British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 7, Issue 2, 2024 (pp. 1-10)
www.abjournals.org

DETECTING RANDOMNESS EFFECT AMONG RATERS IN PHYSICS ESSAY ITEMS USING MANY-FACET RASCH MEASUREMENT

Adeosun Praise Kehinde (Ph.D.)[1] and Ekwere Ndifreke Soni (Ph.D.)[2]

[1]Department of Educational Evaluation and Counseling Psychology, University of Benin, Edo State, Nigeria.
E-mail: praise.adeosun@uniben.edu, Tel.: +234-8060481557

[2]Department of Educational Evaluation and Counselling Psychology, University of Benin, Edo State, Nigeria.
Email: ndifreke.soni@educ.uniben.edu

Cite this article:

Adeosun P. K., Ekwere N. S. (2024), Detecting Randomness Effect among Raters in Physics Essay Items using Many-Facet Rasch Measurement. British Journal of Education, Learning and Development Psychology 7(2), 1-10. DOI: 10.52589/BJELDP-XPOKQLJY

Manuscript History

Received: 13 Nov 2023

Accepted: 21 Feb 2024

Published: 18 Mar 2024

ABSTRACT: *This study sought to detect randomness effects among raters in physics essay items using Many-Facet Rasch Measurement. The research design adopted for this study is descriptive research design based on survey method. The population of the study comprised eighty-eight (88) public schools in all the local government areas with a physics student population of 3,642 students and ninety-four (94) physics teachers in all the Senior Secondary Schools in Uyo Senatorial District for the 2022/2023 academic session. Three hundred and sixty-four (364) SSS3 physics students and 37 physics teachers from the 31 selected secondary schools in Uyo Senatorial District were sampled using multistage sampling technique for effective selection. The multi-stage sampling technique was adopted for the study. The instrument used for data collection was Physics Achievement Test (PAT) obtained from WAEC and NECO 2020 Physics Essay items with reliability coefficients of 0.91 and 0.90 respectively. The finding revealed that most of the individual raters commit randomness effect when rating physics essay Items. It also revealed there is a significant difference at the rater's group level exhibiting randomness effect when rating physics essay items, which implies that there is no group-level randomness effect present among raters when rating physics essay items. We concluded that rater effects are sources of variance in performance ratings that are associated with the raters' behaviour and not the actual performance of the ratee. It was recommended that raters should follow the rating guidelines to reduce the impact of randomness in their ratings and provide more accurate and objective evaluations.*

KEYWORDS: Randomness effect, Raters, Many Facet-Rasch Measurement.

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

British Journal of Education, Learning and Development Psychology
ISSN: 2682-6704
Volume 7, Issue 2, 2024 (pp. 1-10)

www.abjournals.org

INTRODUCTION

Many Facet-Rasch Measurement (MFRM) is a model-based psychometric analysis that has a broad range of applications in performance assessment. It is an extension of the basic Rasch one-parameter item response theory model, a class of psychometric models used to estimate examinees' ability and the difficulty of test items on the same scale (Downing, 2003). The basic Rasch model uses only one parameter, item difficulty, to estimate the examinees' ability. MFRM extends the basic Rasch model by adding parameters describing facets of measurement interest other than item difficulty (such as rater severity or task difficulty) to the model (Linacre & Wright, 2004). Thus, it allows researchers to obtain measures of examinees' levels of ability while controlling for variability in rater severity, task difficulty, or any other facet of measurement (Linacre & Wright, 2004). Myford and Wolfe (2003) categorized five major types of rater effects: leniency/severity, randomness, halo, and central tendency.

Performance ratings may contain a variety of sources of variance that have more to do with the rater's own rating practices than the ratee's actual performance. The assessment results' construct-irrelevant variance is largely a result of these inaccuracies. Rater mistakes can have a variety of causes. Different causes lead to various grading patterns, which call for various treatments. Researchers must comprehend the many types of rater errors in order to create successful control measures. Myford and Wolfe (2003) categorized five major types of rater effects: leniency/severity, randomness, halo, and central tendency.

