British Journal of Education

Vol.7, No.8, pp.66-80, August 2019

Published by European Centre for Research Training and Development UK (www.eajournal.org)

# EXAMINING ITEM DIFFICULTY AND STUDENT ABILITY PARAMETERS OF NATIONAL EXAMINATIONS COUNCIL'S BIOLOGY EXAMINATIONS USING THE RASCH MEASUREMENT MODEL IN NIGERIA

**Tommy, Udeme Ezekiel & Udo, Ekerete Mathais**
Faculty of Education, University of Uyo, Uyo, Nigeria

**ABSTRACT:** *This study used the Rasch Model in examining item difficulty and student ability parameters of NECO Biology examination in Nigeria. To achieve this purpose, two research questions were formulated to guide this study. Ex-post facto research design was adopted and a sample of 2,500 Senior Secondary III students drawn through multi-stage proportionate random sampling technique across Nigeria was used for the study. The NECO Biology objective questions from 2016 to 2018 were used as the instruments for data collection. The data were analyzed using item difficulty and ability parameter logits of the WinstepsRasch Measurement Model computer program. The findings of the study showed that item difficulty parameters were appropriate but were not arranged hierarchically from the least difficult to most difficult item and students' ability parameters were appropriately estimated by the examination items. It was recommended among others that NECO should consider using the Rasch Model in developing her examination items so as to have valid and reliable items with appropriate item difficulty parameters and person ability parameters that measure the intended unidimensional construct.*

**KEY WORDS**: item difficulty, student ability, national examinations council, biology examinations, rasch, measurement model

## INTRODUCTION

The Senior School Certificate Examinations (SSCE) has been conducted since 2000 by the National Examinations Council (NECO). The vision of NECO is to prepare and administer credible, standardized, nationally and internationally acceptable examinations which can enable the Nigerian child to further his/her education without any hindrance after secondary school education. NECO is therefore charged with the responsibility of conducting examinations in various subjects including Biology. These NECO Biology examinations are prepared in order to determine the extent students have learnt Biology. Test experts are expected to generate good items that can be used to examine the ability of students from whether homogenous or heterogeneous settings, as the value of such a measure would be domiciled in its quality. To ensure quality of the examination items, such items should measure just one construct as it is assumed the NECO Biology examination items do. If an item does not measure just one construct, the item reduces the validity of the measure for that construct.

The purpose of testing is to estimate students' abilities on a construct of measurement through their responses to a set of items. The scores obtained from the items are used to determine the extent students have learnt, for grading and certification of students. In other words, the items on the examinations go further to determine the success level of students. Since the items on examinations are generally prepared targeting the gains of the subject, the items are expected to be qualified to measure the knowledge and skills that such subject try to develop. In these examinations, decisions are expected to be error-free since they shape the future of the



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

individuals. In test situations involving IRT, examinees' performance on examination items can be explained by defining examinees traits, estimating scores for examinees on these traits and using the scores to explain performance (Opasina, 2009; Adedoyin, 2010). The modern Rasch Measurement Model developed in 1960 by George Rasch is a unidimensional model that belongs to the item response theory (IRT) models.Test experts are expected to generate good items that can be used to examine the abilities of students from whether homogenous or heterogeneous settings, as the value of such a measure would be domiciled in its quality. This will be primarily possible through measuring tools (tests or examinations) whose item difficulty and person parameters are appropriate for the level of the students. But most at times these parameters are wrongly estimated during test construction because most examining bodies including NECO continue to rely on the classical test theory for test development which has a lot of disadvantages including sample and item dependency despite of the strong presence of item response theory. When this happens, it is difficult to actually estimate the true difficulty parameter of examination items as well as true ability parameters of students who write the examinations. Item and person parameters will in turn be properly estimated if tests are developed using the modern Rasch Measurement Model. This study therefore examined the item difficulty and person ability parameters of the NECO Biology examinations using the modern Rasch Measurement Model in order to ascertain if they are appropriate or not since they were developed based on the classical test theory framework.

**Purpose of the Study**

The purpose of the study was to examine the item difficulty and person ability parameters of the NECO Biology examinations using the modern Rasch Measurement Model. Specifically the objectives of the study were to:

1. Ascertain the item difficulty parameters of the NECO Biology examination items using the Rasch Measurement Model.
2. Examine the students' ability parameters at the NECO Biology examination items using the Rasch Measurement Model.

**Research Questions**

The following research questions were formulated to guide this study:

1. What are the item difficulty parameters of the NECO Biology examination items using the Rasch Measurement Model?
2. What are the students' ability parameters on the NECO Biology examination items using the Rasch Measurement Model?

## LITERATURE AND THEORETICAL UNDERPINNING

In the modern Rasch Model, the probability of a specified response (example, right and wrong) is modeled as a function of person ability and item difficulty parameters (Rasch, 1960). Specifically, the probability of a correct response is modelled as a logistic function of the difference between examinee ability and item difficulty (Chong, 2011; Aliyu, 2015).The use of the Rasch Measurement Model in item difficulty and person ability estimation is to ensure specific objectivity. Specific objectivity is one of the theoretical merits of the Rasch Model. It requires that a person's ability is independent of the specific set of items used to measure it, and also ensure invariant scores, as a good measurement should yield invariant scores (that is, the ratio of difficulties between different items should remain constant across the same ability



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

level of examinees). These stated requirements of specific objectivity in measurement are not inherent in most of the measurements carried out in education as a discipline generally, and in our school systems in particular in Nigeria (Joshua, 2005). A modern measurement approach which has been developed and proposed by measurement experts to address this vexing issue of specific objectivity and other shortcomings of the CTT is the IRT (Adedoyin, 2010).

