APPLICATION OF MANY-FACET RASCH MEASUREMENT IN COMPLEX PROBLEM SOLVING SKILLS IN 2019 NECO PHYSICS ESSAY ITEMS

ADEOSUN, Praise Kehinde, EKWERE, Ndifreke Soni,

Department of Educational Evaluation and Counselling Psychology,
University of Benin, Edo State, NIGERIA.

praise.adeosun@uniben.edu, ndifreke.soni@uniben.edu

ABSTRACT

The study sought to examine application of Many-Facet Rasch Measurement in Complex Problem Solving Skills in 2019 NECO Physics Essay Items. This became necessary because severity / leniency, halo effect and central tendency error can affect scores which need to be satisfied before proper judgement can be made when evaluating student's scores.

The descriptive survey design was employed in this study. The population of the study covers public senior secondary school physics students and physics teachers in Uyo Local Government Area, Akwa Ibom state. 50 physics students and 10 physics teachers were sampled using multistage sampling technique for effective selection. The instrument used in this study was 2019 NECO physics easy questions. However, the reliability coefficient 0.86 was obtained.

The findings revealed that the extent of in test scores among raters in the assessment of complex problem-solving skills in 2019 NECO Physics Essay Items was low. Also, there was no halo effect committed among the raters when rating complex problem-solving skills in 2019 NECO Physics Essay Items. We concluded that the extent of severity / leniency exist among raters was low while halo effect was not committed by raters in the rating of complex problem-solving skills in 2019 NECO Physics Essay Items. It was recommended that Raters should be strictly monitored by the head teacher, Principals and the ministry of education during rating process and also, raters should be given adequate training to be aware of these errors and possible way to reduced or avoid these errors.

Keywords: Rasch Measurement, severity, leniency, halo effect

INTRODUCTION

Background of the Study

Physics has features that are generally accepted and believed to widen the knowledge and increase the level of understanding of the learners. In other to achieve learning efficiency and effectiveness in physics, the teaching should be guided discovery so as to enable the student solve problems on their own.

On a yearly basis, National Examination Council (NECO) conducts Senior Secondary Certificate Examination (SSCE) for students all over the nation. NECO, an examination body in Nigeria that conducts the SSCE for candidates in their third year of the Senior Secondary School, this examination provides opportunity for students to transit from the Senior Secondary level three to tertiary institution. Physics is one of the core science curricular subjects taught at senior school level in Nigeria. Different subsets of the universe of items representing physics as well as other subjects are sampled, responses of students are scored based on their abilities and certificates are awarded. This examination is administered after the physics curriculum has been covered, which the students might have gone through proper learning process in the subject area.

Learning process is unique to each learner, some learn faster, some scanning information and mastering skill seems effortless, others require several exposures over a long period to grab skill while some learn best through text. Therefore, learning styles are unique as the personalities varies.

The process of changing in person’s behavior is refers to as learning. Change occur due to development of new skills, understanding certain scientific theories, changing an attitude and embracing new ideas towards concepts. Change do not occur accidentally or in it natural state as the body system changes and people grow older. Learning is a change that occur intentionally when attending a course, reading books or paper, learning automatically takes place. In the learning process, learners or students acquired abilities such as analyzing, classifying, comparing & contrasting, describing, evaluating, explaining, complex problem solving and so on.

Complex problem-solving skills is said to be an important skill for citizen of todays’ changing technological society. Physics context presents a great opportunity for students to engage in problem solving. Complex problem-solving skills are considered a primary goal in physics instruction for both secondary and university physics courses. Problem solving is the process of moving toward a goal when the path to the goal is not known or uncertain.

Therefore, complex problem-solving skills are those steps or strategies taken to aids attain a clearly conceived aim or goal in solving problems. Complex problem solving also depends on a student’s prior experience, as some skills can be learned well enough that they become automatic and require minimal effort. The degree of well-defined or ill – defined problem depends on an individual expertise, therefore solvers differ in their problem-solving approaches and skills.

