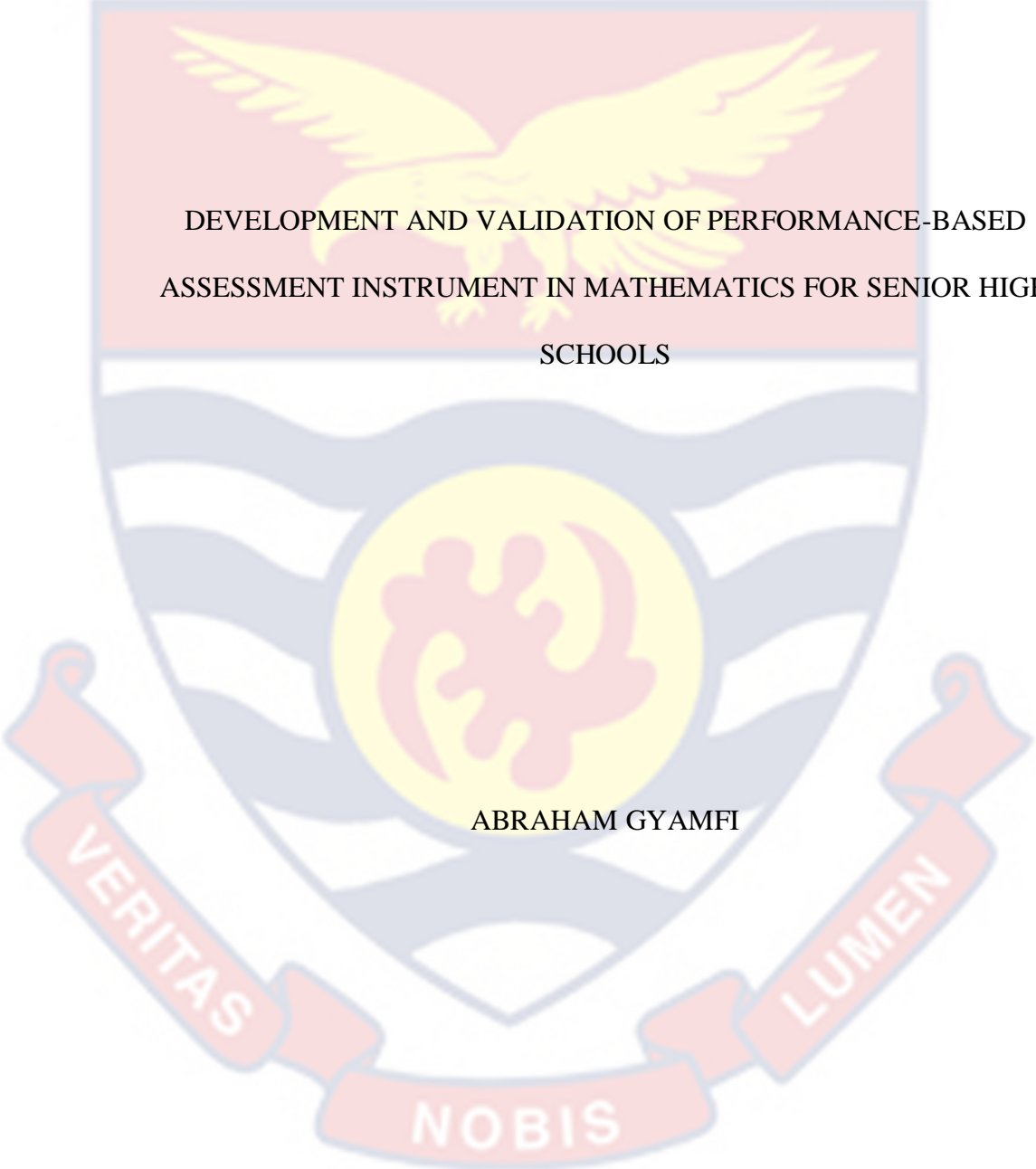


UNIVERSITY OF CAPE COAST



DEVELOPMENT AND VALIDATION OF PERFORMANCE-BASED
ASSESSMENT INSTRUMENT IN MATHEMATICS FOR SENIOR HIGH
SCHOOLS

ABRAHAM GYAMFI

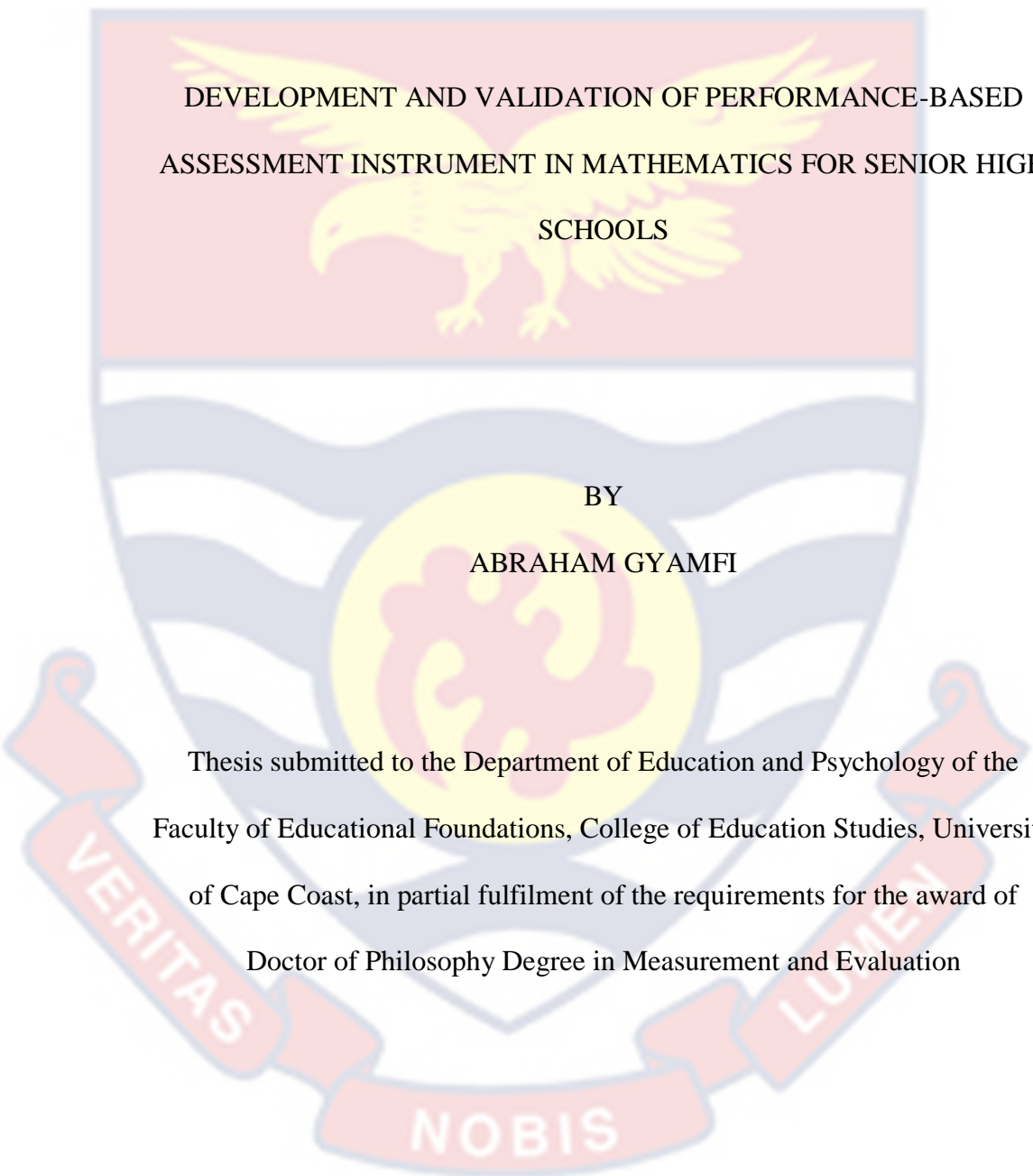
2023



© Abraham Gyamfi

University of Cape Coast

UNIVERSITY OF CAPE COAST



DEVELOPMENT AND VALIDATION OF PERFORMANCE-BASED
ASSESSMENT INSTRUMENT IN MATHEMATICS FOR SENIOR HIGH
SCHOOLS

BY
ABRAHAM GYAMFI

This thesis submitted to the Department of Education and Psychology of the
Faculty of Educational Foundations, College of Education Studies, University
of Cape Coast, in partial fulfilment of the requirements for the award of
Doctor of Philosophy Degree in Measurement and Evaluation

NOVEMBER 2023

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature: Date:

Name:.....

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature: Date:

Name:.....

Co-Supervisor's Signature: Date:

Name:.....

ABSTRACT

The purpose of the study was to develop and validate a performance-based assessment instrument in mathematics for Senior High Schools. The study sought to find out if a newly developed performance-based assessment items in mathematics would be a good assessment instrument for Senior High School examination. The study employed quantitative instrumentation research design with a four-phase instrument development and validation process: planning, construction, qualitative evaluations, and quantitative validation. Stratified, census, simple random sampling and purposive sampling procedures were employed to select 240 mathematics examiners, 150 mathematics teachers and 750 SHS Three students in the Western Region for the validation phase of the self-developed instrument. Questionnaire and performance-based assessment test were used as the main data collection instruments. The Cronbach alpha reliability coefficient of the questionnaire was 0.843. Data were analysed using means and standard deviation, Pearson Product Moment correlation coefficient, modified Kappa statistics, Principal Component Analysis and four-way ANOVA. It was found that the performance-based assessment instrument (designed by the author) is feasible and credible, has educational and catalytic effects. It was also found that the developed performance-based assessment has a high inter-rater reliability (0.879-0.988), good content validity ratio (0.834-1.00) and good construct validity. Based on the findings, it was recommended that the performance-based assessment should be an integral part of the methods of assessment lessons in mathematics at the Senior High Schools.

KEYWORDS

Catalytic effect

Credibility

Differential Item Functioning (DIF)