In most contexts, the parameters of the model characterize the proficiency of the respondents and the difficulty of the items as locations on a continuous construct of measurement. For example, in educational tests, item parameters represent the ability or attainment level of people who are assessed. The higher a student's ability relative to the difficulty of an item, the higher the probability of a correct response on that item. When a person's location on the construct is equal to the difficulty of the item, there is by definition a 0.5 probability of a correct response in the Rasch Model (Wu & Adams, 2007). Location of items and persons along the measurement scale is estimated by the model from the proportion of response of each person to each item. The scale resulting from the Rasch analysis of ordinal response of each person to each item has the properties of an interval scale. Interval scales have known and equal interval between two graduations. On interval scales, numbers tell how much more of the construct of interest is present. These scales are linear and quantitative. Raschmodeling puts particular emphasis on covering the entire construct of measurement and requires the inclusion of items with different intensity to achieve acceptable measures (Soutar& Garry, 2001). This feature is considered particularly useful for developing a measurement like the NECO Biology examination as the concept is designed to cover the entire width of possible responses of students' learning experience.

Examination items are unidimensional, that is, they measure the same underlying construct for all examinees if the item difficulty order is stable for different subclasses of examinees and constant item difficulty order indicates that performance on the examination items requires the same skills, knowledge and strategies for all examinees (Rasch, 1960; Vigneau&Bors, 2005). Different order of item difficulties shows that different mechanism and skills are employed to solve the items, and therefore, in this case, the nature of the construct depends on the class to which an examinee belongs. Hence, the correlation of the examination with external criteria may also change, that is, class membership acts as a moderator variable which is further evidence of the change of the construct (Embretson, 2007; Obinne*et al.*, 2013). It is important to note that, under the Rasch Model, not only the order of items should remain constantly hierarchical across sub-populations but their estimated difficulty parameters and the distances among them should also remain invariant (Amuche& Fan, 2014). Changes in item difficulty parameters across sub-populations of the same ability level indicate lack of unidimensionality, as there is no uniformity in the underlying construct of measurement (Smith, 2004a). According to Wyse &Mapuranga (2009), when the items are not unidimensional, one cannot compare the abilities of the examinees, the difference in degrees but differences in kind.

Investigating the invariance of item difficulty across examinees is a well-documented way of checking the appropriateness of the items in unidimensionality assumption (Kubinger, 2005; Adedoyin, Nenty&Chilisa, 2008). However, the requirement for invariant item difficulty parameter gets violated quite often. It is very common to check the invariance of item difficulty parameter of a test such as the NECO Biology examination items through appropriate model like the Rasch Measurement Model. The Rasch Measurement Model helps to identify which item difficulty estimate differs most and can direct the test developer to a more accurate


Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

hierarchical order of items. Inappropriate hierarchical order of examination items based on their difficulties can be due to poor or lack of definition of the construct and poor item construction.In CTT, besides the inadequacy in representing students' abilities in raw scores, statistics obtained from raw scores such as p-value (the proportion of correct answer) are also sampled dependent (Adedoyin*et al.*, 2008; De Champlain, 2010). For example, a higher p-value will be obtained from a sample of above average student. Item with high p-value is considered an easy item. In contrast, below average sample will provide a lower p-value that indicates a more difficult item. Therefore, it can be seen that different interpretation can be made from the same single item. As a consequence, if the sample does not reflect the population, the item statistics obtained from the sample are limited in their usefulness. Similarly, since the raw score is defined in terms of number of correct response, it is highly influenced by the test difficulty. Easier test will produce students with higher abilities and vice versa. In short, since students' abilities are test dependent, comparison among different students who sit for different tests does not provide a meaningful interpretation.

In contrast, the modern Rasch Model helps to establish the consistency of a set of examination items. Estimates of students' abilities are independent of which items are used for comparisons and likewise estimates of item difficulties that are used for comparisons. The model also requires invariance in the unit of measurement, and it is the production of these constant units of measurement that result in equal-interval scale scores for students. The Rasch Model analysis indicates whether a set of examination items like the NECO Biology examination items can be considered to comprise a unidimensional measurement scale with equal-interval level properties, and whether scale scores remain invariant across different groups. Invariance is the core measurement principle on which the model rests, with the analysis seeking to identify anomalies in the items which may undermine such invariance of measurement. Anomalies can lead to a better understanding of how the items are ordered and the property being measured and the task is to work towards a better fit of the items to the model's requirements until the match is sufficient to provide invariant measures (Bond & Fox, 2013). This may be achieved by the rearrangement or modification of items and the development of new items.

An item's location on the measurement scale is interpreted as the relative difficulty respondents, as a whole, have in responding correctly to the item. Items located to the right of the scale are more difficult to endorse than those to the left, with the item content helping to define what more or less of the construct signifies. More difficult items are likely to be endorsed positively only by students possessing higher ability, whereas easier or less difficult items are likely to be positively endorsed by many students, including those with lower ability (Cavanagh &Romanoski, 2006). Logits possess several advantages over raw scores. Firstly, as these measures share a common unit on a scale, researchers can readily visualize the order of difficulty of items in an examination relative to each other and can easily ascertain where any individual student is located in relation to all items (Hagquist&Andrich, 2004). Secondly, the conversion of ordinal data to equal-interval data means that any difference in logits implies equal difference in ability on the construct measured by the items (Smith &Plackner, 2009). Item or student logit locations can therefore be summed and used in standard statistical analyses. Finally, unlike raw student and item scores, these measures allow comparisons between subjects from the same group to be made independently of which students are used for the comparison (Andrich& Styles, 2004).


Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

Nopiah*et al.,* (2010) also analyzed the items' difficulty parameters in their research and reported that the students' mean squares were around 1.52 logits and the items were appropriately arranged in terms of difficulty from the easiest to the most difficult items, and all the students responded to the items accordingly. The MNSQ infit of all items .7 and 1.20 as recommended. They concluded that the item difficulty arrangement for the 20-item test increased from the first to the last item, meeting the unidimensionality assumption. The researchers also investigated separation and reported that the items summary gave a good summary with separation G = 6.81 which indicated that the items were sufficiently well separated in difficulties, though some items were redundant by having the same difficulties. Khairani*et al.,* (2012) also studied items' difficulty parameters in their work and reported that all items were within the acceptable range of 0.7 - 1.3 infit and outfit MNSQ, and that the scale was unidimensional because the item difficulty ordering was well positioned except for 3 items which deviated. The most difficult item, item 5 (2.53 logit) was twice as difficult compared to item 38 (1.27 logit). Yang *et al.,* (2011) in the same vane analyzed the difficulty parameter in their study. They reported that the distribution of item difficulty distinctly diverged from normality as they did not follow hierarchical order from least difficult to most difficult items. 21 items were with difficulties lower than the ability of the least able students and only 2 items with difficulties higher than the ability of the most proficient students while 25 items possessed difficulty within the range of examinees' abilities distribution. Yang *et al*., (2011) concluded that the scale was not unidimensional in nature since it could not appropriately aligned the items from their increasing order of difficulty and advised that in subsequent testing; some items should be revised or deleted.

Herrmann-Abell and DeBoer (2011) also examined the difficulty parameter in their research and reported that the infit and outfit MNSQ values for the majority of the items were within the acceptable range of 0.7 to 1.3. They also that the items were placed according to their increasing difficulty, except for few items which showed redundancy, but concluded that on the overall the instrument was a unidimensional one. Phillipson (2008) examined students' ability parameters and reported that the mean student logit scores for the RPM across the four grade levels were increasing accordingly. It showed that, on average, the RPM was becoming increasingly easier and this change was obvious when examining the item and student maps for each grade level. The findings further showed that the most capable students and most difficult items were located at the top end of each scale. He concluded that a good relationship between a test and a student is evident when there is an adequate spread of items according to their increasing order of difficulty and there is a close alignment between items and students at both ends of the scale. To him, the increased uniformity in students' abilities according to increased item difficulties showed that the scales were unidimensional instruments.

Similarly, Khairani*et al.,* (2012), analyzed students' ability parameters and reported that with regards to students' mathematical abilities, 335 students showed responses that were within the expectation of the model, suggesting that the items were unidimensional and contributed usefully to the measurement of mathematical ability with most students' ability estimates from 0.8 to 1.2. Yang *et al.,* (2011) also studied students' ability estimates. They reported that the instrument did not properly estimate students' abilities and as such was not unidimensional. Although, 62% of students had ability measures ranged between 0.7 - 1.3, there were only five items with difficulty in this range and separation index was low indicating that the instrument was not reliable in separating students into ability levels and some items could not differentiate high ability from low ability students. Hermann-Abell&DeBoer (2011) examined the students' ability parameters and reported that the item-person map showed that low performance was


Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

represented at the bottom of the map and high performance was represented at the top of the map. Most students' logits estimates were between .52 - .74, showing that the students' chemistry abilities were appropriately estimated by the CAT and the CAT was unidimensional.

Nopiah*et al.,* (2010), investigated the students' ability parameters in their study and reported that student number 118 who scored 13 over 20 was found to be a misfit (too unpredictable) where MNSQ outfit was 3.58, exceeded the recommendation. This meant that the student responded in the reverse direction where he answered more of difficult questions when others could not and vice versa because the student answered items 14 and 16 correctly, two difficult items while got wrong two easier items, 10 and 11. They stated that this student outcome did not meet the Rasch model expected outcomes, and raised some conclusions that the student underestimated the easiest items and miscalculated the matrix operations; conversely, for the difficult items, the student probably had special interest or knowledge on the topic and/or comfort answering statement-based question, and on the other hand, the student simply guessed the answers. But on the whole, the researchers concluded that the students' abilities were appropriately estimated through the items as only one student response pattern deviated from the Rasch measurement pattern indicating unidimensionality in measurement.

## METHODOLOGY

The ex-post facto research design was adopted for the study. Ex-post facto research design is a design where the researcher carries out an empirical inquiry into a phenomenon and does not have control of the occurrence of the data because their manifestations have already occurred. This design was suitable for this study since there was no treatment and manipulation of the data. The population of the study comprised all the Senior Secondary School III students in Nigeria who enrolled for the 2018 ordinary level examinations. These students served as the respondents for the study. A sample size of 2,593 students drawn through multi-stage proportional random sampling technique was used for the study. The ages of the students ranged from 15 to 19 with a mean and standard deviation of 16 and 4.89 respectively.The sample consisted of 1,192 males and 1,401 females.The research instruments used in data collection for this study were the NECO 2016, 2017 and 2018 Biology objective items. The NECO Biology objective items for each year are made up of 60 items, making it a total of 180 items. Each of the items has options A to E, with one as a correct option and four distracters.The items that made up the research instruments were pooled from the NECO Biology examinations which are standardized examinations conducted by a reputable examination body. The items were deemed valid and reliable because they were validated by NECO and as such did not warrant any further validation as one of the thrusts of this study was to examine if the NECO Biology examination items were valid and reliable. The instruments were administered on the sampled SSIII students in the sampled schools by the researcher with the help of research assistants and teachers. The researcher explained to the students the purpose of the study and the need for them to respond appropriately and candidly. But 2,500 students completed and returned the instruments administered, thereby giving it a return rate of 96%. Rasch Measurement Model software, Winsteps version 3.81.0 was used for data analysis. In Winsteps, the measures are determined through iterative calibration of both person and item using the NECO Biology examinations. The recommended item difficulty and ability logits ranged from -3 to +3 for appropriate item difficulties in an examination and ability parameters. A negative logit difficulty estimate indicates items that have low difficulty index (easy items), and a positive logit difficulty estimate shows items with a high difficulty index (hard items);


Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

positive logit means that an examinee is more able in a test while negative logit indicates less ability (Brentari and Golia, 2007; Phillipson, 2008; Linacre, 2010).