The randomness effect is commonly known as inconsistency. The randomness effect is a rater's tendency to apply one or more trait scales in a manner inconsistent with how the other raters apply the same scales. When a rater awards a score of 25% or 15% to an examinee in an item that other raters award 20% to the same item, then it can be said that the rater is inconsistent in the rating scale. Rater inconsistency describes the variation or lack of uniformity in the assessments provided by various raters when assessing the same subject or item. This discrepancy may be caused by variables including subjective interpretations, variations in how well raters grasp the rating criteria, and levels of expertise. Inconsistency among raters might produce incorrect and inaccurate outcomes in educational assessment.

Under the MFRM framework, there are three ways to evaluate the fit between data and model: Global model fit, group-level fit statistics, and individual-level fit statistics. For global model fit, a log-likelihood chi-square χ^2 (sum of natural logarithms of the model probabilities of all observations), which approximates df = (number of responses used for estimation) – (number of parameters estimated)) is typically output from MFRM analysis (Eckes, 2015). This is known as Fixed-effect Chi-Square. The fixed-effect chi-square is a significance test used to test the null hypothesis that there are no differences in the logit values for an object of measurement (for instance student, and rater), after controlling for measurement error (Eckes, 2015).

The fixed-effect chi-square is defined as: $X^2 = \sum (w_o * D_o^2) - \frac{(\sum Wo * Do)^2}{\sum Wo}$

where D_o is the estimated logit of the object of measurement (i.e., rubric element, severity/leniency of rater, or student ability) and $w_o = \frac{1}{SE^2{}_O}$. Degrees of freedom equal D – 1, where D = the number of observations of the object of measurement. Note that the fixed-effect chi-square is sensitive to sample size. Thus, in large samples, the fixed-effect chi-square may

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

be statistically significant, even with small differences in the object of measurements' logits (Eckes, 2015).

The extent to which the observed ratings match or deviate from the expected ratings generated by the MFRM can be evaluated either globally (for a group of raters) or individually (for individual raters). Eckes (2015) provided detailed calculations for the global fit indices and individual fit statistics related to the rater facet. For example, the rater separation ratio measures the spread of rater severity estimates relative to their precisions. For a particular rater j, the rater separation ratio: $G_J = \frac{SD_{t(g)}}{RMSE_J}$

where $SD^2_{t(J)} = SD^2_{0(J)} - MSE_J$

and

$MSE_j = \sum_{J=1}^{J} \quad St_j^2$

The mean-square error (MSE) is the average of the standard errors estimated for each rater j and the true variance of the severity estimates equals the observed variance minus the MSE. The rater separation ratio is formed by taking the square root of the ratio between the true variance and MSE. The higher the separation rater, the more spread the rater severity measures.

According to a systematic review of methodologies applied in different areas of rater studies, Wind and Peterson (2018) argued that, to inform the interpretation and use of rating scores and improve the quality of rater-mediated assessment, the rating quality indices should go beyond group-level indicators or inter-rater reliability to provide individual-specific information, and incorporate diagnostic information from other facets of the assessment. MFRM offers individual-specific information about raters based on standardized residuals or the differences between observed and expected ratings. Suppose X_{nij} is the observed rating for examinee n evaluated by rater j on criterion i, and e_{nij} be expected rating based on the MFRM model's parameter estimates, the standardized residual, in this case, can be expressed as:

