Creative Education, 2022, 13, 3483-3489
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Determining the Difficulty and Discrimination Parameters of a Mathematics Performance-Based Assessment

Abraham Gyamfi[1], Douglas G. Wren[2]

[1]Department of Educational Studies, Wesley College of Education, Kumasi, Ghana
[2]Darden College of Education & Professional Studies, Old Dominion University, Norfolk, Virginia, USA
Email: abrahamgyamfi84@gmail.com, dwren@odu.edu

How to cite this paper: Gyamfi, A., & Wren, D. G. (2022). Determining the Difficulty and Discrimination Parameters of a Mathematics Performance-Based Assessment. *Creative Education, 13,* 3483-3489.
https://doi.org/10.4236/ce.2022.1311223

Received: October 17, 2022
Accepted: November 12, 2022
Published: November 15, 2022

Abstract

Performance-based assessment (PBA) is different from traditional testing methods in that PBA presents real-life problems for students to solve by integrating critical thinking with their content knowledge and skills. Implementing PBA regularly in mathematics classes is associated with improved student achievement and motivation to learn; however, there are concerns about the general lack of psychometric data to support the use of performance assessments. To address such concerns, this study applied item response theory to estimate the difficulty and discrimination indices of items that comprised a newly developed mathematics PBA. Data were collected by administering the PBA to 750 senior high school students in the Western Region of Ghana. The results indicated that the difficulty and discrimination levels of each item were satisfactory, which suggests that well-designed and properly vetted math PBA items would improve classroom assessments as well as high-stakes tests administered on a large scale. Additional recommendations are included at the end of this paper.

Keywords

Performance Assessment, Mathematics, Item Difficulty, Item Discrimination, Item Response Theory

1. Introduction

Student achievement in mathematics has been hindered by various challenges related to assessment (Gao, 2012; Nortvedt & Buchholtz, 2018; Suurtamm et al., 2016). Two challenges are "a focus on recall of isolated items of knowledge…

[and] inadequacy in aligning assessment tasks with students' real-life situations" (Gao, 2012: p. 63). Performance assessment, also called performance-based assessment (PBA), addresses these concerns in ways that traditional assessments do not. On a PBA, examinees perform the actual skills that the test was designed to measure (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In contrast to most traditional tests, PBA requires students to think critically and apply their knowledge to create a response or a product. Furthermore, PBA incorporates real-world tasks and contexts, thus making the assessment more relevant to students than multiple-choice tests (Wiggins & McTighe, 2005). Implementing performance assessment in mathematics classes not only enhances the quality of instruction and improves student achievement (Mulana et al., 2021; Stone & Lane, 2003), the practice motivates students in their learning as well (Arhin, 2015; Balik, 2012).

The use of performance assessment comes with its own set of challenges. For decades, experts have reported that "the usual suspects" of administering PBA on a large scale are costs and scoring subjectivity (e.g., Darling-Hammond & Adamson, 2010; Madaus & O'Dwyer, 1999; Stiggins & Bridgeford, 1984; Wren & Gareis, 2019). Another problem is the typical lack of psychometric evidence to support the use of PBA in educational settings (Davey et al., 2015; DiCerbo et al., 2016; Stecher et al., 2000). A related issue is developing performance assessments that are appropriate for students with diverse ability levels, as "it is quite conceivable that a well-designed performance assessment could be administered and yet fail to provide useful data if the assessment is delivered at a level that is either too difficult or too easy for the student being assessed" (Bahr, 2007: p. 34). Performance assessments are seldom analyzed to determine the degree of difficulty or how well they discriminate between students at different proficiency levels.

Classical test theory (CTT) has served the educational and psychological measurement community admirably for over 100 years, but CTT models have shortcomings due to the theory's assumptions. One problem associated with CTT is that item statistics—specifically difficulty and discrimination indices—are dependent on the general ability level of the sample of examinees employed during the test development process (Hambleton, 2000). For example, a homogeneous sample of high-ability examinees might yield data that indicates the items are easier and less variant than if data from a heterogeneous sample was used to develop the test. Item response theory (IRT), which appeared on the educational and psychological measurement scene much later, involves the relationship between examinees' ability levels and their responses to test items. For this reason, contemporary researchers and practitioners in the measurement field tend to prefer IRT models to calculate item difficulty and item discrimination parameters.