Educational effect

Examination malpractice

Feasibility

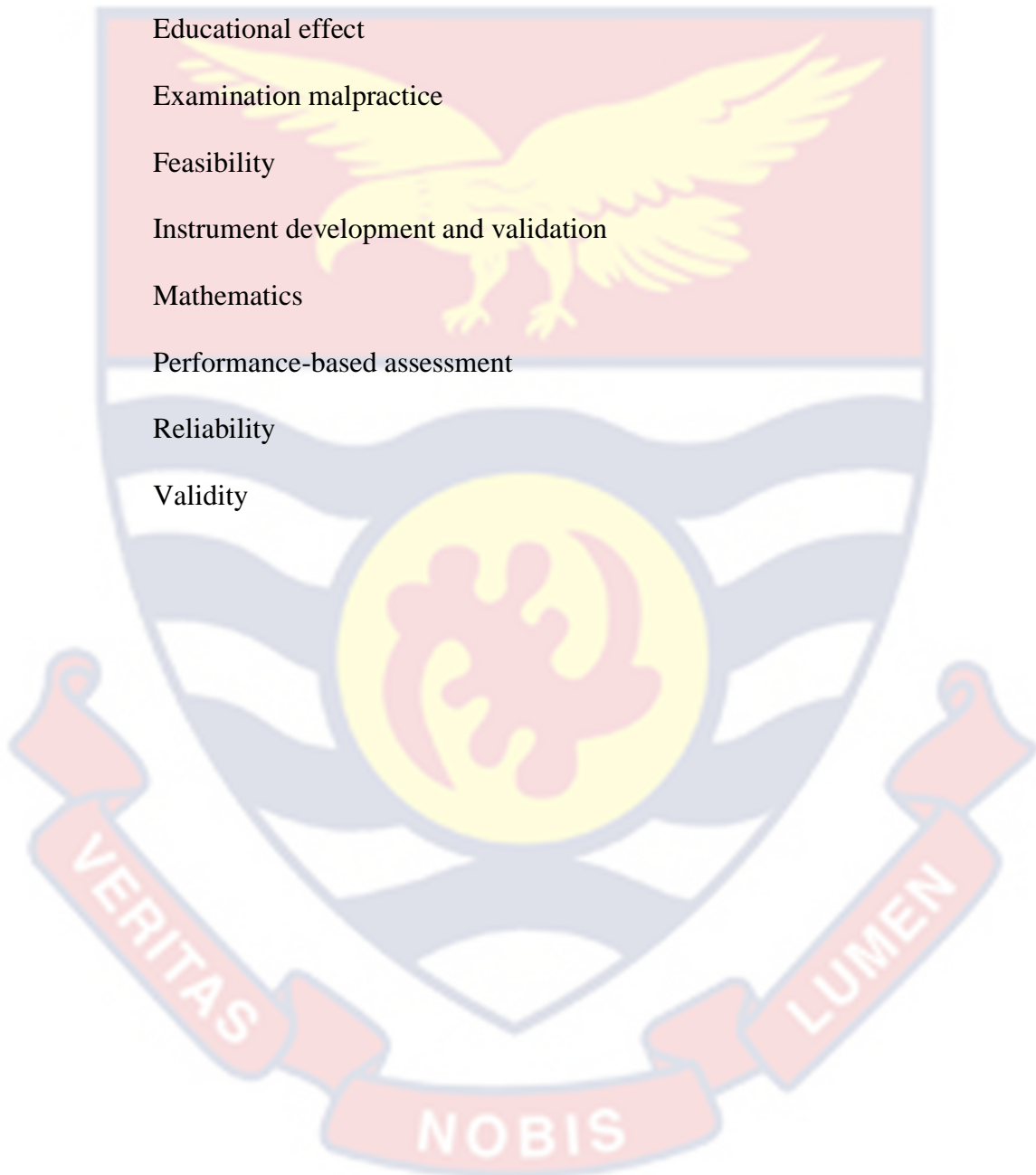
Instrument development and validation

Mathematics

Performance-based assessment

Reliability

Validity



ACKNOWLEDGEMENTS

Firstly, I would like to say thanks to my level-headed, level-minded and avuncular supervisors, Prof. Eric Anane and Dr. Kenneth Asamoah-Gyimah, for their commitment, dedication, assistance and directions for my research. You have immensely contributed to the success of this research.

I would also like to thank the Headmasters and the teachers of the selected schools for their support during the research in their schools. I really appreciate it all.

To my wife, Rosemary Acquaye and Children, Rita Efua-Duaba Cudjoe, Ayebineba Offeibea Ayebine-Gyamfi, Ayebineba Gyamfiwaa Ayebine-Gyamfi and Ayebineba Kwame-Dapaah Ayebine-Gyamfi, I say God bless you.

Again, I cannot forget the support of parents and siblings throughout the research. I say I appreciate it all. Thanks.

Finally, to all friends who in diverse ways rendered suggestions and encouragement, especially Mr. Abraham Yeboah. Rev. Isaac Nana Sam, Mr. Kingsley K. Erzoah and Miss Patience Lange I say thank you very much.

DEDICATION

To my family



TABLE OF CONTENTS

Content	Page
DECLARATION	ii
ABSTRACT	iii
KEYWORDS	iv
ACKNOWLEDGEMENTS	v
DEDICATION	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	xii
LIST OF FIGURES	xiv
CHAPTER ONE: INTRODUCTION	1
Background to the Study	2
Statement of the Problem	14
Purpose of the Study	21
Research Objectives	21
Research Questions	22
Research Hypotheses	22
Definition of Terms	23
Basic Assumptions	24
Significance of the Study	24
Delimitation	25
Limitations	26
Organisation of the Study	26
CHAPTER TWO: LITERATURE REVIEW	28
Conceptual Flow Chart	28

Concept of Assessment	30
Relationship between Assessment and Teaching and Learning	31
Forms of Assessment	34
Formative assessment	36
Strengths of formative assessment	37
Weakness of formative assessment	38
Summative assessment	39
Strengths of summative assessment	41
Weakness of summative assessment	42
Performance-based Assessment	43
Performance-based assessment tasks	46
Scoring Performance Assessment	47
How mathematics should be taught and assessed	48
Theoretical Review	52
Reliability	52
Reliability Theory	52
Methods of Estimating Reliability	55
Standard Error of Measurement (SEM)	57
Validity	58
Validity Theory	61
Contemporary view of validity	67
Principles of validation	68
Categories of validity evidence	70
Factors that affect validity	79
Test Theories and Test Development	80

