International Journal of Evaluation and Research in Education (IJERE)
Vol. 14, No. 3, June 2025, pp. 1590~1596
ISSN: 2252-8822, DOI: 10.11591/ijere.v14i3.30351

Examining the ‘hawk-dove effects’ in portfolio assessment using the multi-facet Rasch model

Andrews Cobbinah[1], Jephtar Adu-Mensah[2]
[1]Department of Education and Psychology, Faculty of Educational Foundations, University of Cape Coast, Cape Coast, Ghana
[2]National Council for Curriculum and Assessment, Accra, Ghana

Article Info

Article history:

Received Feb 20, 2024
Revised Oct 14, 2024
Accepted Oct 30, 2024

Keywords:

‘Hawk and dove effects’
Many-facet Rasch
Portfolios
Rater behavior
Validity

ABSTRACT

Concerns among students have increased due to the use of test scores in decision-making, leading them to question whether their results accurately reflect their abilities, especially when they perceive subjectivity in rater scoring. This study explores the effects of rater bias on portfolio assessment scores among student teachers in the colleges of education in Ghana. A sample of 207 student portfolios, scored by tutors, was analyzed using a three-facet design model and the FACET software. The findings revealed that tutors exhibited varying rating behaviors, including severity, leniency, and halo effects. These differing rating patterns were found to impact the students’ portfolio scores, suggesting that the subjectivity of raters plays a crucial role in the assessment process.

Corresponding Author:

Andrews Cobbinah
Department of Education and Psychology, Faculty of Educational Foundations, University of Cape Coast
Cape Coast, Ghana
Email: andrews.cobbinah@ucc.edu.gh

1. INTRODUCTION

Portfolio assessment has gained significant popularity in the educational field. Alongside this increase in usage, concerns have been raised by educators, policymakers, and researchers about the potential flaws associated with using human scorers. Although portfolios are used to gather evidence to inform decisions on the professional development of pre-service and in-service teachers [1], the assessment is predominantly carried out by human raters, which introduces subjectivity when grading students’ portfolio work. Rater subjectivity in this study reflects any actions of the rater that systematically influence test or assessment scores [2]. Such actions may include individual biases, understanding of the scoring rubrics, and the influence of extraneous factors such as student behavior, handwriting, or relationship with the student. Raters might unintentionally let their personal beliefs or experiences influence their evaluations, resulting in inconsistent grading practices [3].

The emphasis on test results as the primary measure of student achievement or ability has heightened student concerns [4]. Studies have reported that students frequently express concerns about the assessment process, particularly their grades or test scores [5]. These concerns often stem from the belief that teachers or raters exhibit subjectivity when grading [6]–[8]. In contrast to multiple-choice questions, the scoring of portfolio tasks is subject to various human-related factors that can influence the consistency and reliability of scoring [9]–[11]. These factors include but are not limited to the scoring methods employed by raters [12], [13], the gender and professional backgrounds of raters [14], [15], understanding of the scoring criteria [15], the number of raters involved in the scoring process [16]–[18] and the extent of rater training [19]. Considering the extensive amount of written work generated by students and the inconsistencies in grading practices among raters, there is general concern regarding the fairness, reliability, and validity of portfolio scores [20].

The preference for performance-based assessment coupled with the strong advocacy that education should be student-centered in the 21st century justifies the current changes in educational practices and philosophy [5], [7]. This has made portfolio-based learning and assessment more popular among educators [21]. As a tool for evaluation, it operates on the principles of experiential learning, that enables students to document, reflect on and learn from events. According to Black and Wiliam [20], using a portfolio provides detailed information on learning objectives, resources, and strategies to ensure that learning has taken place. In Ghana, for instance, portfolio assessment has been implemented across all public colleges of education to prepare student-teachers for the constantly evolving world. Student teachers are expected to undergo practical training by visiting a local school at the end of each semester. During these visits, they observe and learn from in-service teachers. After the visit, student teachers are to document their field experience, include relevant artifacts as evidence, and submit this documentation as their portfolio for scoring. Tutors evaluate the portfolios using an analytic scoring rubric. Since human raters are involved in scoring such constructed response tasks, the scores are likely influenced by the raters' subjectivity. In a face-to-face discussion with students from four colleges of education, they expressed concerns about their grades, feeling that their scores did not accurately reflect their performance. Some students were unsure about how the tutors graded them and believed that certain peers were given preferential treatment.