## RESULTS AND FINDINGS

### Research Question One

What are the item difficulty parameters of the NECO Biology examination items based on the Rasch Measurement Model?

In order to answer research question three, item difficulty logits were employed. The recommended item difficulty logits ranged from -3 to +3 for appropriate item difficulties in an examination. A negative logit difficulty estimate indicates items that have low difficulty index (easy items), and a positive logit difficulty estimate shows items with a high difficulty index (hard items) (Green and Frantom, 2002; Linacre and Wright, 2004; Brentari and Golia, 2007; Phillipson, 2008; Linacre, 2010). The item difficulty parameters of the NECO Biology examinations for the five years under consideration are presented on Table 1.

**Table 1: Item Difficulty Parameters of the NECO Biology Examination Items for 2016, 2017 and 2018 based on the Rasch Measurement Model.**

| Item | Item Difficulty Parameter and Each Item Total Score | | | | | |
|---|---|---|---|---|---|---|
| | 2016 | Total Score | 2017 | Total Score | 2018 | Total Score |
| 1 | -1.96 | 424 | -1.39 | 377 | -1.41 | 378 |
| 2 | -1.90 | 420 | -1.72 | 410 | -1.56 | 390 |
| 3 | -.64 | 315 | -.91 | 334 | -.94 | 337 |
| 4 | -.93 | 343 | -.81 | 324 | -.42 | 285 |
| 5 | .38 | 204 | .25 | 211 | -.86 | 329 |
| 6 | .47 | 195 | .18 | 219 | -.15 | 257 |
| 7 | -.46 | 296 | -.20 | 260 | -.64 | 307 |
| 8 | -.49 | 299 | -.07 | 246 | .63 | 176 |
| 9 | -1.63 | 402 | -1.40 | 378 | -.52 | 295 |
| 10 | .01 | 244 | -.11 | 250 | -.09 | 250 |
| 11 | -1.51 | 393 | -1.19 | 360 | -.85 | 328 |
| 12 | .36 | 206 | .67 | 168 | 1.31 | 116 |
| 13 | -1.44 | 388 | -.96 | 339 | -1.16 | 357 |
| 14 | .76 | 165 | .85 | 150 | .96 | 145 |
| 15 | -1.23 | 371 | -.60 | 303 | -.63 | 306 |
| 16 | .49 | 192 | .42 | 193 | .53 | 186 |
| 17 | -.09 | 255 | -.37 | 278 | -.72 | 315 |
| 18 | -.32 | 281 | -.99 | 342 | -1.00 | 342 |
| 19 | -.27 | 275 | -.55 | 298 | -.41 | 284 |
| 20 | .44 | 198 | .29 | 207 | -.14 | 256 |
| 21 | .53 | 188 | .54 | 181 | .44 | 195 |
| 22 | 1.53 | 98 | 1.43 | 101 | .96 | 145 |



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

| | | | | | | |
|---|---|---|---|---|---|---|
| 23 | -.32 | 281 | -.18 | 258 | .67 | 172 |
| 24 | -.44 | 294 | -.52 | 294 | -.48 | 291 |
| 25 | .58 | 183 | .87 | 148 | .70 | 169 |
| 26 | -.27 | 275 | -.46 | 288 | -.48 | 291 |
| 27 | .68 | 173 | .61 | 174 | .75 | 165 |
| 28 | 1.05 | 137 | 1.00 | 136 | .82 | 158 |
| 29 | .43 | 199 | .55 | 180 | .35 | 204 |
| 30 | -.11 | 258 | .04 | 234 | .13 | 227 |
| 31 | -.43 | 292 | -.72 | 315 | -.81 | 324 |
| 32 | -.20 | 268 | -.49 | 291 | -.45 | 288 |
| 33 | -.46 | 296 | -.47 | 289 | -.38 | 281 |
| 34 | .41 | 201 | -.19 | 259 | .01 | 240 |
| 35 | .22 | 221 | .36 | 200 | 1.15 | 129 |
| 36 | -.78 | 329 | -.53 | 295 | -.30 | 272 |
| 37 | -.36 | 285 | -.24 | 264 | -.63 | 306 |
| 38 | -.32 | 281 | -.22 | 262 | -.44 | 287 |
| 39 | .62 | 179 | .37 | 198 | -.26 | 268 |
| 40 | .27 | 216 | -.01 | 239 | -.34 | 277 |
| 41 | .13 | 231 | .31 | 205 | 1.12 | 131 |
| 42 | -.80 | 331 | -.46 | 288 | -.40 | 283 |
| 43 | .52 | 189 | .52 | 183 | .37 | 202 |
| 44 | -.44 | 294 | -.52 | 294 | -.64 | 307 |
| 45 | 1.52 | 99 | 1.30 | 111 | 1.44 | 106 |
| 46 | .64 | 177 | .90 | 146 | .80 | 160 |
| 47 | .56 | 185 | .34 | 202 | .51 | 188 |
| 48 | 1.09 | 134 | .81 | 154 | .94 | 147 |
| 49 | .00 | 246 | -.19 | 259 | -.26 | 268 |
| 50 | 1.30 | 116 | .99 | 137 | 1.12 | 131 |
| 51 | .25 | 218 | -.03 | 241 | -.17 | 259 |
| 52 | -.55 | 305 | .06 | 232 | -.26 | 268 |
| 53 | .63 | 178 | 1.21 | 118 | .67 | 172 |
| 54 | .83 | 158 | .48 | 187 | .41 | 198 |
| 55 | .25 | 218 | .23 | 213 | -.14 | 256 |
| 56 | 1.18 | 126 | .99 | 137 | 1.02 | 140 |
| 57 | -.35 | 284 | .00 | 238 | -.10 | 251 |
| 58 | 1.21 | 123 | 1.29 | 112 | 1.42 | 107 |
| 59 | .01 | 245 | -.49 | 291 | -.55 | 298 |
| 60 | -.62 | 313 | -.89 | 332 | -.72 | 315 |
| Mean | .00 | 244. | .00 | 238.9 | .00 | 241.9 |
| SD | .80 | 8 | .73 | 74.2 | .74 | 74.1 |
| | | 80.0 | | | | |