A good understanding of the concept of physics can be a reference for students to solve various problems that exist on the subject of physics. The problem that students often encounter is the difficulty in interpreting concrete and abstract concepts of physics. Students become unable to track the right solution to solve the problem. Also, the physics issues presented today contain high level thinking skills. Students then have difficulty in identifying, understanding, and analyzing the physical difficulties they face. The ability to identify and analyze concept is necessary to solve a physics problem.

Many-Facet Rasch Measurement (MFRM) belong to a whole family of models that have their roots in the dichotomous Rasch model. Rasch models share assumptions that set them apart from other psychometric approaches often used for the analysis and evaluation of tests and assessments. Many-facet Rasch measurement refers to a class of measurement models suitable for a simultaneous analysis of multiple variables potentially having an impact on assessment outcomes.

Many-facet Rasch measurement models, also known as facets models, incorporate more variables, or facets (such as raters, scoring criteria and tasks), than the two that are included in a classical testing situation that is examinees and items.

The many-facet Rasch measurement model is an extension of the Master’s partial credit model that makes possible the analysis of data from assessments that have more than the traditional two facets associated with multiple-choice tests. The Many-facet Rasch

Measurement Model is basically an additive linear model that translate the ordered observations to a logit scale, making calibration of all parameters on the same equal – interval scale possible. The Many-facet Rasch Measurement Model has some assumptions and requirements (Downing, 2003). It is unidimensional, that is all assessment criteria on a rating scale should measure the sa4me single underlying variable, or construct. The Many-facet Rasch Measurement Model is invariant, this means that each facet can be separated out and estimated independently of other facets to determine how various facets are functioning as intended. In other words, the invariance properties of the Facets model are only achieved when there is good model-data fit.

Raters can introduce various sources of variance into performance ratings that are associated with their own rating behavior and not with the actual performance of the ratee. According to Scullen, Mount and Goff (2000) the sources of errors are called rater effect while Myford and Wolfe (2003) called it rater errors. These errors are important sources of construct irrelevant variance in the assessment results. These errors are leniency, severity, halo effect and central tendency effect (Myford & Wolfe, 2003).

Leniency has been a consistent tendency of a rater to give examinees higher ratings than what they should receive on the basis of their actual performance. That is to say when the teacher (rater) tends to assign higher score (rating) to examinee (student) when the score (rating) are not earned. For example, when a student (examinee) score is 60% and the teacher (rater) award 75% to the student. At the other hand, severity occur when the teacher (rater) tend to assign low score (rating) to examinee (student) when the score is actually higher. For instant when a student score 50% but the teacher (rater) award 35% to the student.

Halo effect occur when a rater (teacher) allow an individual’s performance on one item or question or trait to influence the performance in other item or question or trait. For example, when an examinee score 70% in question one and score 40% in question two but the teacher (rater) award 60% in question two base on the high score the examinee got in the question one, then halo effect has occurred.

Statement of the Problem

In assessing students’ performance when given complex problem (item), there are unexpected characteristics that affects the scores other than student’s factor and environmental factor. These unexpected characteristics are collectively known as rater errors. Rater errors result from teacher, assessor’s inaccuracy that affect the reliability and validity of the ratings. As a result of this threaten the fairness of the scores awarded to students are threaten.

In order to express the complex problem solving skills in physics, the items are presented in an essay or subjective form. The most compelling disadvantage of subjective items that allow the valid measurement of upper level cognitive behaviours is scoring. The reduction of reliability of the scores is as a result of difficulty associated with objectively scoring the answer to the items.

When measurement involves more than one source of error as in the case of scoring subjective items, Many-facet Rasch Measurement Model is adopted which removes the restriction of the classical test theory.

The error-prone nature of most measurement facets, in particular the fallibility of human raters, raises serious concerns regarding the psychometric quality of the scores awarded to examinees. These concerns need to be addressed carefully, particularly in high-stakes assessments like WACE and NECO the results of which heavily influence examinees’ career or study plans.

Purpose of the Study

The purpose of this study is to detect severity/leniency and halo effect errors among raters in complex problem solving in 2019 NECO Physics Essay Items using Many Facet Rasch Measurement.

Research Questions

First, to what extent does severity/leniency in test score occur among raters in the assessment of complex problem solving skills in 2019 NECO Physics Essay Items?