$Z_{nij} = \frac{X_{nij} - e_{nij}}{W_{nij}^{\frac{1}{2}}}$

where $e_{nij} = \sum_{k=0}^{m} \quad KP_{nijk}$

P_{nijk} is the probability of person n, when rated on item i by judge j being awarded a rating of k,

and $W_{nij} = \sum_{k=0}^{m} \quad (K - e_{nij})^2 P_{nijk}$

Squaring the standardized residuals averaging over the elements of the other facets (for instant examinees and tasks) for each rater yields the residual-based indices of data-model fit, which takes the form of Mean Squared Error (MSE) fit statistics that are asymptotically distributed as scaled chi-square statistics divided by their degrees of freedom (Eckes, 2015). The unweighted MSE fit statistic for rater j averaged overall all examines n = 1, N and criteria i = 1, I can be obtained by:

$MS_{u(j)} = \frac{\sum_{n=1}^{N} \quad \sum_{i=1}^{I} \quad Z_{nij}^2}{N.I}$

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

The unweighted MSE fit statistic calculated above is also called outfit statistic (short for outlier sensitive fit statistic). An example of an outlying situation can be a severe rater assigning a lenient rating to a highly proficient examinee on a medium difficulty criterion, which will increase the outfit statistic. Weighting the Z_{nij} by the model variance W_{nij} results in the weighted MSE fit statistic:

$$MS_{w(j)} = \frac{\sum_{n=1}^{N} \sum_{i=1}^{I} W_{nij} Z_{nij}^2}{\sum_{n=1}^{N} \sum_{i=1}^{I} W_{nij}}$$

This statistic is also called infit statistic (information weighted fit statistic) because it is sensitive to "inlying" unexpected responses or situations where the location of the rater is close to those of the other facets on the measurement scale. The infit statistic usually has higher estimation precision and is considered more important than outfit statistic (Linacre, 2002; Myford & Wolfe, 2003, in Eckes, 2015). The outfit and infit MSE statistics can be used as a diagnostic tool to evaluate the extent to which the ratings assigned by a particular rater match or deviate from the model's expectations because they both have an expected value of 1.0 and range from 0 to +∞. Raters with fit values greater than 1.0 show more variation than expected in their rating; this is called misfit (or underfit). By contrast, raters with fit values less than 1.0 show less variation than expected, indicating that their ratings are too predictable or provide redundant information; this is called overfit (Eckes, 2015).

Statement of the Problem

The written responses of essay physics items are far more complex than responses to multiple-choice items, and are traditionally scored by raters. Raters typically gauge an essay's quality aided by a scoring rubric that identifies the characteristics an essay item must have to merit a certain score level. Due to the fact that some raters might lack cognitive process of the information given in student's responses while some raters may connect students' responses with their prior knowledge that is not in the marking guide based on their understanding of the content as a result introduced error to students' feedback, thereby increasing the impact of inconsistency in scores.

When physics raters are not accurate in identifying and assessing the strengths and weaknesses in students' essay items, the feedback given may not be helpful or constructive. This can impede students' learning and growth in problem-solving skills especially in physics. It is therefore necessary to apply the many-facet Rasch measurement model in detecting rater errors such as randomness effect in rating of physics essay items if objectivity and reliability of scores is to be obtained.

Purpose of the Study

The purpose of this study is to detect randomness effect among raters in Physics Essay Items using Many Facet Rasch Measurement.

Research Questions

First, to what extent does the individual rater produce a randomness effect when rating physics essay items in Uyo Senatorial District?

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

Second, is there a difference in randomness at the rater's group level when rating physics essay items in Uyo Senatorial District?

Hypothesis

There is no significant difference at the rater's group level exhibiting randomness effect when rating physics essay items in Uyo Senatorial District.

Significance of the Study

The results of research may be useful to students as they reduce manipulation of results caused by rater error. The finding could be beneficial to test constructors as students' performance could be enhanced positively by detecting multiple errors caused by the rater, which will increase the reliability of the scores.

The result of this study will be beneficial to psychometricians by expanding the knowledge of estimating individual elements' invariant calibrations across a level of facets such as individual raters and demographic subgroups. It will also provide information on particular rating patterns employed by the raters in evaluating students' responses.