2. Methodology

The aim of the study was to apply item response theory to determine the diffi-

culty and discrimination parameters of items on a performance-based assessment developed by the primary author. The PBA comprised five mathematics computation items presented in real-life scenarios. Each item was designed to assess the proficiency of senior high school (SHS) students in these math domains: transformation, descriptive statistics, mensuration, geometric construction, and linear equations. Item 4, the geometric construction item, is shown in Figure 1.

A total of 750 SHS students in the Western Region of Ghana were administered the PBA. The sample consisted exclusively of SHS 3 students (i.e., students in their third and final year of high school) from two randomly selected classes at 15 high schools, stratified according to the Ghana Education Service's SHS categorization system. There were five schools from each of the three highest SHS categories: A, B, and C. Senior high schools in the highest categories are government or public schools, which are considered the best schools in the traditionally competitive Ghanaian secondary school system.

Using a standardized scoring rubric developed by the primary author, three different examiners—one each for the A, B, and C categories—scored the students' responses to the PBA items. Data were analyzed with Samejima's (1969) two-parameter logistic (2PL) graded response model, a popular IRT model used to estimate the difficulty and discrimination levels of polytomously scored items such as those on performance assessments.

3. Results

Data analyses yielded values that indicated the discrimination and difficulty parameters of the PBA items (see Table 1). Discrimination, known in IRT as location, is represented by a, and difficulty, or slope, is denoted by b in the table. Although the theoretical range of both parameters in IRT is −4 to +4, ranges that are observed in practice are usually $-2.8 \leq a \leq +2.8$ and $-3 \leq b \leq +3$ (Baker, 2001).

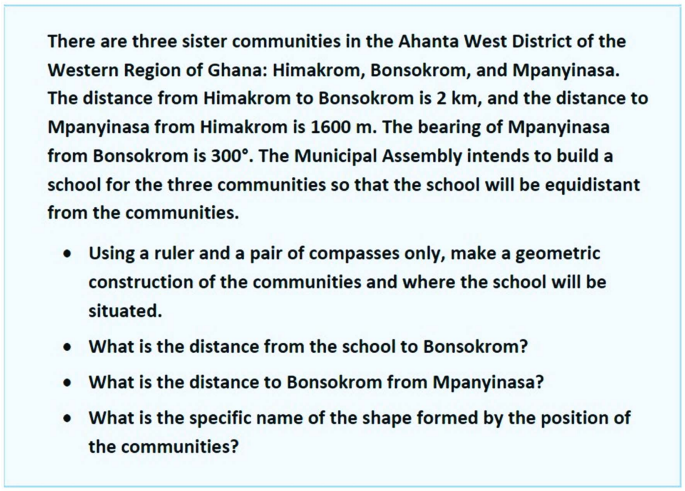

Figure 1. Geometric construction item for mathematics performance-based assessment.

Table 1. Parameter estimates for mathematics performance-based assessment items.

Parameter	Item 1 *transformation*	Item 2 *descriptive statistics*	Item 3 *mensuration*	Item 4 *geometric construction*	Item 5 *linear equations*
a	0.969	1.799	2.836	1.163	2.710
b_1	−2.888	−1.502	−2.278	−2.313	−2.278
b_2	−1.379	−1.123	−1.892	−1.754	−1.553
b_3	−0.804	−0.711	−1.325	−1.191	−0.893
b_4	−0.022	−0.707	−1.131	−0.680	0.135
b_5	0.109	0.927	−0.956	−0.660	0.224
b_6	0.672	1.629	−0.527	0.525	0.449
b_7	0.914	1.769	0.384	1.214	2.053
b_8	2.172	2.683	0.935	1.487	2.306
b_9	2.459	2.945	0.950	2.811	2.741

a = item discrimination (slope), b_i = item difficulty (location).

