a difficult subject.

In relations with this, another interview question that talks about what particular topic in multiple choice test did they find it difficult and why they considered this as a difficult one. The student's responses were:

B1: Siguro dyan sa punnett square because siguro hindi pa ako masyadong marunong mag punnett square.
(Maybe the punnett square because somehow I'm not that good in solving the punnett square.)

B2: The Mendelian genetics, difficult siya kung maraming gametes ang involve. Tapos nakakalito na pag solve.
(The Mendelian genetics; it is difficult if more gametes are involved and it is confusing to solve.)

B3: Yung mahirap lang talaga namin ay yung problem solving.
(The most difficult part is the problem solving.)
B4: Mahirap yung solving.
(Problem solving is difficult.)
B5: Probability, the most difficult.

B6: Probably the mioses and mitosis. I'm having hard time there.
B7: Blood typing. I find it hard remembering the said topic.
B8: Yung about sa karyotyping.
(It's on karyotyping.)

In lieu with this, the researchers found out that the Genetics subject has wide scope where it includes quantitative analysis and principles but require a deep working understanding. Which was affirmed *according to (Griffiths, n.d.), Genetics demands higher levels in Bloom's taxonomy.*

The researchers could tell that students find difficulty maybe because they don't find the Genetics as an interesting subject and in addition, it involves principles as well as numbers which in return, students' manifest inability to perform simple quantitative analysis. And besides, Genetics requires higher thinking ability.

II. Item fitness of the test items using the 2 PL model

Table 4.2 ***The Item Fitness of the 30 items using 2PL Model***

Test Item Number	P (theta)	Quantitative Description
1	0.1517	fit
2	0.4503	fit
3	0.8951	fit
4	0.4117	fit
5	0.2045	fit
6	0.6145	fit
7	0.2986	fit
8	0.1239	fit
9	0.348	fit
10	**0.034**	**misfit**
11	0.1942	fit
12	0.6831	fit
13	0.1447	fit
14	0.5138	fit
15	0.8516	fit
16	0.7611	fit
17	0.3214	fit
18	0.4163	fit
19	0.6067	fit
20	0.6491	fit
21	0.7936	fit
22	0.3761	fit
23	0.5803	fit
24	0.2893	fit
25	0.6768	fit
26	**0.0036**	**misfit**
27	0.3341	fit

28	0.1045	fit
29	0.1236	fit
30	**0.0455**	**misfit**

Based on Table 4.2, it shows the item fitness of the 30-item test which were calibrated whether a fit or misfit item depending upon the parameter P (theta). From these data, there were misfitted 3 items to 2PL model and these are Items 10, 26 and 30. These items do not fit to the 2 PL model and this can be attributed to two things. First, the wrong item characteristic curve model may have been employed. Second, the values of the observed proportions of correct response are widely scattered that a good fit, regardless of model, cannot be obtained. *In this sense, it is suggested to eliminate the item. Those items that misfit might consider rephrasing or deleting this item. The assumption can be tested through item analysis. Item-fit studies may assist in identifying faulty item construction (e.g., incorrect item keying)* (Mastura, 2016). There were 27 items which fitted to the 2-parameter logistic model these items namely; 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28 and 29.

The assessment of data-model fit is important because the application of an IRT model can be justified only when data fits the model. *The most general approach for assessing model-data fit of IRT models is to compare an observed score distribution with an expected score response distribution across discrete ability levels for each item (Yang, 2007).*

In lieu with this, another interview question that talks about whether there any chance/s that the students find it difficult to comprehend the given question. Some students answered that;

B9: In some instances, I find it difficult in comprehending the given questions given by the instructor.

B10: Nakakalito ang questions ni Ma'am.

(Ma'am questions were confusing.)