METHODOLOGY

Research Design

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

Target Population

The population of the study comprised eighty-eight (88) public schools in all the local government areas with a physics student population of 3,642 students and ninety-four (94) physics teachers in all the senior secondary schools in Uyo Senatorial District for the 2022/2023 academic session.

Sample and Sampling Techniques

A sample of three hundred and sixty-four (364) senior secondary school three (SS3) physics students and thirty-seven (37) physics teachers from the thirty-one (31) selected secondary schools in Uyo Senatorial District was used for the study. The multi-stage sampling technique was adopted for the study.

Firstly, proportionate sampling was used to select 31 schools representing 35% of the total number of schools in each of the local government areas used in this study.

Secondly, simple random sampling was used to select 364 physics students representing 10% of the physics students from the 31 schools in the Uyo senatorial district.

Thirdly, stratified sampling technique was used to obtain the sample size of 37 physics teachers in 31 schools. This was done by dividing the population into strata based on gender, and then randomly sampling 40% of each of the strata.

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

Data Collection Instrument

The instrument used for data collection was Physics Achievement Test (PAT) obtained from WAEC and NECO 2020 Physics Essay items.

Validity and Reliability of the Instrument

The easy items were validated by the Examination Development Department of WAEC and NECO. The reliability of the instruments was established using a sample of 30 students from public Senior Secondary (SS 3) who were not part of the sample but of the main study population. The reliability of NECO and WAEC was determined using fit statistics in Winsteps version 4.8.2.0 package to obtain a reliability coefficient of 0.91 and 0.90 respectively.

Data Collection

The researcher administered the instrument (Physics Achievement Test) to SSS 3 students with the assistance of their physics teachers in the selected schools. The duration for the test was 1 hour 30 minutes. After administering the test, the responses were retrieved from the students. The students' responses were rated by the 37 physics teachers. A 6-point scale was used. 0–39 'fail' (1), 40–44 'fair' (2), 45–49 'pass' (3), 50–59 'good' (4), 60–69 'very good (5), and 70 and above 'excellent' (6). After scoring and rating the test by the physics teachers, the researcher retrieves all the student's responses from the raters (physics teachers) for proper analysis.

Data Analysis

The data collected were analyzed using Winsteps version 5.1.1.0 software for FACET. Research questions 1 was answered using descriptive statistics while hypothesis 1 was tested using fixed Chi-Square at the alpha level of 0.05.

RESULTS

Research Question 1: To what extent does the individual rater produce a randomness effect when rating physics essay items in Uyo Senatorial District?

Table 1: Category Statistics That Show the Extent to Which Individual Raters Produce Randomness Effect When Rating Physics Essay Items

Rater (R)	Measure(logits)	Standard Error	Infit Mean Square (MNSQ)	Outfit Mean Square (MNSQ)
1	-1.27	0.54	1.21	1.14
2	-3.07	0.55	2.44	2.33
3	-3.68	0.55	1.36	4.68
4	-0.99	0.53	0.85	0.84
5	-0.72	0.52	0.48	0.44
6	-1.86	0.55	1.02	0.98
7	-4.68	0.58	0.81	0.67
8	-1.27	0.55	0.36	0.30
9	-5.85	0.71	0.59	0.38
10	-0.72	0.52	0.38	0.34
11	-1.27	0.54	1.21	1.14

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

12	-3.07	0.55	2.50	2.33
13	-3.68	0.55	1.36	4.68
14	-0.99	0.53	0.85	0.84
15	-0.72	0.52	0.48	0.44
16	-1.86	0.55	1.02	0.98
17	-4.68	0.58	0.81	0.67
18	-1.27	0.55	0.36	0.30
19	-5.85	0.71	0.59	0.38
20	-0.72	0.52	0.38	0.34
21	-1.27	0.54	1.21	1.14
22	-3.07	0.55	2.51	2.33
23	-3.68	0.55	1.36	4.68
24	-0.99	0.53	0.85	0.84
25	-0.72	0.52	0.48	0.44
26	-1.86	0.55	1.02	0.98
27	-4.68	0.58	0.81	0.67
28	-1.27	0.55	0.36	0.30
29	-5.85	0.71	0.59	0.38
30	-0.72	0.52	0.38	0.34
31	-1.27	0.54	1.21	1.14
32	-3.07	0.55	2.53	2.33
33	-3.68	0.55	1.36	4.68
34	2.94	0.51	0.42	0.40
35	-0.72	0.52	0.48	0.44
36	-1.86	0.55	1.02	0.98
37	-4.68	0.58	0.81	0.67