Research on portfolio assessment in Ghana has primarily focused on teachers' assessment practices [22], [23], student teachers' perspectives on portfolio assessment [1], teachers' knowledge of assessment [24], and teachers' grading practices [25]. However, the influence of raters on students' portfolio scores has not been thoroughly explored. Based on this context, this study examines the rater effects in portfolio assessment in the colleges of education in Ghana. The current study examines the diverse rating patterns of tutors when scoring portfolios.

2. THEORETICAL BASIS

2.1. The many-facet Rasch model (MFRM)

The many-facet Rasch model (MFRM) is an extension of Georg Rasch's basic two-facet model of person ability and item difficulty to consider other aspects of assessment settings that could produce a construct-irrelevant variation. The MFRM was created with examiner-mediated evaluations in mind; as a result, an aspect for examiner severity/leniency was introduced to account for examiners' propensity to apply rating scales per their internal norms of relative leniency or severity. Like the Rasch model generates candidate ability estimates independent of the assessment items employed, the MFRM provides candidate ability estimates independent of the assessment items and examiners. Also included in the MFRM is the use of rating scale analysis.

As an extension of the generic Rasch model to account for various aspects of measurement that could generate construct-irrelevant variation, Linacre [26] released his initial work on the MFRM in 1989. Application of the MFRM, for instance, can offer individual estimates of examiner severity, station difficulty, and candidate competence in a typical performance evaluation, such as portfolio scoring. The MFRM facilitates the substitution of real proficiency measurements for observed scores that include bias or another concept of irrelevant variation. In (1) depicts a three-facet MFRM in which candidate ability, rater severity, and item (or station) difficulty are functions of the log of the likelihood of obtaining rating k v. k - 1 on a rating scale.

$$\ln\left(\frac{p_{njk}}{p_{nj(k-1)}}\right) = \theta n - a_J - \delta l - \tau k \qquad (1)$$

The MFRM detects examiners as well as items and candidates that do not follow the predicted pattern of answers contributing to the fundamental assessment of the underlying construct, in keeping with the prescriptive aspect of the Rasch model. This methodology differs from generalizability studies, which are more descriptive and attempt to account for all sources of variation in the data, and choice studies, which offer suggestions on sampling to lessen sources of construct-irrelevant variance [26]. The MFRM detects examiners, items, and candidates that do not follow the predicted pattern of answers contributing to the fundamental assessment of the underlying construct, in keeping with the prescriptive aspect of the Rasch model.

2.2. The hawk-dove effect

One potential vulnerability of portfolio assessment is the variation among examiners in their level of leniency or strictness [27]. This phenomenon, often referred to as the 'hawk-dove' effect, entails hawks being inclined to fail most candidates due to exceptionally high standards, while doves tend to pass most candidates. The hawk-dove nomenclature is widely used in literature to distinguish between lenient and severe raters. The comparability of evaluation scores provided by several examiners is impacted by the harshness of the rater. Rater severity occurs when a rater shows significant resistance towards an agreed scoring pattern during training sessions [28]. In the end, failure to identify and manage rater severity could lead to unjust or inaccurate assessments of student proficiency. Studies by Eckes [28] have consistently shown that it is unrealistic to expect examiners to behave uniformly in terms of rating patterns. Several elements, such as professional or rating experience, attitudes, personality traits, the purpose of the assessment, and workload could influence the rating pattern of the rater making such a rater to be classified as either a 'hawk' or 'dove'. Despite the simplicity of describing the hawk and dove phenomenon, it is crucial to develop a reliable statistical method to tackle such rating patterns or behaviors.

3. METHOD

The study employed the three-facet design within the MFRM framework. These facets included the student, tutor (rater), and item (portfolio). The scoring rubrics used by tutors in rating student's portfolios use varying response categories across each section as a result partial credit modelling (PCM) was considered for this study. According to Masters [29], the PCM is most appropriate when response categories across items use different numbers or when the relative difficulty or complexity of tasks varies across sections.

This study relied on pre-existing secondary data, which were extant data of students' portfolio scores for a semester. Examining archival data for a purpose other than its initial intent has the potential to uncover or reassess emerging trends in assessment practices. It could offer researchers the opportunity to validate findings from other studies [30]. Employing the purposive sampling approach, a total of 207 student portfolios from the 2021/2022 academic year that were scored by tutors were obtained from two colleges of education. O'Mara *et al.* [31] recommend 20-30 observations per facet, therefore, the sample size used for this study was appropriate.