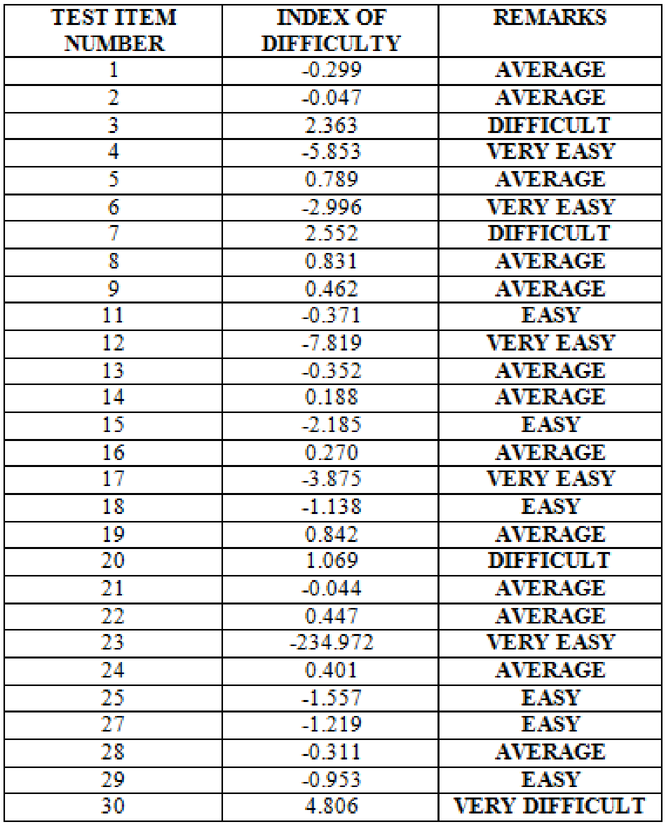
B11: Hindi ko ma gets ang exact problem na gustong ipa answer ni Ma'am.

(I can't exactly grasp the problem of what Ma'am wanted us to answer.)

The researchers could perhaps tell that the software and the students both agreed that some of the formulated questions in multiple choice test given by the Genetics Instructor were faulty items, inappropriate questions or the questions were grammatically incorrect. This would then lead to confusion for the takers in answering the given questions.

III. Item difficulty of the multiple-choice test in Genetics using 2 PL Model

Table 4.3 ***Item Difficulty in 2PL Model***



TEST ITEM NUMBER	INDEX OF DIFFICULTY	REMARKS
1	-0.299	AVERAGE
2	-0.047	AVERAGE
3	2.363	DIFFICULT
4	-5.853	VERY EASY
5	0.789	AVERAGE
6	-2.996	VERY EASY
7	2.552	DIFFICULT
8	0.831	AVERAGE
9	0.462	AVERAGE
11	-0.371	EASY
12	-7.819	VERY EASY
13	-0.352	AVERAGE
14	0.188	AVERAGE
15	-2.185	EASY
16	0.270	AVERAGE
17	-3.875	VERY EASY
18	-1.138	EASY
19	0.842	AVERAGE
20	1.069	DIFFICULT
21	-0.044	AVERAGE
22	0.447	AVERAGE
23	-234.972	VERY EASY
24	0.401	AVERAGE
25	-1.557	EASY
27	-1.219	EASY
28	-0.311	AVERAGE
29	-0.953	EASY
30	4.806	VERY DIFFICULT

Legend:

< - 3.0	(Very Easy)
-2.9 - -1.0	(Easy)
-0.9 – 1.0	(Average)
1.1 – 2.9	(Difficulty)
> 3.0	(Very Difficult)

Based on the tabulation presented in Table 4.3, results showed that there were five (5) items which are considered easy items and these are Items 4, 6, 12, 17, and 23. In addition, there are six (6) easy items and these are Items 11, 15, 18, 25, 27, and 29. However, there were thirteen (13) items which are considered average item; these are Items 1, 2, 5, 8, 9, 13, 14, 16, 19, 21, 22, 24, and 28. There are three (3) items that were difficult items and these are Items 3, 7, and 20 and lastly, one (1) item which is very difficult and this is item number 30. In addition, individual item graph can be seen in Item Characteristic Curve (ICC) shown in Figure 4.3. Out of thirty items in the multiple-choice test, there were only thirteen (13) items that were considered good items which lie in the difficulty index as an average. And seventeen (17) items which need revision, rejection or be eliminated and these were items that were considered not good items which lie and distributed as either very easy, easy, difficult or very difficult. The result shows that more than half of the items needed much attention for revision or modification. Additionally, the interviewees were also asked what they think about the multiple-choice test given by the instructor whether it is easy or difficult. The responses were:

B1: I think it’s easy, kasi hindi mahirap magbigay ng exam si Ma'am.