The result of the 2016 NECO Biology examination in Table 1 reveals that the item difficulty ranges from -1.96 to 1.53. It shows that 28 items, making 46.67% of the items have difficulty estimates with negative logits are fairly easy while 32 items or 53.33% of the items have difficulty estimates with positive logits are fairly difficult. The result also shows that there is a perfect match between easy and difficult items with a mean of .00 and low standard deviation



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

of .80. It also reveals that the difficulty parameters are reasonably appropriate for the abilities of the students and shows little variability in the scores of the students. Apart from items 1 and 2 which are at their appropriate positions as the two easiest items, other items are not hierarchically positioned in terms of their difficulties. For example, item 22 which is the most difficult item is supposed to be the last item on the scale as demanded by the Rasch Model for unidimensional scales. The result of 2017 NECO Biology examination indicates that the item difficulty ranges from -1.72 to 1.43. It shows that 31 items which is 51.67% of the items have difficulty parameters with negative logits which means they are fairly easy items and 29 items, making 48.33% of the items have difficulty parameters with positive logits implying fairly difficult items. The result also reveals that there is a match between the easy items and difficult items with a mean of .00 and low standard deviation of .73. The item difficulty parameters are appropriate for the abilities of the students and there is little variation in the scores of the students as indicated on the table. The item difficulty parameters are not hierarchically arrange as item 2, the easiest item is suppose to be item 1, followed by other items and item 22 with difficulty of 1.43 logit is suppose to be the last item on the scale to be a perfect unidimensional instrument.

The table also indicates the result of the 2018 NECO Biology examination with item difficulty that ranges from -1.56 to 1.44. It reveals that 35 items or 58.33% of the items have difficulty parameters with negative logits, meaning fairly easy items and 25 items which is 41.67% of the items have difficulty parameters with positive logits, implying fairly difficult items. The result also shows that there is a balance between easy and difficult items with a mean of .00 and low standard deviation of .74. The difficulty estimates are appropriate for the abilities of the students and it also shows that there is little variation in the scores of the students. Hierarchically, item 2 is suppose to be the first item on the scale while item 45 is suppose to be the last item on the scale as it is the most difficult item which is required by the Rasch Measurement Model for unidimensional scales. Overall, the item difficulties for the five years under study are appropriate for the students' abilities but are not ordered hierarchically from least to most as recommended by the Rasch Model for unidimensional scales.

**Research Question Two**

What are the students' ability parameters at the NECO Biology examinations items based on the Rasch Measurement Model?

In order to answer research question four, ability parameter logits were employed. Ability parameter is expressed in logits and the recommended range is from -3 to +3. Positive logit means that an examinee is more able in a test while negative logit indicates less ability (Green and Frantom, 2002; Linacre and Wright, 2004; Brentari and Golia, 2007; Phillipson, 2008; Linacre, 2010). The students' ability parameters at the NECO Biology examination items for the three years under study are presented on Tables 2.



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

**Table 2: Students' Ability Parameters at the NECO Biology Examination Items for 2016, 2017 and 2018 based on the Rasch Measurement Model.**