Development and Validation of Instrument	83
Criteria for Evaluating Assessment Instrument	94
Empirical Studies	106
Educational effect	106
Catalytic effect	107
Feasibility	108
Reliability and Validity	109
Experience and educational effect	111
Experience and feasibility	112
Chapter Summary	113
CHAPTER THREE: RESEARCH METHODS	114
Introduction	114
Research Philosophy	114
Research Design	114
Study Area	116
Population	117
Sampling Procedures	117
Data Collection Instrument	119
Validity of Instrument	121
Pilot-testing of the instrument	122
Reliability of Instrument	123
Ethical Consideration	123
Data Collection Procedures	124
Stages of the development and validation of the PBA	125
Data Processing and Analysis Procedure	130

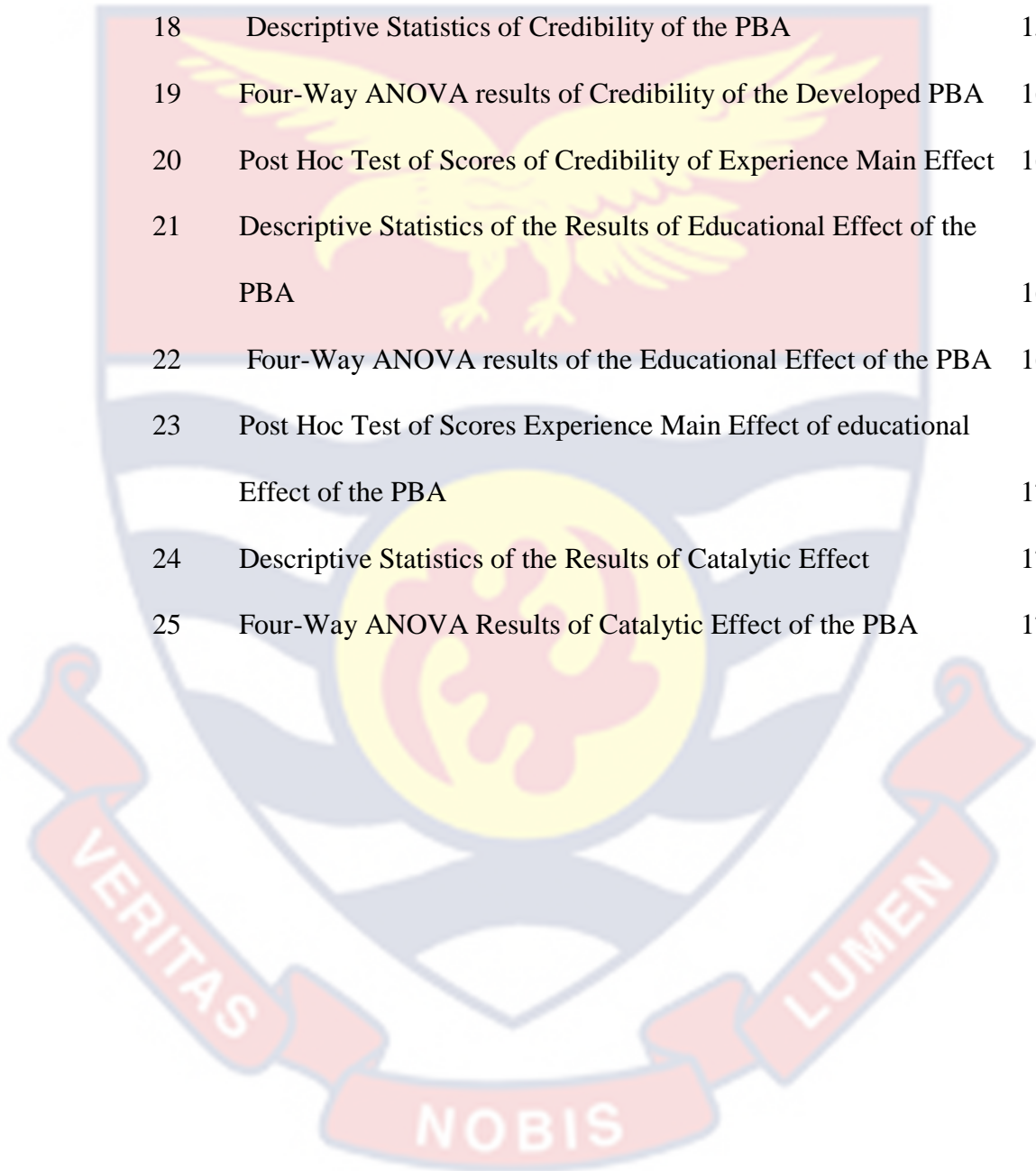
Chapter Summary	131
CHAPTER FOUR: RESULTS AND DISCUSSION	132
Introduction	132
Analysis of Bio-Data of Respondents	132
Analysis of Data on Research Questions	135
Research Question One	136
Research Question Two	138
Research Question Three	139
Research Question Four	141
Research Question Five	142
Research Question six	144
Analysis of Hypotheses	148
Checking assumptions for ANOVA	149
Discussions of key Findings	179
Chapter Summary	189
CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS	191
Summary of the Study	191
Conclusion	194
Contribution to Knowledge, Practice and Policy	195
Recommendations	196
Suggestions for Further Research	197
REFERENCES	198
APPENDICES	227