(I think it’s easy, because Ma'am did not give difficult exam.)

B3: It is easy because there’s a question kasi na nandoon lang sa dini discuss nya.

(It is easy because some questions were found during her discussion.)

B5: It is easy if you study, it is easy for me.

B7: There are items which are actually quite difficult, but most of the time it was easy.

Most of the interviewees agreed that the items in multiple choice test were easy for them. Some of the interviewee confidently said that the multiple-choice test were easy to them. The software tells that there were only three (3) items that were difficult and one (1) very difficult item and eleven (11) items which were considered easy items (see Table 4.3) This implies that most of the items needed much attention and should be subjected to revision or deletion. According to Navarro and Santos (2012), items that were on the scale of very easy items, easy items, difficult and very difficult are subjected for revision, replacement or deletion of the said items and the average item retains which then means, it is an ideal item. There were only thirteen (13) items that were on the average scale and this items that remains are good items. In connection to this, according to the study of (Foley, 2010) *“IRT is that the item difficulty and examinee ability parameters are on the same scale”.* In this sense, that the researchers found out that only few students had possessed higher thinking skills and most of the students possess low thinking skills.

IV. Item discrimination of the multiple-choice test in Genetics using 2 PL model

Table 4.4 ***The Estimates of Item Parameter of 2PL Model***

TEST ITEM NUMBER	**DISCRIMINATI ON**	**REMARKS**
1	1.942	**VERY HIGH**
2	-0.585	**-**
3	-0.292	**-**
4	0.171	**VERY LOW**
5	0.833	**MODERATE**
6	0.341	**VERY LOW**
7	-0.291	**-**
8	0.781	**MODERATE**
9	-1.807	**-**
11	0.308	**VERY LOW**
12	0.062	**VERY LOW**
13	-1.129	**-**
14	-2.681	**-**
15	-0.479	**-**
16	2.550	**VERY HIGH**
17	0.246	**VERY LOW**
18	1.000	**MODERATE**
19	2.055	**VERY HIGH**
20	-0.906	**-**
21	0.785	**MODERATE**
22	0.257	**VERY LOW**
23	0.002	**VERY LOW**
24	0.288	**VERY LOW**
25	0.316	**VERY LOW**
27	0.981	**MODERATE**
28	0.764	**MODERATE**
29	-0.316	**-**
30	0.248	**VERY LOW**

Legend:

0	*(No)*
0.01 – 0.34	*(Very Low)*
0.35 – 0.64	*(Low)*
0.65 – 1.34	*(Moderate)*
1.35 – 1.69	*(High)*
> 1.70	*(Very High)*

Table 4.4, shows the item discrimination whether the item possesses no discrimination, very low, low, moderate, high and very high discrimination index. Items 2, 3, 7, 9, 13, 14, 15, 20, and 29 are negative items and tend to discriminate well. This means that all of these items determine the low ability of the examinee that got the correct answer. It can be seen from the table shown that there were nine (9) items with negative values. These items needed to be revised, replaced or deleted. This indicates that 9 items out of 30 in Biology 106 (Genetics) in multiple choice test discriminate the group. This can be seen based on Figure 4.3 (p.36) where some of the Item Charateristics Curves (ICC) of the 30 items were flat or not shaped. According to Baker (2001), *"Items with negative discrimination occur in two ways. First, the incorrect response to a two- choice item will always have a negative discrimination parameter if the correct response has a positive value. Second, sometimes the correct response to an item will yield a negative discrimination index. This tells you that something is wrong with the item: Either it is poorly written or there is some misinformation prevalent among the high- ability students. In any case, it is a warning that the item needs some attention".*

Additionally, the respondents were then asked what they think about the multiple choice test given by the Genetics Instructor whether it is easy or difficult. Here are some of the responses of the students:

B2: For me, somehow difficult because nakakalito pag answer.