In the colleges of education, student portfolios are assessed using the supported teaching in school rubric for college-based grading. This evaluation form has four components with various items scored on distinct rating scales. The component "personal philosophy," underscores the importance of clearly expressing one's values, objectives, and convictions regarding teaching and learning. The subsequent component centers on "student reflection," while the third element pertains to "photographs/artifacts," which necessitates the inclusion of appropriate visual materials accurately depicting the activities conducted during each visit. Lastly, the fourth component, "post observation presentation or report," emphasizes written and oral presentations.

4. RESULTS AND DISCUSSION

4.1. Strictness and leniency behavior among raters

The study investigates how raters behave when evaluating students' portfolios. This section focuses on the raters' levels of strictness and leniency in scoring. The analysis was conducted by examining the rater facet in the measurement report, which aimed to assess the scoring patterns of each rater concerning the scores assigned to each criterion of the scoring rubrics. The results are shown in Table 1.

Table 1. Measurement report obtained for the rater facet

Rater	Logit measure	SE	Infit	Outfit	Fair average	Observed average	CI range	
R4	1.26	.07	.89	.93	4.00	4.32	1.12	1.4
R3	.04	.07	.77	.89	3.41	3.31	-.10	.18
R5	-.29	.09	1.42	1.23	3.27	3.11	-.47	-.11
R1	-.30	.07	1.01	.95	3.27	3.08	-.44	-.16
R2	-.70	.07	1.08	1.09	3.13	2.85	-.84	-.56
Mean	.00	.07	1.03	1.02				
Standard deviation	.75	.01	.25	.14				

Model, sample: root-mean-square deviation (RMSE)=.08, standard deviation =.75
Separation ratio=9.94; Separation index=13.59; Reliability of separation index =.99
Model, fixed (all same) Chi-square=418.9; SD=4; p=.00
Model, random (normal) Chi-square=4; SD=3; p=.26

In the first column of Table 1, raters appear in the order of their strictness (severity), that is, from most severe to most lenient. Each severity measure constitutes an estimate of a rater's "true" location on the latent variable. Thus, each measure is associated with some degree of estimation error. As evident from Table 1, the strictest rater was rater 4 (logit=1.26), and the rater showing the highest leniency was rater 2 (logit=-.70). In addition, the standard error (SE) was used to define an interval around the estimate within which the value of the parameter is expected to fall at a time (the confidence interval (CI)). The strictness measure of rater 4 was estimated to be 1.26 logits, with SE=.07 with a lower limit CI of 1.12, and an upper limit of 1.4. The width of a CI represents the degree of uncertainty inherent in a set of data. Hence, one can be more certain about rater 4's measure than the rest of the raters. In addition, the CIs for rater 4 and the rest of the raters (R2, R3, R5, and R1) do not overlap, which implies that their measures are significantly different. However, the CIs for rater 5 and rater 1 overlap strongly, hence their measures are not significantly different.

The infit and outfit values of the raters appeared to be acceptable (range .7 to 1.4), with the value of each rater being close to the expected value (1). Raters with fit values greater than 1.0 show more variation than expected in their ratings; this is called misfit (or underfit). Raters 1, 5, and 2 show high mean-square fit statistics. This indicates that they are not consistent (too unpredictable) in their judgment of similar performances. For performances of the same ability level, different ratings are awarded. 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 overfitting. Rater 4 and 3 low mean-square statistics suggest that these raters are over-fitting (too predictable). In other words, these raters are highly likely to display a restriction of range. They tend to give similar ratings to performances of different ability levels, thus not discriminating. Additionally, the high values of separation ratio, separation index, and separation index reliability indicate that the raters differed in their scoring of the student's portfolios. Similarly, the fixed-effects Chi-square value was significant, suggesting that the raters exhibited different behaviors when scoring.

4.2. Halo effect

Another rater behavior that is highly likely to occur in portfolio assessment is the halo effect. To determine this effect, the measurement report of the item facet was examined as a statistical indicator at the group level. In this study, during this analysis, it was observed that the separation ratio was 8.23, the separation index was 11.31, the reliability of the separation index was .99, and the fixed-effects Chi-square value was statistically significant (χ2=1051.7, SD=13, *p*=.00), as shown in Table 2. According to these results, the difficulty levels of the criteria differed, and there was no halo effect on the scoring. To determine whether the raters displayed halo behavior, the infit and outfit values of the raters, which are statistical indicators at the individual level, were examined. The measurement report of the items revealed that the differences in the logit values between the difficulty levels of the criteria were greater (2.03–(-1.33)=3.36). If the fit value of a rater significantly differs from 1, that rater is considered to display halo behavior. In this current study, raters 5, 1, and 2 had a fit value above 1; this suggests a halo effect on the scoring of these raters (Table 1).