In the randomness effect, we consider the Infit Mean Square and Outfit Mean Square. The Infit and Outfit Mean Square greater than 1 indicate the presence of randomness. The result from Table 1 shows raters 1, 2, 3, 11, 12, 13, 21, 22, 23, 31, 32, and 33 with the Infit Mean Square and Outfit Mean Square values of 1.21 & 1.14, 2.51 & 2.33, 1.36 & 4.68, 1.21 & 1.14, 2.51 & 2.33, 1.36 & 4.68, 1.21 & 1.14, 2.51 & 2.33, 1.36 & 4.68, 1.21 & 1.14, 2.51 & 2.33, and 1.36 & 4.68 respectively. The average total of Infit Mean Square and Outfit Mean Square are 1.49 and 2.75 respectively. These indicate that individual raters committed randomness. Therefore, there was a randomness effect error committed by individual raters when rating complex problem-solving skills in WAEC and NECO 2020 physics Essay Items.

Hypothesis One: There is no significant difference at the rater's group level exhibiting randomness effect when rating physics essay items in Uyo Senatorial District.

Table 2: Group–Level Indices of Randomness Effect in Complex Problem-Solving Skills Between WAEC and NECO 2020 Physics Essay Items

Indices	Value
Ratee Separation Ratio	2.75
Ratee Separation Index	4.0
Ratee Separation Reliability	0.89
Fixed-effect Chi-Square Statistics	37.2 (df=36, P<0.05), Sig. =0.01

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

Table 2 shows a Chi-square value of 37.2 and a P-value of 0.01. Testing at an alpha level of 0.05, the P-value is less than the alpha level; thus, the null hypothesis which states that there is no significant difference at the rater's group level exhibiting randomness effect when rating physics essay items in Uyo Senatorial District is rejected. Therefore, there is a significant difference at the rater's group level exhibiting randomness effect when rating physics essay items in Uyo Senatorial District. This implies that there is no group-level randomness effect present among raters when rating physics essay items.

In addition, Table 2 shows the Ratee Separation Ratio of 2.75. This implies that the indicator did not suggest a group-level randomness effect among raters when rating student scores in physics essay items. Also, Table 2 also shows the Ratee Separation Index of 4.0. This indicator does not suggest a group level of randomness effect among raters when rating student scores in physics essay items. Table 9 also shows a Ratee Separation Reliability of .89. The high degree of ratee separation reliability indicates that there was no group-level of randomness effect among raters when rating student scores in physics essay items.

DISCUSSION OF FINDINGS

The findings from research question one revealed there was a randomness effect error committed by individual raters when rating complex problem-solving skills in WAEC and NECO 2020 physics Essay Items. This may have been as a result of de-motivation of raters during the evaluation process or may not be invested in providing accurate ratings as this can lead to random or inconsistent ratings. The finding is in agreement with the findings of Wang et al. (2020). These researchers investigated rater performance on the Canadian English Language Benchmark Assessment for Nurses (CELBAN) speaking component using a Many-Facets Rasch Measurement (MFRM). The result reveals that grammar, among the eight speaking criteria, was identified as the most difficult criterion on the scale and the one demonstrating the most randomness. The finding is supported by the findings of Tarakol and Pinner (2019) who examined the extent to which the facets modelled in an OSCE can contribute to scoring variance and how they fit into a Many-Facet Rasch Model (MFRM) of OSCE performance. The results did suggest that examiners were lenient and that some behaved inconsistently. The finding was not in agreement with the study of Kondo-Brown (2002) who investigated whether trained native Japanese-speaking (JNS) raters would rate certain types of students and certain criteria more severely or leniently than others in assessing Japanese L2 compositions for norm-referenced purposes such as placement. The result revealed self-consistency in rater bias patterns.