| 2016 | | | | 2017 | | | | 2018 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Count | Raw Score | Ability para-meter | Ass. S E | Count | Raw Score | Ability para-meter | Ass. S E | Count | Raw Score | Ability para-meter | Ass. SE |
| 2 | 5 | -2.64 | .48 | 4 | 5 | -2.67 | .48 | 2 | 5 | -2.61 | .48 |
| 2 | 6 | -2.43 | .41 | 3 | 8 | -2.11 | .39 | 2 | 7 | -2.22 | .41 |
| 1 | 7 | -2.23 | .36 | 2 | 9 | -1.96 | .37 | 7 | 10 | -1.78 | .36 |
| 1 | 8 | -2.08 | .35 | 4 | 10 | -1.82 | .36 | 3 | 11 | -1.66 | .35 |
| 1 | 9 | -1.93 | .33 | 3 | 11 | -1.70 | .35 | 9 | 12 | -1.54 | .33 |
| 5 | 10 | -1.79 | .33 | 13 | 12 | -1.58 | .34 | 18 | 13 | -1.43 | .33 |
| 3 | 11 | -1.67 | .32 | 12 | 13 | -1.47 | .33 | 10 | 14 | -1.33 | .32 |
| 5 | 12 | -1.55 | .31 | 7 | 14 | -1.36 | .32 | 13 | 15 | -1.23 | .31 |
| 8 | 13 | -1.44 | .31 | 18 | 15 | -1.26 | .31 | 12 | 16 | -1.13 | .31 |
| 5 | 14 | -1.34 | .30 | 10 | 16 | -1.16 | .31 | 14 | 17 | -1.04 | .30 |
| 9 | 15 | -1.24 | .30 | 3 | 17 | -1.07 | .30 | 14 | 18 | -.95 | .30 |
| 10 | 16 | -1.14 | .29 | 18 | 18 | -.98 | .30 | 20 | 19 | -.87 | .29 |
| 6 | 17 | -1.05 | .29 | 5 | 19 | -.89 | .29 | 20 | 20 | -.78 | .29 |
| 9 | 18 | -.96 | .29 | 20 | 20 | -.81 | .29 | 15 | 21 | -.70 | .29 |
| 11 | 19 | -.88 | .28 | 10 | 21 | -.72 | .29 | 10 | 22 | -.62 | .28 |
| 13 | 20 | -.79 | .28 | 8 | 22 | -.64 | .28 | 11 | 23 | -.54 | .28 |
| 12 | 21 | -.71 | .28 | 16 | 23 | -.56 | .28 | 8 | 24 | -.46 | .28 |
| 5 | 22 | -.63 | .28 | 4 | 24 | -.48 | .28 | 21 | 25 | -.39 | .28 |
| 23 | 23 | -.55 | .28 | 20 | 25 | -.41 | .28 | 16 | 26 | -.31 | .28 |
| 9 | 24 | -.47 | .28 | 13 | 26 | -.33 | .28 | 13 | 27 | -.23 | .28 |
| 20 | 25 | -.40 | .28 | 14 | 27 | -.25 | .27 | 9 | 28 | -.16 | .28 |
| 30 | 26 | -.32 | .27 | 15 | 28 | -.18 | .27 | 9 | 29 | -.08 | .27 |
| 19 | 27 | -.24 | .27 | 38 | 29 | -.08 | .27 | 9 | 30 | -.01 | .27 |
| 25 | 28 | -.17 | .27 | 18 | 30 | .00 | .27 | 12 | 31 | .07 | .27 |
| 23 | 29 | -.09 | .28 | 12 | 31 | .09 | .27 | 13 | 32 | .14 | .28 |
| 22 | 30 | -.02 | .28 | 19 | 32 | .17 | .27 | 11 | 33 | .22 | .28 |
| 17 | 31 | .08 | .28 | 16 | 33 | .24 | .27 | 10 | 34 | .30 | .28 |
| 18 | 32 | .15 | .28 | 16 | 34 | .33 | .28 | 7 | 35 | .37 | .28 |
| 13 | 33 | .23 | .28 | 15 | 35 | .40 | .28 | 16 | 36 | .45 | .28 |
| 13 | 34 | .31 | .28 | 17 | 36 | .48 | .28 | 13 | 37 | .53 | .28 |
| 19 | 35 | .38 | .28 | 13 | 37 | .55 | .28 | 32 | 38 | .61 | .28 |
| 16 | 36 | .46 | .29 | 29 | 38 | .64 | .28 | 19 | 39 | .69 | .29 |
| 8 | 37 | .54 | .29 | 13 | 39 | .72 | .29 | 14 | 40 | .78 | .29 |
| 28 | 38 | .62 | .28 | 15 | 40 | .81 | .29 | 23 | 41 | .86 | .29 |
| 12 | 39 | .70 | .29 | 17 | 41 | .89 | .29 | 12 | 42 | .95 | .30 |
| 13 | 40 | .79 | .29 | 8 | 42 | .98 | .29 | 5 | 43 | 1.04 | .30 |
| 19 | 41 | .87 | .29 | 6 | 43 | 1.07 | .30 | 10 | 44 | 1.13 | .31 |
| 6 | 42 | .96 | .30 | 7 | 44 | 1.16 | .30 | 8 | 45 | 1.23 | .31 |
| 6 | 43 | 1.05 | .30 | 5 | 45 | 1.26 | .30 | 6 | 46 | 1.33 | .32 |
| 8 | 44 | 1.14 | .31 | 7 | 46 | 1.36 | .29 | 15 | 47 | 1.43 | .33 |
| 4 | 45 | 1.24 | .31 | 4 | 47 | 1.46 | .27 | 3 | 48 | 1.54 | .34 |
| 6 | 46 | 1.34 | .32 | 1 | 48 | 1.57 | .25 | 3 | 49 | 1.66 | ..35 |
| 11 | 47 | 1.44 | .33 | 1 | 49 | 1.69 | .22 | 3 | 50 | 1.79 | .36 |
| 2 | 48 | 1.55 | .34 | 1 | 50 | 1.82 | .21 | | | | |
| 1 | 49 | 1.67 | .35 | | | | | | | | |
| 1 | 50 | 1.80 | .36 | | | | | | | | |
| Mean | 29.4 | -.15 | .29 | | 28.7 | -.81 | .29 | | 29.0 | -.09 | .30 |
| SD | 9.3 | .78 | .03 | | 9.8 | .81 | .03 | | 10.9 | .90 | .02 |