Table 2. Measurement report obtained for the item facet

Item	Logit measure	SEM	Infit	Outfit
Item 13	2.03	.12	1.37	1.37
Item 1	1.87	.14	.72	.69
Item 11	1.51	.15	1.25	1.27
Item 14	1.18	.08	1.09	1.09
Item 8	.42	.07	.90	1.00
Item 6	-.40	.14	.73	.68
Item 12	-.44	.13	1.08	1.05
Item 10	-.51	.13	1.11	1.14
Item 2	-.61	.16	.84	.83
Item 5	-.64	.17	1.55	1.58
Item 4	-.66	.14	.75	.70
Item 3	-1.21	.15	.81	.79
Item 7	-1.22	.16	1.07	1.09
Item 9	-1.33	.13	.83	.76
Mean	.00	.13	1.01	1.00
Standard deviation	1.14	.03	.25	.27

Model, sample: RMSE=.14, SD=1.17
Separation ratio =8.55; Separation index =11.73
Reliability of separation index =.99
Model, fixed (all same) Chi-square =1051.7, SD=13, p=.00
Model, random (normal) Chi-square =12.8, SD=12, p=.38

4.3. Discussion

The current inquiry examines the rating patterns of tutors when scoring students' portfolios in the colleges of education in Ghana. The findings indicate that tutors demonstrate rating behaviors such as severity, leniency and halo effect when scoring student's portfolios. The findings of this study are novel as they offer insights that enhance the understanding of rater variability in portfolio scores within the colleges of education. These findings imply that rater characteristics impact portfolio scores, leading to variability in the results [12], [13], [16], [18]. This may explain why many students hold negative perceptions about their scores. As a result, students may question the decisions based on such scores, believing they do not accurately reflect their true potential, increasing concerns about the validity and reliability of portfolio scores.

For example, the findings revealed that raters differed in their levels of strictness and leniency when evaluating students' portfolios. This finding corroborates Huang and Whipple [27] observation that raters significantly differ in their severity degree when assessing student's constructed responses. Additionally, it supports Sata and Karakaya [32] assertion that raters exhibit varying levels of strictness when scoring. The variations observed among raters based on their strictness or leniency when scoring influence students' scores on the trait being measured [32]. This variation among raters may be influenced by their emotions, beliefs, preferences, perceptions, backgrounds, and experiences [33]. Since tutors are involved in various academic activities, including teaching, research, counselling, and grading students' scripts and portfolio work, these demands can create time constraints, leading to less thorough assessments and potentially more lenient or inconsistent grading. Tutors facing the pressure of managing multiple academic responsibilities may encounter cognitive overload, which can impair their ability to consistently apply scoring rubrics, thus affecting the reliability of the grades given [7]. Furthermore, the stress from high academic demands can intensify subjective biases, leading tutors to depend more on heuristics than on detailed analysis when evaluating student work. These observed strictness and leniency among raters impact the assessment process, leading to inconsistent judgments and unequal treatment of students, threatening the validity of portfolio assessment results [33]. Implementing quality control measures like rater training reduces variability among raters. Literature has shown that taking raters through training enhances understanding of the assessment criteria, standards, and expectations [7], [20], [21].

5. CONCLUSION

This study examined the rating behavior of tutors when scoring students' portfolios in the colleges of education in Ghana. It can be concluded from this study that tutors in the colleges demonstrated strictness and leniency in their scoring which can impact portfolio scores. Therefore, the management of the colleges should implement training programmed for raters. These programmed should focus on increasing rater awareness of potential biases and providing strategies to minimize their impact. Raters should be encouraged to adopt a systematic and objective approach when evaluating portfolios or constructed response tasks, emphasizing the importance of scoring rubrics.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Andrews Cobbinah	✓	✓		✓	✓	✓		✓	✓	✓			✓	
Jephtar Adu-Mensah		✓	✓		✓			✓	✓	✓	✓	✓		

C :	**C**onceptualization	I :	**I**nvestigation	Vi :	**Vi**sualization
M :	**M**ethodology	R :	**R**esources	Su :	**Su**pervision
So :	**So**ftware	D :	**D**ata Curation	P :	**P**roject administration
Va :	**Va**lidation	O :	Writing - **O**riginal Draft	Fu :	**Fu**nding acquisition
Fo :	**Fo**rmal analysis	E :	Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author [AC], upon reasonable request.