APPENDIX A- DISTRIBUTION OF THE TARGET AND ACCESSIBLE POPULATION	227
APPENDIX B1-TEST SPECIFICATION FOR THE PERFORMANCE BASED ASSESSMENT ITEMS	229
APPENDIX B2-PERFORMANCE-BASED ASSESSMENT TEST	235
APPENDIX B3-SCORING RUBRIC FOR THE PERFORMANCE BASED ASSESSMENT ITEMS	238
APPENDIX C-QUESTIONNAIRE FOR TEACHERS AND EXAMINERS	244
APPENDIX D- EXPLORATORY AND CONFIRMATORY FACTOR ANALYSIS	248
APPENDIX E-INTRODUCTORY LETTER	250
APPENDIX F-UNIDIMENTIONALITY AND LOCAL INDEPENDENCE ASSUMPTIONS OF THE PERFORMANCE BASED ASSESSMENT	251
APPENDIX G1-HOMOGENITY AND LEVENE TEST OF FEASIBILITY	252
APPENDIX G2-HOMOGENITY AND LEVENE TEST OF CREDIBILITY	253
APPENDIX G3-HOMOGENITY AND LEVENE TEST OF EDUCATIONAL EFFECT	254
APPENDIX G4-HOMOGENITY AND LEVENE TEST OF CATALYTIC EFFECT	255

LIST OF TABLES

Table	Page
1 Comparison of Traditional and PBA Task	7
2 Statistics on Examination Malpractice in WASSCE from 2016 to 2019	10
3 Minimum value of CVR	74
4 Cronbach alpha of the Questionnaire	123
5 Distribution of Respondents by Gender	133
6 Distribution of Examiners and Teachers by Years of Experience	134
7 Descriptive statistics of the Results by Mathematics Teachers and Examiners on the Feasibility of PBA items for SHSs (N = 390)	136
8 Results of More Feasible Statements of the PBA by Mathematics Teachers and Examiners (N = 390)	137
9 Descriptive Statistics of the Credibility of the Developed PBA items in Mathematics by the Mathematics Teachers and Examiners (N = 390)	138
10 Descriptive Statistics of the Educational Effect of the PBA items in Mathematics by the Mathematics Teachers and Examiners (N = 390)	140
11 Descriptive Statistics of the Catalytic Effect of the PBA items in Mathematics by the Mathematics Teachers and Examiners (N = 390)	141
12 Pearson Product Moment Correlation for Inter-rater Reliability	143
13 Modified Kappa Statistics for Content Validity Ratio for Item and the Scale	145

14	Eigenvalues of Total Variance Explained	147
15	Factor Loadings of the PBA Items	148
16	Descriptive Statistics of the Results of the Feasibility of the PBA	151
17	Four-Way ANOVA Results on Feasibility of the PBA	154
18	Descriptive Statistics of Credibility of the PBA	157
19	Four-Way ANOVA results of Credibility of the Developed PBA	160
20	Post Hoc Test of Scores of Credibility of Experience Main Effect	163
21	Descriptive Statistics of the Results of Educational Effect of the PBA	166
22	Four-Way ANOVA results of the Educational Effect of the PBA	169
23	Post Hoc Test of Scores Experience Main Effect of educational Effect of the PBA	171
24	Descriptive Statistics of the Results of Catalytic Effect	174
25	Four-Way ANOVA Results of Catalytic Effect of the PBA	177



LIST OF FIGURES

Figure		Page
1	Flow chart for the validation of the PBA	29
2	Conceptual Framework for the development and validation of the PBA	126
3	Stages of validation of the instrument	130
4	Scree Plot for the Items	147



CHAPTER ONE

INTRODUCTION

Mathematics is considered as one of the critical subjects at the Senior High School (SHS) level. According to the Ministry of Education (MOE) (2018), tackling societal issues can be aided by having a solid understanding of mathematics. Many people think that society is not affected by mathematics since students' understanding of the subject has not been applied very often to address societal issues. The comprehensive discussions of the issue were carried with people that matter in this study. Mathematics teachers and West Africa Examination Council (WAEC) examiners in the Western Region were involved in the study. The purpose was to have a good assessment instrument that will produce opportunity to students to perform the knowledge they have acquired in mathematics. Most of the questions students' respond to, both in the classroom and in external examinations are mostly not authentic assessment. The questions encourage rote learning without application. This reduces the expected impact of mathematics on the society. Performance-Based Assessment (PBA) which is application-centred is thus desired to bridge the gap between knowing and doing (Arhin, 2015). Classical test theory (CTT) was used to validate the item and test characteristics of the newly developed PBA. Generally, the aim of educational assessment is to bring out students' true performance for decision making. This is revealed in tasks where students demonstrate by performing how well they have mastered a particular content.

Background to the Study

In the assessment of students' learning, particularly in mathematics, there have been various changes made globally. This time period calls for a focus on assessment since, despite improvements in mathematics education and curricula, significant advancements in assessment in the subject have not been noticed (Suurtamm, et al, 2016; Bahr, Monroe & Mantilla, 2018). As Mpuangnan and Adusei (2021, p.2) stated, “as an issue of policy, the implementation of standards-based curricula should always be accompanied by the implementation of standards-based assessment”.

In fact, “incremental change in assessment systems will foster concurrent improvement in professional and curriculum development” (Mpuangnan & Adusei, 2021, p.3). There should be assessment of broader range of mathematical abilities in addition to what is being assessed. Such abilities include representing, problem solving and understanding. Many nations have prepared curriculum guides to support a broader view of mathematical assessment. The Australian frameworks, for instance, suggested that numeracy should not include basic calculations competences only, but should also comprehensive and foster a link between comprehension and operation of number (Leonelli & Schmitt, 2012). Traditionally, mathematics assessments have tended to mean that mathematics is an activity connected to determining a quick answer using an already established, method that has been memorized (Bahr, Monroe & Mantilla, 2018; Gao, 2012). This has thus failed to represent the true nature of mathematics (Galbraith, 2016). “The measurement of decontextualized technical skills should be replaced with

measures that reflect what is known about what it means to understand and do mathematics” as stated in the Assessment Standards for School Mathematics [AAMT] (National Council of Teachers of Mathematics [NCTM], 2002, p. 32).