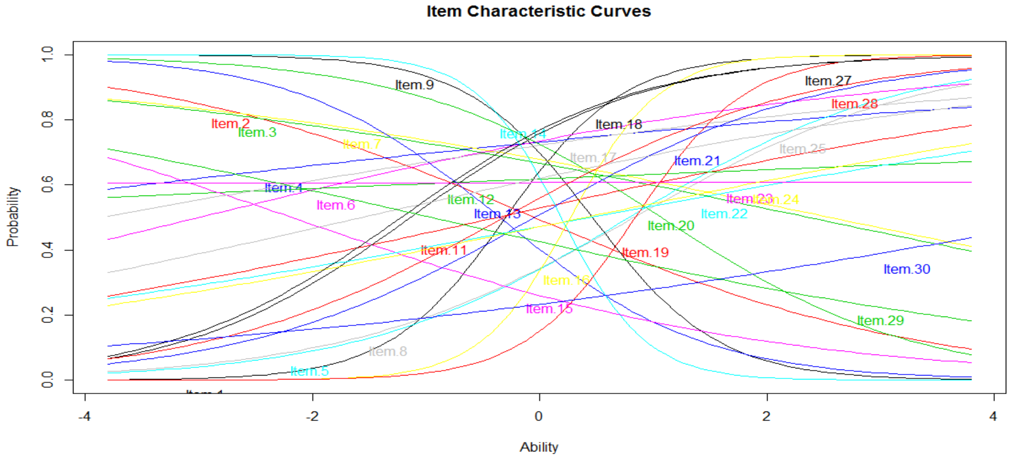
(For me, somehow difficult because I find it confusing to answer.)

B11: Hindi ko ma gets ang exact problem na gustong ipa answer ni Ma'am. (I can't

With this, the results of Table 4.4 and the interviewee's responses agreed with each other that some of the formulated test questions were confusing. It also indicates that the students were confused in comprehending the formulated questions maybe because the items given by the instructor is poorly written.

V. Latent traits based on the used of 2 PL model

Figure 4.3 ***The ICC of the 30 items using 2PL model***



Legend:

S-Shape	*(Good Items)*
Inverted S-Shape	*(Discriminating Item)*
Flat or linear graph	*(Misfit Item)*

Figure 4.3 shows the Item Characteristic Curve (ICC) where it indicates the conditions and location of every item and the latent traits. Above figure presents an individual graph of every item. The *y*-axis indicates the probability of the students of correct item and *x*-axis is the ability or latent traits of the students. The results showed that only few items are considered good items since the shape of the graph are S-shaped and these items are 5, 8, 18, 21, 27 and 28. And these six (6) items were considered the ideal items out of thirty items in multiple choice test. These items also indicate that the higher ability students had, the higher probability to answer these items correctly. And the rest of the items were inverted S-shaped in the graph which shows that the students whose ability is in the low scale had the higher chances to get the correct answer. And these items need some attentions. The inverted S-shape or the negative values mean that these items result to the confusion to the students who has higher thinking skills and the low ability students who got greater chances to answer the difficult items. According to Yu (2013), *"if the item is misfit one reason is that the item is not well-written when low skilled students got the higher scores than high skilled students have given the right answer. By the same token, if a supposedly low skilled student answers many difficult items correctly in a block of questions; one explanation for this could be an instance of cheating"*.

Perhaps, students who have possessed higher thinking skills could not answer most of the items because the test items are poorly written or be considered inappropriate questions which then tend to confusion on the part of the takers. And the researchers also could tell that most of the low ability students had the higher chances to correct the items maybe some of the questions thrown by Instructor was too easy for them.

SUMMARY

Summary of Findings

This study was intended to calibrate the item difficulty and individual responses of the students in Genetics subject in the Marawi City, Mindanao State University- College of Natural Sciences and Mathematics (CNSM). It further attempted to describe the difficulty of an item, item discrimination, and ability of the students, what are their perceptions, concerns, and factors concerning the Genetics subject. After sufficient analyses of the data, the following findings emerged:

1. In terms of parameter logistics models, it was shown that the 2 PL model in an R ltm package can be used in calibrating the multiple-choice test.
2. Out of 30 items in the multiple-choice test in Genetics, there were three items that were eliminated to the 2 PL model because these does not fit to the chosen parameter logistic.