Baker's (2001) cut-off ranges were used to determine how well the items discriminated between examinees with different levels of proficiency in each domain. The cut-off ranges were as follows: *very low discrimination* = 0.01 to 0.34, *low discrimination* = 0.35 to 0.64, *moderate discrimination* = 0.65 to 1.34, *high discrimination* = 1.35 to 1.69, and *very high discrimination* > 1.70. While the transformation and geometric construction items discriminated moderately between students at different points on the proficiency continuum, the descriptive statistics, mensuration, and linear equation items all demonstrated *very high* discrimination power.

Unlike CTT with its single-value approximations of item difficulty, IRT models provide varied estimates of an item's difficulty depending on the estimated ability of examinees. In other words, the values at each level (b_i) explain how an item performs along the ability scale. The statistical definition of item difficulty in IRT is the point on the ability scale at which the probability of answering the item correctly is 0.5 (Baker, 2001).

The b values in **Table 1** show that each item performed as expected, as b increased progressively in value from the lowest ability level to the highest ability level, (i.e., b_1 to b_9). The items with the greatest range of difficulty were the transformation and linear equations items, with b_1 and b_9 values approaching the typical minimum and maximum b values of −3 to +3. The least difficult and most difficult items were the descriptive statistics and mensuration items, respectively. Overall, the functioning of items along the ability scale indicated that the difficulty level of every item was acceptable.

4. Recommendations

The results of the study have implications for improving classroom and large-scale mathematics assessments in Ghana. At present, the degree of difficulty and dis-

crimination of items on SHS classroom tests is largely unknown. Well-designed performance-based assessments would provide better information about students' proficiency in mathematics than the traditional assessments currently do. In addition, high-quality PBA items would enhance the core mathematics section of the high-stakes *West African Senior Secondary Certificate Examination* (WASSCE) by effectively discriminating between examinees who are proficient in a specific domain and examinees who are not.

Ghanaian mathematics teachers and assessors associated with the West African Examinations Council (WAEC) who have access to software for analyzing polytomous test items could use this study to guide them as they create and calibrate new PBA items for their tests. The WAEC's vision is "To be a world-class examining body adding value to the educational goals of its stakeholders" (WAEC, 2022), so it seems logical that the organization would embrace the best assessment methods and statistical models for obtaining comprehensive psychometric evidence for items on the WASSCE and other assessments.

Other recommendations mirror those made by Arhin (2015) in earlier research conducted with SHS mathematics students in Ghana. After finding that performance assessment-driven instruction "had an encouraging effect on students' attitude towards mathematics especially on students' motivation, independent thinking and understanding in solving mathematical problems" (p. 114), Arhin suggested that SHS mathematics teachers include PBA tasks in their lessons. He also proposed that math teachers receive training on the use of performance assessment-driven instruction. Along the same lines, the authors of the present study recommend that PBA become an integral part of methods and assessment courses at colleges and universities where students are trained to teach mathematics. These recommendations pertain not only to Ghana, but to other countries where performance-based learning and assessment have not been fully actualized in secondary classrooms. Providing current and future educators with the tools and knowledge to develop lessons and tests that incorporate PBA would have a catalytic effect on mathematics education by "improving the quality, realism, and utility of instruction for students" (Baker, 2019: p. vii).

Acknowledgements

The authors would like to thank the heads of the senior high schools in the Western Region of Ghana that participated in the study, as well as Stephen Court for his help during the preparation of this paper.

Conflicts of Interest

The authors declare no conflicts of interest with respect to the research, authorship, and publication of this article.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and*

Psychological Testing. American Educational Research Association.

Arhin, A. K. (2015). The Effect of Performance Assessment-Driven Instruction on the Attitude and Achievement of Senior High School Students in Mathematics in Cape Coast Metropolis, Ghana. *Journal of Education and Practice, 6,* 109-116. https://files.eric.ed.gov/fulltext/EJ1083838.pdf

Bahr, D. L. (2007). Creating Mathematics Performance Assessments that Address Multiple Student Levels. *Australian Mathematics Teacher, 63,* 33-40. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2352&context=facpub

Baker, E. L. (2019). Foreword. In D. G. Wren, & C. R. Gareis (Ed.), *Assessing Deeper Learning: Developing, Implementing, and Scoring Performance Tasks* (pp. vii-ix). Rowman Littlefield.