REFERENCES

[1] P. A. Atiah, "Newly qualified teachers' perceptions and experiences about portfolio assessment in WA municipality, Ghana," *African Journal of Teacher Education (AJOTE)*, vol. 11, no. 2, pp. 140–163, Jan. 2023, doi: 10.21083/ajote.v11i2.7043.

[2] L. F. Bachman, *Statistical analyses for language assessment*. Cambridge: Cambridge University Press, 2004, doi: 10.1017/CBO9780511667350.

[3] D. R. Sadler, "Indeterminacy in the use of preset criteria for assessment and grading," *Assessment & Evaluation in Higher Education*, vol. 34, no. 2, pp. 159–179, Apr. 2009, doi: 10.1080/02602930801956059.

[4] M. E. Nungari and K. T. Odhiambo, "Teachers and students perception on portfolio assessment in Kenyan secondary schools: towards preparedness for competency based education," *Journal of Research & Method in Education (IOSR-JRME)*, vol. 12, no. 2, pp. 12–22, 2022, doi: 10.9790/7388-1202041222.

[5] H. Komatsu, J. Rappleye, and I. Silova, "Student-centered learning and sustainability: solution or problem?" *Comparative Education Review*, vol. 65, no. 1, pp. 6–33, Feb. 2021, doi: 10.1086/711829.

[6] W. Meeus, P. van Petegem, and N. Engels, "Validity and reliability of portfolio assessment in pre-service teacher education," *Assessment & Evaluation in Higher Education*, vol. 34, no. 4, pp. 401–413, Aug. 2009, doi: 10.1080/02602930802062659.

[7] R. K. Shah, "Concepts of learner-centred teaching," *Shanlax International Journal of Education*, vol. 8, no. 3, pp. 45–60, Jun. 2020, doi: 10.34293/education.v8i3.2926.

[8] E. Schaefer, "Rater bias patterns in an EFL writing assessment," *Language Testing*, vol. 25, no. 4, pp. 465–493, 2018, doi: 10.1177/0265532208094273.

[9] K. Barkaoui, "Effects of marking method and rater experience on ESL essay scores and rater performance," *Assessment in Education: Principles, Policy & Practice*, vol. 18, no. 3, pp. 279–293, Aug. 2011, doi: 10.1080/0969594X.2010.526585.

[10] J. Huang, "Factors affecting the assessment of ESL students' writing," *International Journal of Applied Educational Studies*, vol. 5, no. 1, pp. 1–17, 2009.

[11] J. Huang, D. Zhu, D. Xie, and T. Shu, "Examining the reliability of an international Chinese proficiency standardized writing assessment: implications for assessment policy makers," *Assessing Writing*, vol. 55, p. 100693, Jan. 2023, doi: 10.1016/j.asw.2023.100693.

[12] T. Han and J. Huang, "Examining the impact of scoring methods on the institutional EFL writing assessment: a Turkish perspective," *PASAA*, vol. 53, no. 1, pp. 112–147, Jan. 2017, doi: 10.58837/CHULA.PASAA.53.1.5.

[13] H. Li and L. He, "A comparison of EFL raters' essay-rating processes across two types of rating scales," *Language Assessment Quarterly*, vol. 12, no. 2, pp. 178–212, Apr. 2015, doi: 10.1080/15434303.2015.1011738.

[14] J. Huang and C. J. Foote, "Grading between the lines: what really impacts professors' holistic evaluation of ESL graduate student writing?" *Language Assessment Quarterly*, vol. 7, no. 3, pp. 219–233, Jul. 2010, doi: 10.1080/15434300903540894.

[15] S. C. Weigle, H. Boldt, and M. I. Valsecchi, "Effects of task and rater background on the evaluation of ESL student writing: a pilot study," *TESOL Quarterly*, vol. 37, no. 2, pp. 345–354, Jul. 2003, doi: 10.2307/3588510.

[16] Y. Liu and J. Huang, "The quality assurance of a national English writing assessment: policy implications for quality improvement," *Studies in Educational Evaluation*, vol. 67, p. 100941, Dec. 2020, doi: 10.1016/j.stueduc.2020.100941.

[17] Y.-W. Lee, R. Kantor, and P. Mollaun, "Score dependability of the writing and speaking sections of new TOEFL," in *Annual Meeting of National Council on Measurement in Education*, 2002.