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

The result of the 2016 NECO Biology examination in Table 2 reveals that the students' ability parameters range from -2.64 for a raw score of 5 which is an exceedingly low score to a parameter of 1.80 for a raw score of 50, a substantially high score. From the result, 279 students, making 55.8% are less able at the items with logits range of -2.64 to -.02 and 221 students, making 44.2% of the students are more able at the items with logit range of .08 to 1.80. It also indicates that the standard error associated with each ability range from .28 to .48. For instance, the standard errors associated with ability estimates of -2.64, -1.05, 1.34 and 1.80 are .48, .30, .32 and .36 respectively. These values mean that 52%, 70%, 68% and 64% of the total variance associated with these ability estimates of students could be attributed to true variance in a unidimensional scale. Table 2 shows that the students' ability estimates for 2017 NECO Biology examination range from -2.67 assigned to a raw score of 5, a very low score to 1.82 assigned to a raw score of 50, a substantially high score. From the table, 260 students or 52% of the students are less able at the items with logit range of -2.67 to -.08 while 240 students which is 48% of the students are more able at the items with a logit range of .00 to 1.82. The result also reveals the standard errors associated with each ability estimate and it range from .21 to .48. For example, the standard errors associated with the ability parameters of -2.11, -.18, .40 and 1.82 are .39, .27, .28 and .21 respectively. These values indicates that 61%, 73%, 72% and 79% of the total variance associated with each of these ability estimates could be attributed to true variance in a undimensional scale.

The table also indicates that the result of the students' ability parameters for the 2018 NECO Biology examination range from -2.61 for a raw score of 5, an exceptionally low score against 1.79 for a raw score of 50, a substantially high score. From the result, 275 students which is 55% of the students are less able at the items with logit range of -2.61 to -.01 and 225 students, making 45% of the students are more able at the items with logit range of .07 to 1.79. The result also shows that the standard errors associated with each ability parameter and it ranged from .27 to .48. For instance, the standard errors associated with ability parameters of -1.78, -1.04, .78 and 1.66 are .36, .30, .29 and .35 respectively. These values indicate that 64%, 705, 71% and 65% of the total variance associated with each of these ability parameters could be attributed to true variance in a unidimensional scale. The results of the five years under study reveal that the students' abilities were appropriately estimated as they all fall within the values recommended in Rasch analysis.

## DISCUSSION

The objective of research question one was to assess the item difficulty parameters of the NECO Biology examinations based on the Rasch Measurement Model. Though the result has shown that the item difficulty parameters were appropriate for estimating the students' abilities, they were not arranged hierarchically from the least difficult item to the most difficult item as demanded by Rasch Measurement Model for unidimensional scales. The implication of this is that the students were not given the opportunity to answer from less difficult items to most difficult items as demanded by the Rasch Model for unidimensional examinations which could result in loss of interest and lack of motivation by the students concerning the examinations. This finding is not supported by the finding of Phillipson (2008) because she found out in her study that the most difficult items that constituted the scales for her study were located at the lower end while the easier items were located at the beginning of the scales as demanded by the Rasch Model for uindimensional scales. This finding disagrees with the finding of Herman-Abell and DeBoer (2011) since they reported in their study that the items that made up their


Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

instrument were placed according to their increasing difficulty, though few items showed redundancy, but concluded that on the overall, the instrument was a unidimensional one.

The objective of research question two was to examine the students' ability parameters at the NECO Biology examination based on the Rasch Measurement Model. The result has shown that the students' abilities were appropriately estimated as their logits were within the values recommended in Rasch analysis and also seen in the ranking of the students from least capable to most capable according to their scores on the examinations. The result has also revealed that the items were fair to both the less able students and the more able students as shown in their scores. Ability parameter indicates the amount of construct possessed by a student as measured by the items or scales. An examination should have items that allow different composites of ability to correctly respond to the items. Different students' abilities can best be estimated when the items that make up an examination measure a unidimensional construct. This finding agrees with the finding of Nopiah*et al* (2010) who found out in their study that students' abilities were appropriately estimated through the items that made up the examinations. This finding is also in line with the finding of Khairani*et al* (2012) as they reported that the students' ability estimates were within the expectation of the Rasch Measurement Model and the items contributed usefully in estimating students' abilities.

**Implications to Research and Practice**

1. The Rasch Measurement Model collapses the score categories to only two dichotomous categories. This increases the precision of the model on the items and persons involved in the examinations. The implication for this study is that there would be very precise item and person parameter estimates and other psychometric qualities measured.

2. Students' performance, strength and weaknesses in tests could be adequately explained using the findings of this nature. The examining bodies and test analysts can use the various psychometric properties possible in Rasch Measurement Model like item difficulty and person ability parameters

3. NECO should henceforth employ the Rasch Model in developing her examination items so as to have valid and reliable items with appropriate item difficulty parameter and person ability parameter that measure the intended unidimensional construct.
4. Since the study permits the identification of each examinee's strength and weakness in testing situations, this diagnosis should be used to improve the quality of Biology instructions. Two major methods to improve the quality of Biology instructions are recommended.

a. Instructional treatment: Biology teachers should develop instructional materials to fit the student's ability patterns. For instance, if a student has low ability in environmental contents, then a presentation emphasizing the use of diagrams, models and concrete demonstrations are suggested.
b. Instruction with feedback and corrective procedures: Feedback devices for example, formative evaluation and diagnostic tests should be built into the instruction to identify deficiencies in the students' learning of a given Biology unit. The most corrective techniques include: re-teaching of a selected Biology unit, small group study session and individualized tutoring. Essentially, the corrective devices will provide the students with


Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

the instructional cues, the learning participation and the reinforcements which are best suited to their characteristics and needs.

## CONCLUSION

The Rasch Model provides rich interpretation regarding item and person parameters on a test. From the findings of the study, it was concluded that the item difficulty parameters from the NECO Biology examinations were in line with the item difficulty specification of the Rasch Measurement Model, but were not arranged hierarchically from less difficult to most difficult items as required by the Rasch Measurement Model for unidimensional scales while the students' ability parameters at the NECO Biology examination items were appropriately estimated as they were within the required range demanded by the Rasch Measurement Model for unidimensional scales.