Second, to what extent does raters produce halo effect when rating complex problem solving skills in 2019 NECO Physics Essay Items?

Significance of the Study

The findings of the study could be beneficial to student because it will reduce score manipulation that occurs as a result of rater error. The study could be great importance to public examination bodies like West Africa Examination Council (WACE), and National Examination Council (NECO) as it will provide them with the specific information about measurement error and ways of reducing these error effect.

METHODOLOGY

Research Design

The research design adopted for this study is descriptive research design based on survey method.

Target Population

The population of this study consist of 207 Senior Secondary Three (SS3) physics students and all the 17 physics teachers in all the public schools in Uyo local Government Area in Akwa Ibom State.

Sample and Sampling Techniques

The sample for study comprised of 50 SS3 physics students and 10 physics teachers from the 10 selected public secondary schools in Uyo. The multi-stage sampling technique was adopted for the study.

Firstly, the simple random sampling was used to select 10 senior secondary three schools representing 70% from 14 schools in Uyo. This was achieved by balloting and replacement of the school picked to ensure a representative sample of the schools chosen.

Secondly, Cluster sampling techniques was used to select all the students offering Physics intact in all the arms of SS3 due to the limited number of physics students in SS 3. Any class or group of similar characteristics in a cluster.

Data Collection Instrument

The instrument used for data collection is the 2019 NECO Physics Easy items.

Validity and Reliability of the Instrument

The easy items were validated by the Examination Development Department of NECO. The reliability of the instrument was established using a sample of 20 SS3 students that were not part of study population. The reliability of the instrument was determined using fit statistics in Winsteps version 5.1.1.0 package to obtain a reliability coefficient of 0.86.

Data Collection

The instrument for data collection was administered to the physics students and collected by researcher with the assistance of invigilators who were physics class room teacher.After administrating the test, the responses were retrieved from the students. 10 copies of those all the responses were made, each copy (50 scripts) were given to physics teachers (raters) for scoring and rating of student's performance. 5-point scale were used. 0-39 'poor' (1), 40 – 59 'fair' (2), 60– 79 'Good' (3), 80-89 'very good' (4) and 90 – 100 'excellence (5). After scoring and rating of the test by the physics teachers, the researcher retrieved all the student's responses from the raters (physics teachers) for proper analysis.

Data Analysis

The data collected was analyzed using Winsteps version 5.1.1.0 software for FACET. Research questions 1and 2 were answered using descriptive statistics (variable map, and person statistics).

RESULTS

Research Question 1: To what extent does severity/leniency in test score occur among raters in the assessment of complex problem solving skills in 2019 NECO Physics Essay Items?

In analyzing MFRM, the first information that is usually checked is the variable map (see figure 1), which display every facet on a ruler-like variable map. The facets display a table that provides the individual rater severity / leniency measures (logits) and the standard error of each severity estimate indicating the precision with which a rater's severity / leniency has been measured. Raters are ordered in a variable map in terms of levels of severity / leniency each exercised. Most severe rater appear at the top of column 3, while the most lenient rater appear lower in the same column. The rater severity values appear in the rater column, with rater 5 being the most severe (2.84 logits) which is two standard deviation above the mean and rater 9, the most lenient (-2.30 logits) which is two standard deviation below the mean. Each asterisk (X) represents 2 individuals.

```
MEASURE       ITEM - MAP -     PERSON
                        <rare>     |    <more>
   6                             +
                   X |
                       |
  5                          +
                       |
                           X      | T
  2                             T+
                       |
                       |
 3                   +      R5 male   1
                             |      R4 male   1
                             | S    R10 male   1
 2                             S+   R1 female 2
                                  |R6 female 2
                   XXXX      |
       1                         +  R8 male   1
                        XX   | M
                      XXX    |
       0              XXX M+
```

ISSN: 2186-845X ISSN: 2186-8441 Print
www.ajmse. leena-luna.co.jp

```
            XX  |
          XX    |   R2 male   1
-1              X  + S     R7 male   1
           X   |    R3 male   1
                X      |
-2        XX S+
                |   R9 female 2
            |
-3              X   +T
            |
                X    |
-4                 T+
         <freq>|<less>
```

Figure 1. Variable map for raters, examinee and scale facets.