3. In the discrimination parameter of the item analysis, there were nine (9) items that have negative values, which then mean that these items need attention whether for revision, modification, replacement or deletion because these items were either poorly written or there is some misinformation among the high-ability students.
4. Regarding the percentage of the correct and wrong answer of the responses of the students, it was found out that almost half of the students find difficulty in answering the multiple-choice test examination. The respondents generally agreed that Genetics subject is a tough subject.
5. In terms of item difficulty, out of 27 items, there are five (5) items which are labelled as very easy; six (6) easy items, thirteen (13) average items and three (3) difficult items.
6. In terms of item calibration in 2PL model, most of the low ability students got the correct answers on the difficulty level of the items.
7. In terms of the ideal item there were six (6) items which are considered good items.

CONCLUSION

Based on the findings of the data gathered through quantitative and descriptive approach, the following conclusions were drawn:

In view to the findings of this study, the researchers concluded that the teacher is one of the prime problems why students failed to pass the subject. And almost of the students, who were enrolled in Genetics subject, has difficulty in answering during examination. The test item formulated by the teacher has a great factor also that could affect the individual responses to the test items. Based on this study, the researchers found out that some of the multiple test examination items given by the instructor were inappropriate and need revision or deletion. The result also shows that the low ability students had greater chances to answer on the items that were difficult. In contrast, the high ability students had lesser chances in answering the problematic items. According to Baker (2001) in relation to test item, "*it is poorly written or there is some misinformation prevalent among the high-ability students. In any case, it is a warning that the item needs some attention*".

It also concluded that the taker's ability varies from low to high thinking skills. But most of the population possesses low ability and only few test takers possess higher thinking skills.

As a final note the IRT 2PL model fitted to this study and can be an effective and powerful tool to calibrate the item in multiple choice test. It helps the students and instructors in terms of test assessment, questionnaire development, evaluation, refinement and to minimize response burden.

RECOMMENDATIONS

In the light of the findings, this study generated recommendations such as the following:

1. Teachers are recommended to:

1.1 Make a well-written test making use of the table of specification. They must also employ holistic teaching strategies that would cater the learning needs of all their students, may it be about the learning styles and other aspects in learning.
1.2 They must design classroom management technique that will minimize cheating, if not totally eradicate cheating of students during examination.

2. Students;

2.1 Students must do their responsibility in studying their lessons and they must not also hesitate to ask for clarification if they are confused with the lecture. Above all, students must not cheat during examination.
2.2 Encourage not to cheat during examination.

3. Future researchers are recommended to:

3.1 Conduct a comparative study about Genetics or science courses among the different instructors and evaluate if the result would be the same as what is found in this study.
3.2 For more understanding, it is recommended that further research on the same study should be undertaken by future researchers adding the information characteristic curve (performance ability). The item characteristic curve is focused on the ability of the respondents.
3.3 Extend this study to 3 PL Model and replicate the results and analyses of this study using other software which includes but not limited to SAS/STAT 13.1, BILOG and other commercial IRT software packages and fit statistics software models.
3.4 Evaluate the item parameter and ability in R package software in a polytomous context would be a natural extension of this work.

REFERENCES

An, X. & Yung, Y.F. (2014). Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It. Paper SAS364-2014.

Atar, B. & Aktan, D. C. (2013). Person Explanatory Item Response Theory Analysis: Latent Regression Two Parameter Logistic Model. 38, 168.

Baker, F. B. (2001). The basics of item response theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Original work published in 1985. Retrieved from http://echo.edres.org:8080/irt/baker/

Bastari, B. (2000). Linking multiple-choice and constructed-response items to a common proficiency.University of Massachusetts Amherst. Thesis dissertation

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. Quality of Life Research, 16, 95–108.

Brzezińska, J. (2016). Polytomous Item Response Theory Models Using R. *Econometrics 2(52).*

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software May 2012, Volume 48, Issue 6.*

Chen, J. (2012). Applying item response theory models to design a learning progression- based science assessment. Thesis dissertation. Michigan State University.