Baker, F. B. (2001). *The Basics of Item Response Theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Balik, I. W. (2012). Pengaruh Implementasi Asesmen Autentik Terhadap Prestasi Belajar Matematika dan Motivasi Berprestasi (Eksperimen pada Peserta Didik Kelas VIII SMP Negeri 3 Gianyar). [The Effect of Authentic Assessment Implementation on Mathematics Learning Achievement and Achievement Motivation (Experiments on Class VIII Students of SMP Negeri 3 Gianyar)]. *Jurnal Penelitian dan Evaluasi Pendidikan, 2,* 1-26. https://ejournal-pasca.undiksha.ac.id/index.php/jurnal_ep/article/view/380

Darling-Hammond, L., & Adamson, F. (2010). *Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning.* Stanford University, Stanford Center for Opportunity Policy in Education. https://edpolicy.stanford.edu/library/publications/1462

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment.* Center for K-12 Assessment & Performance Management, Educational Testing Service. https://www.ets.org/Media/Research/pdf/psychometric_considerations_white_paper.pdf

DiCerbo, K. E., Shute, V., & Kim, Y. (2016). The Future of Assessment in Technology-Rich Environments: Psychometric Considerations. In J. M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy* (pp. 1-21). Springer. https://doi.org/10.1007/978-3-319-17727-4_66-1

Gao, M. (2012). Classroom Assessments in Mathematics: High School Students' Perceptions. *International Journal of Business and Social Science, 3,* 63-68. https://ijbssnet.com/journal/index/937

Hambleton R. K. (2000). Emergence of Item Response Modeling in Instrument Development and Data Analysis. *Medical Care, 38,* II-60-II-65. https://doi.org/10.1097/00005650-200009002-00009

Madaus, G. F., & O'Dwyer, L. M. (1999). A Short History of Performance Assessment: Lessons Learned. *The Phi Delta Kappan, 80,* 688-695. http://www.jstor.org/stable/20439537

Mulana, I. M. B., Candiasa, I. M., Jampel, I. N., & Suma, K. (2021). The Effects of Performance Assessment on Mathematics Learning Outcomes. *Academy of Entrepreneurship Journal, 27,* 1-13. https://www.abacademies.org/articles/the-effects-of-performance-assessment-on-mathematics-learning-outcomes.pdf

Nortvedt, G.A., & Buchholtz, N. (2018). Assessment in Mathematics Education: Responding to Issues Regarding Methodology, Policy, and Equity. *ZDM Mathematics Education, 50,* 555-570. https://doi.org/10.1007/s11858-018-0963-z

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika, 34,* 1-97. https://doi.org/10.1007/BF03372160 http://www.psychometrika.org/journal/online/MN17.pdf

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The Effects of Content, Format, and Inquiry Level on Science Performance Assessment Scores. *Applied Measurement in Education, 13,* 139-160. https://doi.org/10.1207/S15324818AME1302_2

Stiggins, R. J., & Bridgeford, N. J. (1984). *The Use of Performance Assessment in the Classroom.* Northwest Regional Educational Laboratory. https://files.eric.ed.gov/fulltext/ED242718.pdf

Stone, C. A., & Lane, S. (2003). Consequences of a State Accountability Program: Examining Relationships between School Performance Gains and Teacher, Student, and School Variables. *Applied Measurement in Education, 16,* 1-26. https://doi.org/10.1207/S15324818AME1601_1

Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., Silver, E. A., Ufer, S., & Vos, P. (2016). *Assessment in Mathematics Education: Large-Scale Assessment and Classroom Assessment.* Springer. https://doi.org/10.1007/978-3-319-32394-7

West African Examinations Council (2022). *Our Vision*. https://www.waecgh.org

Wiggins, G. P., & McTighe, J. (2005). *Understanding by Design* (2nd ed.). Association for Supervision and Curriculum Development.

Wren, D. G., & Gareis, C. R. (2019). *Assessing Deeper Learning: Developing, Implementing, and Scoring Performance Tasks*. Rowman Littlefield.