### Future Research

The authors have presented a thorough primer on how to carry out studies using the RaschModel. There are, however, a number of unresolved areas that need further research which are:

1. Further studies should be carried out on item and person parameters of other examinations conducted by NECO using the Rasch Measurement Model.
2. Dimensionality analysis of WAEC, NABTEB and JAMB examinations could be carried out using the Rasch Measurement Model
3. The use of the Rasch Measurement Model in the equating of Biology examinations conducted by NECO needs to be studied.
4. The Rasch Measurement Model should be used in the development of criterion-referenced tests.

### References

Adedoyin, O. (2010)*I nvestigating the invariance of person parameter estimates based on classical test and item response theories*, International Journal of Education, 2 (2) 107-113.

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

Aliyu, R. T. (2015) *Construct validity of Mathematics test items using the Rasch Model*, International Journal of Social Science and Humanities Research, 3 (2) 22-28.

Amuche, C. I. and Fan, A. F. (2014) *An assessment of item bias using differential item functioning techniques in NECO Biology conducted examination in Taraba State, Nigeria*, American International Journal of Research in Humanities, Arts and Social Sciences, 6 (1) 95-100.

Andrich, D. and Styles, I. P. (2004) Final report on the psychometric analysis of the early development instrument (EDI) using the Rasch Model, A technical paper commissioned for the development of the australian early development instrument (AEDI), Murdoch University, Australia.

Bond, T. G. and Fox, C. M. (2013), Applying the Rasch Model: Fundamental measurement in human sciences (3rd Ed.),Lawrence Erlbaum Associates, Mahwah, New Jersey.

Brentari, E. and Golia, S. (2007) *Unidimensionality in the Rasch Model: How to detect and interprete*. Statistica, 67 253-261.



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

Chong, H. (2011). A simple guide to item response theory (IRT) and Raschmodeling, Available at http;//www.creative-wisdom.com, Retrieved January 4, 2012.

Cavanagh, R. F. and Romanoski, J. T. (2006) *Rating scale instruments and measurement*, Learning Environments Research, 9 (3) 273-289.

De Champlain, A. (2010) *A primer on classical test theory and item response theory for assessments in medical education*, Medical Education, 44 109-117.

Embretson, S. E. (2007) Mixed Rasch models for measurement in cognitive psychology,In M. Von Davier, and C. H. Carstensen (Eds.), Multivariate and mixture distribution Rasch Models, Extensions and applications, Springer Verlag, New York.

Green, K. E. and Frantom, C. G. (2002) Survey development and validation with the Rasch Model, A paper presented at the international conference on questionnaire development, evaluation, and testing, Charleston.

Hagquist, C. and Andrich, D. (2004) *Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Raschmodeling*, Personality and Individual Differences, 36 955-968.

Herrmann-Abell, C. F. and DeBoer, G. E. (2011) *Using Distractor-driven standards-based multiple-choice assessments and Raschmodeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items*, Chemistry Education Research and Practice, 12 184-192.

Joshua, M. T. (2005) Fundamentals of Test and Measurement in Education, University of Calabar Press, Calabar, Nigeria.

Khairani, A. Z., Razak, N. A. and Shamsuddin, H. B. (2012) *Modelling students' mathematical ability and items' difficulty parameters, Application of the Rasch measurement model*, International Journal of Scientific and Engineering Research, 3 8 2229-5518.

Kubinger, K. D. (2005). *Psychological test calibration using the Rasch model, Some critical suggestions on traditional approaches*, International Journal of Testing, 5 377-394.

Linacre, J. M. (2010) A user's guide to Winsteps, Available at http://www.winsteps.com/winman/index.htm?guide.htm, Retrieved December, 2010.

Linacre, J. M. and Wright B. D. (2004) Winsteps: Multi-choice, rating scale, and partial credit Rasch analysis,MESA Press, Chicago.

Nopiah, Z. M., Osman, M. H., Razali, N., Ariff, F. H. M. and Asshaari, M. F. (2010) *How good was the test set up? From Rasch analysis perspective*, International Journal of Scientific and Engineering Research, 2 (4) 116-129.

Obinne, A. D. E., Nworgu, B. G. and Umobong, M. E. (2013) *An investigation into differential item functioning of test conducted by the two major examinations in Nigeria*, Advances in Educational Research, 2 (1) 1-8.

Opasina, O. C. (2009) Development and validation of alternative to practical physics test using item response theory model, An Unpublished Ph.D Thesis, University of Ibadan, Ibadan, Nigeria.

Phillipson, J. N. (2008) *The optimal achievement model and underachievement in Hong Kong, An application of the Rasch Model*, Psychology Science Quarterly, 50 (2) 147-172.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. MESA Press, Chicago.

Smith, E. V. (2004) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals, In E. V. Smith and R. M. Smith (Eds.), Introduction to Rasch Measurement, JAM Press, Minnesotea.

Smith, S. R. and Plackner, C. (2009)*The family approach to assessing fit in Rasch measurement, Journal of Applied Measurement*, 10 (4) 424-437.



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)

Soutar, G. N. and Garry, M. (2001)Raschmodeling: An alternative to summated scales, ANZAM, Auckland.

Vigneau, F. and Bors, D. A. (2005) *Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices*, Educational and Psychological Measurement, 65 (1) 109-123.

Wu, M. and Adams, R. (2007) Applying the Rasch model to psycho-social measurement, A practical approach, Educational Measurement Solution, Melbourne.

Wyse, A. and Mapuranga, R. (2009) *Differential item functioning analysis using Rasch item information functions*, International Journal of Testing, 9 333-357.



Print ISSN: ISSN 2054-6351 (print), Online ISSN: ISSN 2054-636X (online)