The findings from hypothesis one revealed that there is a significant difference at the rater's group level exhibiting randomness effect when rating physics essay items in Uyo Senatorial District. This implies that there is no group-level randomness effect present among raters when rating physics essay items. This is because raters may have followed the assessment guidelines and specific instructions when rating student scores, thereby reducing the likelihood of arbitrary or random scoring. The Ratee Separation Ratio connotes that the spread of ratee performance measures is 3 times larger than the precision of those measures. Ratee Separation Index implies that there are over 4 statistically distinct strata of ratee performance in the sample of ratees while the Ratee Separation Reliability implies that raters could reliably distinguish among the ratees in terms of their performance.

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

The findings were against the study of Ihli et al. (2016) that compares risk preferences elicited from two different methods and the resulting inconsistency rates in response behaviour. The result reveals significantly different risk results of group raters. However, the study conformed to the findings of Koçak (2020) who investigated rater tendencies and reliability in different assessment methods. The result from the indices reveals the absence of randomness among group raters.

CONCLUSION

Based on the findings, it was concluded that most individual raters committed randomness effects when rating physics essay items. However, at the group-level rating, raters did not commit randomness effects when rating physics essay items. It can be inferred that rater effects are causes of variation in performance evaluations that are connected to the behavior of the rater and not the ratee's actual performance.

RECOMMENDATION

Based on the findings, researchers recommended that raters should follow the rating guidelines to reduce the impact of randomness in their ratings and provide more accurate and objective evaluations.

REFERENCES

Eckes, T. (2005). Examining rater effects in Test DIF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221. https://doi:10.1207/s15434311laq0203_2.

Ihli, H. J., Hiputwa, B. & Musshoff, O. (2016). Do Changing Probabilities or Payoffs in Lottery-Choice Experiments Affect Risk Preference Outcomes? Evidence from Rural Uganda. *Journal of Agricultural and Resource Economics 41(2).* https://www.researchgate.net/publication/303790257_Do_Changing_Probabilities_or_Payoffs_in_Lottery

Koçak, D. (2020). Investigation of Rater Tendencies and Reliability in Different Assessment Methods with Many Facet Rasch Model. International Electronic Journal of Elementary Education, 12(4), 349 – 358.

Kondo-Brown, k. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *SAGE Journal. 19(1*). https://doi.org/10.1191/0265532202lt218oa

Linacre, J. M. (2004). Optimizing rating scale effectiveness. In., E. V., Smith, Jr, & R. M. Smith. (Eds.). *Introduction to Rasch model*. Maple Grove, Minnesota: JAM press, 258-278.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using Many-Facet Rasch measurement: *Part I. Journal of Applied Measurement, 4, 386- 422.*

Tavakol, M. & Pinner, G. (2019). Using the Many-Facet Rasch Model to analyses and evaluate the quality of objective structured clinical examination: a non-experimental cross-sectional design. BMJ Open. 9(9), 1-9. doi:10.1136/bmjopen-2019-029208

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY

Wang, P., Coetzee, k., Strachan, A., Monteiro, S. & Cheng, L. (2020). Examining Rater Performance on the CELBAN Speaking: A Many Facets Rasch Measurement Analysis Canadian. *Journal of Applied Linguistics, Special Issue,* (23), 73-95.

Article DOI: 10.52589/BJELDP-XPOKQLJY
DOI URL: https://doi.org/10.52589/BJELDP-XPOKQLJY