Table 1: Descriptive Person Statistics showing the extent of severity/leniency in test score among rater in the assessment of complex problem solving skills in 2019 NECO Physics Essay Items.

Rater (R)	Average total score	Total count	Measure(logits)	Model Standard Error
1	81.0	50	1.99	0.39
2	65.0	50	-0.72	0.43
3	62.0	50	-1.27	0.44
4	85.0	50	2.57	0.37
5	87.0	50	2.84	0.36
6	79.0	50	1.68	0.40
7	63.0	50	-1.09	0.43
8	75.0	50	1.03	0.41
9	57.0	50	-2.30	0.48
10	84.0	50	2.43	0.38

Table 1 shows that rater 5 is the most severe rater with a logit of 2.84 and standard error of 0.36. this is followed by raters 4, 10, 1, 6, and 8 with a logit values of 2.57, 2.43, 1.99, 1.68, and 1.03 respectively.as well as their corresponding standard error value of 0.37, 0.38, 0.39, 0.40, and 0.41 respectively. The grand mean of 2.09 logits and the grand mean standard error of 0.39 was obtained. It was observed that, the lager the measure (logit), the more severe the rater. Also, as the standard error value get smaller, the more severe the rater. From the grand mean, it implies that extent of rater severity was low. This is because the mean logits of 2.09 was one standard error above the mean.

Table1 also shows that rater 9 is the most lenient rater with the logit (measure) of -2.30 and standard error of 0.48 below the mean. Followed by rater 3, 7, and 2 with measure (logit) of -1.27, -1.09 and -0.72 respectively and standard error value of 0.44, 0.43, and 0.43 respectively. These values gave the grand mean logit of -1.35 and grand mean standard error of 0.45 respectively. This implies that raters leniency was low because it was one standard error below the mean. Therefore, the extent of severity/leniency in test score among raters in the assessment of complex problem solving skills in 2019 NECO Physics Essay Items was low.

Research Question 2: To what extent does raters produce halo effect when rating complex problem solving skills in 2019 NECO Physics Essay Items?

Table 2: Category Statistics that shows the extent raters produce halo effect when rating complex problem solving skills in 2019 NECO Physics Essay Items.

Rater (R)	Measures (logits)	Standard Error	Infit Mean Square (MNSQ)
3	-1.27	0.44	1.38
2	-0.72	0.43	2.42
7	-1.09	0.43	0.92
9	-2.30	0.48	0.91
10	2.43	0.38	0.57
6	1.68	0.40	0.69
8	1.03	0.41	0.54
4	2.57	0.37	0.66
1	1.99	0.39	1.11
5	2.84	0.36	0.48

In order to detect halo effect error from table 2, we used the infit mean square value. 0.8 – 1.2 indicate that no halo effect which means that the value fit into the model. O.4-0.6 indicates that the presence of halo effect is mild while 0.3 and above infit mean square indicate a severe halo effect. From table 5 above, it can be seen that the infit mean square value of rater measures ranged from 0.48 (R5) – 2.42 (R2). Rater 7, 9, and 1 with infit mean square value of 0.92, 0.91, and 1.11 fit into the model indicating that there was no halo effect committed when rating when rating complex problem-solving skills in 2019 NECO Physics Essay Items. Also, table 5 shows that rater 10, 6, 8, 4 and 5 with infit mean square value of 0.64, 0.54, 0.66 and 0.48 respectively shows that the extent of halo effect error committed by the raters was mild. The grand mean of Infit Mean Square for rater 10, 6, 8, 4 and 5 is 0.96. this indicate no halo effect. Therefore, there was no halo effect among the raters when rating complex problem-solving skills in 2019 NECO Physics Essay Items.

DISCUSSION OF FINDINGS

The finding research question one revealed that the extent of severity/leniency in test score among raters in the assessment of complex problem-solving skills in 2019 NECO Physics Essay Items was low. This is to say, when ratings rely on unadjusted raw score, students who were rated by lenient raters would have an unfair advantage over those who were rated by severe raters.