The Assessment Standards for School Mathematics (National Council of Teachers of Mathematics [NCTM], 1995) and the Principles and Standards for School Mathematics (National Council of Teachers of Mathematics [NCTM], 2000) stated that assessment items should elicit the type of knowledge and performance of mathematics that are valued and expected to be exhibited. Therefore, standards-based instruction and programme is completed by standards-based assessment policy (Dunbar et al, 2017).

For authentication of students’ assessment, a convergent reform in the curriculum and instruction of mathematics is what is needed. As such terms such as, “authentic assessment,” “alternative assessment,” and “performance assessment” have become cause of campaign for efforts to change the pattern of the nature and purpose of assessment. Mathematics teachers, particularly, are focusing on the use of PBA as not only a means to link assessment with new curriculum reform (Suurtamm et al., 2016) but also to improve the links between instruction and assessment of mathematics (Pelegriano, Chubowsky & Glaser, 2013).

The traditional assessment in mathematics has constantly revealed disparities in students’ performance in Ghana. Male students have constantly been found to perform better in mathematics than their female counterparts (Etsey & Gyamfi, 2017). Also, the WASSCE results that are published every year indicate

that students in the Category A schools outperform their colleagues in the other categories in all subjects including mathematics (WAEC, 2017 & 2018). There is therefore the need to look for an assessment method that could address the disparities which PBA has been suggested by Pelegrino, Chubowsky and Glaser (2013) and Suurtamm et al. (2016).

One of the obvious difficulties with PBA is the inability to design the assessment so that the items may be delivered at the level of each individual student (Pegg, 2013). Children of all abilities levels and background can be found in today's schools, and the teacher's job is to teach them. It is conceivable to employ a performance assessment that's well-designed and still get the wrong information. When the assessment task is either too challenging or too simple for the student being evaluated, this may occur. The purpose of this study is to develop and validate an assessment instrument in mathematics taking into considerations the traits typical of high-quality performance assessments in mathematics.

A number of difficulties have been encountered in mathematics classroom assessments in connection to PBA student achievement. Gao (2012) has outlined these difficulties, which among others include limiting mathematical ability to only recall of discrete pieces of mathematical information. Gao (2012) proposed that assessment be integrated into planned instruction and connected to students' real-world experiences in order to increase student achievement in mathematics. Sun-Geun and Eun-Hui (2015), Kone (2015) and Sung-Eun (2015) reported in

their studies that PBA has educational value as far as teaching and learning in the classroom are concerned.

According to Ghana's profile dimension of mathematics education, knowledge and comprehension account for 30% and knowledge application for 70%, respectively (Ministry of Education [MOE], 2012). This suggests that the application of knowledge is really what mathematics is all about. The general objectives of mathematics, which serve as the compass for mathematics education, specify that by the time students have completed their mathematics coursework at the SHS level, they should be able to apply their knowledge to real-world circumstances (MOE, 2012). This suggests that assessments of students in mathematics education must allocate 70% of their time to showing how well they can apply their knowledge of mathematics to practical problems (PBA) (MOE, 2012). This is the spirit of PBA indicating that mathematics education at the SHS level was designed to be that of PBA.

Hibbard (2017) stated “performance-based instruction and assessment represent a set of activity for the acquisition and application of knowledge, skills and work habits through the performance of tasks that are meaningful to real life situations and engrossing to students” (pg. 43). As an extension of the conventional fact-and-skill training to real-world application of the knowledge obtained, performance-based instruction and assessment result in a balanced approach. According to Brennan (2006), the potential usefulness of PBA resides in the test's realism and the fact that different people would approach the test differently, leading to various correct answers. The knowledge is put to use in

practical situations. Performance-based assessment is useful as a formative assessment (Asamoah-Gyimah & Anane, 2018). Nitko (2014), however, said that PBA may be used for summative assessment in a manner similar to that of the West African Senior Secondary Certificate Examinations (WASSCE).

It is believed that performance-based assessment, a modern form of assessment, addresses many of the problems with traditional assessment. The application of knowledge is the main focus of PBA. According to Nitko (2014), PBA is a type of assessment that involves students performing a task that asks them to apply their knowledge and abilities from a variety of learning. Students can demonstrate their level of learning through this. A PBA, in its most basic definition, is a sort of evaluation that calls for students to perform or produce something while demonstrating the "specific skills and competencies" they have learned. Ainsworth and Viegut (2006) defined PBA task as an "activity that requires students to construct a response, create a product, or perform a demonstration" (p.57). Performance evaluation examines a student's overall performance in achieving a learning objective by putting their knowledge and abilities from a variety of disciplines to use. Performance evaluation also supports numerous solutions to a task, leading to multiple correct answers.

Although some of the typical questions students answer at the SHS may be demonstrative or hands-on, it has been observed that they only have one right answer and are not authentic. A comparison of traditional assessments tasks and the newly developed PBA tasks is presented in Table 1.