[18] C. Zhao and J. Huang, "The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: implications for assessment policy makers," *Studies in Educational Evaluation*, vol. 67, p. 100911, Dec. 2020, doi: 10.1016/j.stueduc.2020.100911.

[19] S. C. Weigle, *Assessing writing*. Cambridge: Cambridge University Press, 2002.

[20] P. Black and D. Wiliam, "Classroom assessment and pedagogy," *Assessment in Education: Principles, Policy & Practice*, vol. 25, no. 6, pp. 551–575, Nov. 2018, doi: 10.1080/0969594X.2018.1441807.

[21] D. M. Yoo, A. R. Cho, and S. Kim, "Development and validation of a portfolio assessment system for medical schools in Korea," *Journal of Educational Evaluation for Health Professions*, vol. 17, p. 39, Dec. 2020, doi: 10.3352/jeehp.2020.17.39.

[22] K. Asare, "Exploring the kindergarten teachers' assessment practices in Ghana," *Developing Country Studies*, vol. 5, no. 8, pp. 110–128, 2015.

[23] B. Kankam, A. Bordoh, I. Eshun, T. K. Bassaw, and F. Y. Korang, "An investigation into authentic assessment practices of social studies teachers in the senior high schools (SHSS) in Ghana," *American Journal of Social Sciences*, vol. 2, no. 6, pp. 166–172, 2014.

[24] A. Bordoh, I. Eshun, A. M. Quarshie, T. K. Bassaw, and P. Kwarteng, "Social studies teachers' knowledge base in authentic assessment in selected senior high schools in the Central Region of Ghana," *Journal of Social Sciences and Humanities*, vol. 1, no. 3, pp. 249–257, 2015.

[25] E. Anane and J. Adu-Mensah, "Teachers scoring and grading of students' responses to tasks: the Ghanaian basic school experience," *Academic Journal of Interdisciplinary Studies*, vol. 8, no. 1, pp. 87–94, Mar. 2019, doi: 10.2478/ajis-2019-0009.

[26] J. M. Linacre, "Many-faceted Rasch measurement," Ph.D. dissertation, The University of Chicago, United States, 1989.

[27] J. Huang and P. B. Whipple, "Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: implications for educational assessment policy," *Humanities and Social Sciences Communications*, vol. 10, no. 1, p. 860, Nov. 2023, doi: 10.1057/s41599-023-02385-4.

[28] T. Eckes, *Introduction to many-facet Rasch measurement: analysing and evaluating rater-mediated assessments*, 2nd ed. Frankfurt: Peter Lang, 2015.

[29] G. N. Masters, "The partial credit models," in *Handbook of polytomous item response theory models*, M. L. Nering and R. Ostini, Eds., New York: Routledge, 2010, pp. 109–122.

[30] E. Anane, "Pre-service teachers' motivational orientations and the impact of self-regulated learning on their academic achievement: a mixed method study," Ph.D. dissertation, Durham University, England, United Kingdom, 2014.

[31] L. O'Mara, B. Carpio, C. M Mallette, W. Down, and B. Brown, "Developing a teaching portfolio in nursing education: a reflection," *Nurse Educator*, vol 25, no. 3, pp. 125-130, 2000, doi: 10.1097/00006223-200005000-00012.

[32] M. Şata and İ. Karakaya, "Investigation of the use of electronic portfolios in the determination of student achievement in higher education using the many-facet Rasch measurement model," *Educational Policy Analysis and Strategic Research (EPASR)*, vol. 15, no. 1, pp. 7–21, Mar. 2020, doi: 10.29329/epasr.2020.236.1.

[33] A. van Moere, "Raters and ratings," in *The Companion to Language Assessment*, A. J. Kunnan, Ed., New York: Wiley, 2013, pp. 1358–1374, doi: 10.1002/9781118411360.wbcla106.

BIOGRAPHIES OF AUTHORS

Andrews Cobbinah holds a Ph.D. in research, measurement and evaluation. He is a senior lecturer at the University of Cape Coast-Ghana. His research areas are skill training in peer assessment and critical thinking. He has 12 years of teaching experience at the university. He lectures in statistics, educational assessment and educational measurement. He can be contacted at email: andrews.cobbinah@ucc.edu.gh.

Jephtar Adu-Mensah is a Ph.D. Candidate in Department of Education and Foundation, Graduate School, University of Cape Coast, Ghana. He is a senior assessment officer at the National Council for Curriculum and Assessment (NaCCA). His research focuses on assessment practices, and rater variance in educational assessment. He can be contacted at email: mensahjef@gmail.com.