This study was in accordance with Wolfe, Myford, Engelhard, and Manalo (2007) they found out that only 5% of raters become more severe over the rating period, while 16% of raters became more lenient over the rating period. Thus, it appears that raters' leniency and severity may change over time, and the direction of the change may not always be predictable.

Smith and Kulikowich (2016) were not in support of the study as two judges were used to provide information concerning the judge facet. The fit statistics address intra judge consistency over the other facets. Sebok and MacMillan (2014) is in line with the study. Five raters participated in this study: 3 faculty members and 2 students. The most severe rater (R1) had a measure of 0.15 and the most lenient rater (R5) had a measure of –0.16, producing a spread of ±0.16 logits, which is roughly one third of a logit difference between the most lenient and most severe rater. The Rasch analysis showed that the most severe rater (0.15 logits) was a new faculty member who had a counselling background but no previous experience with the admissions process at this particular institution.

The finding of research question two revealed that there was no halo effect among the raters when rating complex problem-solving skills in 2019 NECO Physics Essay Items. This finding disagrees to Engelhard (1994) who found that two out of his 15 highly trained raters displayed halo effect, but he failed to explain why such halo effect occurred with two of his so highly trained raters. Comparing generalizability theory with many-facet Rasch measurement in determining halo effect, Kozaki (2004) also found that two out of his four professionally and bilingual judges showed signs of halo effect on two categories of grammar and vocabulary. She attributed this halo effect to powerful roles these two categories played in the assessment and judges carried over the impression of competence.

In the same vein, the study is in support of Knock, Read, and von Randow (2007) attributed halo effect to lack of training and feedback to her raters because after training and feedback at least some of the raters did not show halo effect in the face-to-face group, but halo effect remained with raters in the online group even after training and feedback.

Also, the study was not in line with Yorozuya and Oller, (2009). Authors indicated that all 15 raters showed halo effect, according to authors, it could be counted for all conditions under which raters were rating. There were two conditions and the raters in condition one rated all four scales; grammar, vocabulary, pronunciation, and fluency at a single hearing rather than at separate hearings and this led to higher intercorrelations of scales, hence the appearance of halo effect.

In line with this study, Farrokhi and Esfandiari (2011) used 188 undergraduate Iranian English majors to rate the essays their classmates had written. The peer-assessors used a 6-point analytic rating scale to assess the essays on 15 assessment criteria. The results of Facets analysis showed that the individual peer-assessors might display halo effect while the group raters did not detect halo effect.

CONCLUSIONS

Based on the findings, it is concluded that raters were not only severe when assessment of complex problem-solving skills but they are also lenient in their rating process. Also, raters did not commit halo effect when rating complex problem-solving skills in 2019 NECO Physics Essay Items. These rater errors can introduce construct-irrelevant variance into students’ scores and weaken the validity of inferences about their complex problem-solving skills.

REFERENCES

[1]. Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna Eds, *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. Mahwah, NJ: Erlbaum. 261–287.

[2]. Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531-1540.

[3]. Knoch, U., Read, J. & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43.

[4]. Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*(1), 1–27.

[5]. Myford, C., & Wolfe, E. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, *4*, 386–422.

[6]. Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956-970.

[7]. Sebok, S. S., & MacMillan, P. D. (2014). Assessment of a Master of Education Counselling Application Selection Process Using Rasch Analysis and Generalizability Theory. *Canadian Journal of Counselling and Psychotherapy*, *48*(2). https://dev.journalhosting.ucalgary.ca/index.php/rcc/article/view/60970.

[8]. Smith, E. V., & Kulikowich, J. M. (2016). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Journal of Educational and Psychological Measurement*, *64*, 617–639.

[9]. Wolfe, E. W., Myford, C. M., Engelhard, G., & Manalo, J. R. (2007). Monitoring reader performance and DRIFT in the AP English Literature and Composition examination using benchmark essays (College Board Research Report). New York, NY: College Board.

[10]. Yorozuya, R. & Oller, J. W. (2009). Oral proficiency scales: Construct validity and the halo effect. *Language Learning*, *30*(1), 135-153.