Table 1-Comparison of Traditional and PBA Task

Traditional assessment	Performance-based assessment
a. Using a scale of 2cm to 1 unit on both axes, draw on a sheet of graph paper, two perpendicular axes Ox and Oy for $-5 \leq x \leq 5$ and $-5 \leq y \leq 5$.	At the wedding ceremony of Mr and Mrs Ayebine-Gyamfi, the photographer took a picture of the couples. The photographer realised
b. Draw on the same graph sheet, indicating clearly all vertices and coordinates	that the original picture (object) lies within the range of 1 to 5 on a Cartesian plane on both axes.
i) ΔABC with vertices $A(2, 1)$, $B(1, 4)$ and $C(-1, 2)$;	a. Record four possible coordinate of the picture
ii) the image of $\Delta A_1B_1C_1$ of ΔABC under a reflection in the line $y = 0$, where $A \rightarrow A_1$, $B \rightarrow B_1$ and $C \rightarrow C_1$	b. Using an appropriate scale, plot the ordered pairs and join the points to form a shape.
iii) the image $\Delta A_2B_2C_2$ of ΔABC under a translation by the vector $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$, where $A \rightarrow A_2$, $B \rightarrow B_2$ and $C \rightarrow C_2$	c. What is the specific name of the plane shape drawn?
iv) the image $\Delta A_3B_3C_3$ of ΔABC under an anticlockwise rotation of 90° about the origin where $A \rightarrow A_3$, $B \rightarrow B_3$ and $C \rightarrow C_3$	d. Rotate your picture through 90° anticlockwise about the origin to form image 1. Label your image appropriately.
v) what single transformation maps $\Delta A_1B_1C_1$ onto $\Delta A_3B_3C_3$	e. Using a scale factor within the range of -2 to 2 , enlarge your picture to form image 2. Label your image appropriately.
where $A_1 \rightarrow A_3$, $B \rightarrow B_3$ and $C \rightarrow C_3$? (WASSCE 2019, Q9)	f. Reflect your picture in the line $y=2$

Table 1 Cont'd

Traditional assessment	Performance-based assessment
<p>i. Using a ruler and a pair of compass only, construct</p> <p>i.(a) $\triangle ABC$ with $AB =7.5\text{cm}$, $AC =13.5\text{cm}$ and $\angle ABC=120^\circ$</p> <p>(b) locus, l_1 of points equidistant from A and B</p> <p>(c) locus, l_2 of point equidistant from B and C</p> <p>ii. using N, the point of intersection of l_1 and l_2 as centre, draw a circle to pass through the points A, B and C (WASSCE 2020, Q13)</p>	<p>There are three- sister communities in the Ahanta West District of the Western, Himakrom, Bonsokrom and Npanyinasa. The distance from Himakrom to Bonsokrom is 2km, the distance of Npanyinasa from Himakrom is 1600m. The bearing of Npanyinasa from Bonsokrom is 300°. The assembly intends to build a school for the three communities so that the school will be equidistant from the communities.</p> <p>Using a ruler and a pair of compasses only,</p> <p>a. Make a geometric construction of the communities and where the school will be situated.</p> <p>b. What is the distance of the school to Bonsokrom?</p> <p>c. What is the distance of Bonsokrom from Npanyinasa?</p> <p>d. What is the specific name of the shape formed by the position of the communities? (justify your answer)</p>

Source: Authors own construction (2020)

From Table 1, the sets of questions were on transformation and geometric construction respectively. It could be seen that even though both require students to perform the tasks, there is always one correct response to the traditional tasks as compared to the PBA where students' answers depend on their selected coordinates. The on-demand PBA task thus reveals individuals' true

performance. Again, the PBA task is linked to real life situation making it an authentic assessment. This helps students to transfer acquired knowledge and skills to solve real life problems.

According to Stone and Lane (2006), a well-developed PBA could decrease malpractices in examination in light of the PBA's advantages. The PBA, in contrast to the traditional modes of assessment at the SHS, is marked by various correct answers because each examinee's approach may vary, making knowledge sharing and copying challenging. Additionally, because the procedure would need to be developed in the examination room, students cannot use material brought into the examination room. It would be challenging to prepare answers in advance as a result.

Available evidence indicates that the incidence of examination malpractice has been on the fluctuates since 2009, in both the Basic Education Certificate Examination (BECE) and the West African Senior Secondary Certificate Examination (WASSCE) (WAEC, 2016). Table 2 shows the prevalence of examination malpractice in WASSCE from 2006 to 2019.

Table 2- *Statistics on Examination Malpractice in WASSCE from 2016 to 2019*

Year	No. Sat	No. of Cands. involved in Malpractice Cases	% of Cands. involved in Malpractice Cases
2006	120492	9872	8.19
2007	129479	4101	3.16
2008	131353	2160	1.64
2009	152584	3273	2.15
2010*	-	-	-
2011	148697	4209	2.83
2012	173655	3439	1.98
2013	409711	5653	1.38
2014	240662	8051	3.35
2015	267741	12746	4.76
2016	270318	11936	4.42
2017	287353	13793	4.80
2018	316999	2787	0.88
2019	346094	48,855	14.1
Total	2647232	130875	4.37

*Ghana did not present candidates for 2010 WASSCE for school candidates

Source: (WAEC, 2019)

From Table 2, it could be seen that after 2006, there was a reduction in the percentage of students involved in examination malpractice from 8.19% to 1.64% in 2008. However, the percentage rose thereafter to 2.83% in 2011 after which again reduced to 1.38% in 2013. Once again, the malpractice rose again up to 4.80% in 2017. Again, the percentage reduced to 0.88% in 2018. Unfortunately, 14.1% of the students had their results withheld because of alleged involvement in examination malpractices. It therefore means that measures such as cancellation of entire results, and imprisonment adopted to control examination malpractice

have not been effective. Between 2006 and 2019, the nation has registered a total of 130875 WASSCE candidates representing 4.37% of the WASSCE candidates between 2016 and 2019 who are involved in examination malpractice.