Colvin, K. (2015). Evaluation of R Package ltm with IRT Dichotomous Models. University of Albany. de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Dorans, N. J., Pomerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

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

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, N Lawrence Erlbaum.

Foley, B. (2010). Improving IRT parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique. Thesis Dissertations. *University of Nebraska - Lincoln.*

Geremew, N.M. (2014). Applying item response theory to the survey of adult skills (PIAAC). Master thesis of statistician Stolcholm University.

Gonzalez, E.(2008). Principles of multiple matrix booklet designs and parameter recovery in largescale. Assessments. *Educational Testing Service, Princeton, NJ, USA*1. *Indiana University, Bloomington, IN, USA*2.

Griffiths, T. (n.d.).Why do students find genetics so difficult to learn.University of British Columbia Vancouver Canada.

Haberman, S J., & Sinharay, S. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. Western Kentucky University. Psychomertica-78,3,417-440.

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

Hula, W., Fergadiotis ,G.,& Martin, N. (2012). Model choice and sample size in item response theory analysis of alphasia test. American Journal Speech-Language Pathology, 21, S38-S50.

Journal of Statistical Software, (JSS) (2007).

Johns, J., Mahaven, S. & Woolf, B. (2006). Estimating student proficiency using an item response theory model. Computer Science Department University of Massachusetts. Amherst Amherst, MA 01003 U.S.A.

Kolen, M. J. and Brennan, R. L. (2011). *Test Equating: Methods and Practices*. New

Köse, I.A. (2014). Assessing model data fit of unidimensional item response theory simulated data. Educational Research Review, 9(17),642-649.

Le, D., Chalmers, R. P., & Liang, L. (2013). Applying item response theory modelling in educational research. Graduate Thesis and Dissertations. Paper 13410.

Lee, SH (2007). Multidimensional item response theory: A SAS MDIT MACRO and Emprical Study of PIAT MATH Test. Unpublished Doctoral Dissertation. The University of Oklahoma.

Maindonald, J. H. (2013). Using r for data analysis and graphics introduction, code and commentary. Centre for Mathematics and its Applications, Australian National University.

Mastura, F. (2016). Comparison of the three item response theory models.Thesis and Dissertations, Mindanao State University-Marawi City.

Microsoft Encarta, (2009).

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. J. (2015). Fitting Diffusion Item Response Theory Models for Responses and Response Times Using the R- Package diffIRT, *Journal of Statistical Science.*

Navarro, R., & Santos, R. (2012). Assessment of learning Outcomes. Lorimar Publishing, Inc. Oxford university Press, (2004).

Ryan, J., & Brockman, F. (2009). A practitioner's introduction to equating with primers on classical test theory and item response theory. Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Technical Issues in Large Scale Assessment.

Sahin, F., and Colvin, K. (2015). Evaluation of R Package ltm with IRT Dichotomous Models. *NERA Conference Proceedings 2015. 6.*

Smith, D.M. & Venables, W.N. (2005). An introduction to r: software for statistical modelling and computing. CSIRO Mathematical and Information Science. Cleveland, Australia.

Student Handbook, 2008 Edition.Mindanao State University.Marawi City.

Tao, W. (2008). Using the score based testlet method to handle local item dependence.

Thesis dissertation. Boston College Lynch School of Education.

Thorpe, G. L. and Favia, A. (2012). "Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications." *Psychology Faculty Scholarship.*

Vocabulary.comDictionary.http://www.vocabolary.com/dictionary.htm-4/18/16.

Wu, M.,& Adams, R.,(2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. Math. Educ. Res. J. 18(2)93-113.

Yang, S. (2007). A Comparison of Unidimensional and Multidimensional Rasch Models Using Parameter Estimates and Fit Indiceswhen Assumption of Unidimensionality is violated. Unpublished Doctoral Dissertaion. The Ohio State University.

Yu, C.H. (2013). A simple guide to the item response theory (IRT) and rasch modeling. Retrived from **http://www.creative-wisdom.com-**08/20/2013.

Zao,Y. (2008). Approaches for addressing the fit of item response theory models to educational test data. Unpublished Doctoral Dissetation. University of Massachusetts Ambers