The number of students involved in examination malpractices is significant enough to have a negative impact on Ghana's educational system's credibility. The low use of PBA is thought to be a contributing factor in the high frequency of examination practises in SHS mathematics. Students would have to create their own original solutions for the on-demand PBA items based on their chosen strategy (Camilli, 2006; Cohen & Wollack, 2006). Due to the distinctive responses that PBA students provide, it will be challenging for students to copy from one another, teachers to solve questions before sending them to students, or invigilators to assist specific students.

In terms of validating the traditional mathematics items in the SHSs in Ghana, Annan-Brew (2020) reported that the classical test theory approach is mostly used. Validation of the instrument is done by the use of table of specification to check content representativeness. Reliability, discrimination and difficulty indices (for dichotomously scored items) are estimated for validation of the instrument. Also, expert judgement is used to ensure content relevance of the assessment results. Little is done about educational and catalytic effects as well as the feasibility and credibility of the instruments. This might be because they are not aware such important validations ought to be done. At the 2010 Ottawa Conference for assessment and evaluation, it was suggested that in addition to good psychometric properties, assessment instrument should have good

educational and catalytic effects as well as feasibility and credibility. Educational effect implied that the instrument should improve the teaching and learning process while the catalytic effect implied that it should be possible to provide immediate feedback on students' performance to stimulate learning. Feasibility of the instrument explained that it should be easy, flexible and convenient to use the instrument in classroom setting. Credibility on the other hand meant that the instrument should be able to produce trusted results. Initially, the criteria were applied in healthcare, however, they are applicable to classroom assessment instrument.

Validating assessment instrument in mathematics with expert judgement is potentially affected by gender and experience in which the expert found him/herself. For instance, it has been reported that males have better knowledge in mathematics than females (WAEC, 2017 & 2018; Etsey & Gyamfi, 2017). Against that background, there could be difference in the evaluation of mathematics instrument in terms of feasibility, credibility, educational and catalytic effects. Also, one's exposure to knowledge and practice (Experience) in mathematics would influence the way the person evaluates a new mathematics instrument (Iji & Omenka, 2014). Because of the disparities in the performance of students, assessor from a category of school where performance of students is high would likely express different view on the feasibility, credibility, educational and catalytic effects than those in schools where performance is low (Ewetan & Ewetan, 2015).

In validation of the traditional mathematics essay items in Ghanaian SHS, the psychometric for reliability are most estimated using the inter-rater procedure. This is because, Brennan (2002) stated that the strength of an assessment procedure lies in its ability to meet acceptable psychometric values. They are useful for judging the dependability of the assessment results. Despite the potential ability of PBA for SHSs, there are consistency issues with examiner and item. Brennan (2006) stated that PBA is effective when errors due to rater is considered. Performance-based assessment has multiple procedure to the various responses. This is potential for variation in performance.

Again, face validity and content validity are mostly used for essay questions of the traditional mathematics items. Face validity is a measure of the degree to which a procedure, especially a psychological test or assessment, appears effective in terms of its stated aims (Nitko, 2014). Mostly, content validity is subject to expert judgement. Construct validity (degree to which items reflect the construct being measured) of the instrument is also subject to face validity. In the Ghanaian classroom, the 'subject lead' ensures the face validity of the instrument. Content validity (representativeness of items on an assessment and relevance of the items to the content) is also done by expert, mostly the 'subject lead' or the classroom teacher. Chan and Malim (2017) and Hasnida and Ghazali (2016) stated that construct validity is strengthened by subjecting the items to Principal Component Analysis (PCA) and Confirmatory Factor Analysis (CFA) for convergent and divergent validity respectively.

It is therefore important that the PBA is subjected to reliability and validity checks. This is to guarantee that valid and reliable results would be produced from the newly developed PBA for SHSs. Soliciting information from mathematics teachers and examiners on the feasibility, credibility, educational and catalytic effects of PBA for SHSs is an additional way to validate the instrument.

Statement of the Problem

According to Arhin (2015), Brennan (2000), Burkhardt and Swan (2008), and other researchers, PBA in mathematics can help students learn by giving them feedback that encourages learning. Students are better prepared for external examinations when performance-based learning is used in the classroom (Nitko, 2014). Students can also apply their mathematical skills to real-life situations (Kone, 2015).

An observation of the traditional assessment items in mathematics at the SHS level shows that the assessment is mostly made up of two parts; objectives part and essay items. In every set of the traditional test, only about three out of a minimum of nine items have some characteristics of PBA items but the full characteristics: authentic, meaningful, hands on, demonstrative and individualistic. Such items are mostly on concepts such as geometric construction and graphs. Even though the items are performance based in nature, they do not represent real life experience as depicted in Table 1. Students as a result of lack of real-life application of knowledge learned, fail to recognise the significance of the things they are taught (Gyamfi, 2022a). This is due to the fact that educators do

not apply the idea to actual circumstances. Students only think about mathematics as the application of memorised rules to a few complex puzzles. Traditional assessment items in mathematics dominate mathematics assessment at the SHS level. There is no single test in mathematics at the SHS level which all the items are of the PBA type. This calls for the development of PBA for SHSs in Ghana.

Every year incidence of examination malpractices is reported in SHS examinations. This has raised a lot of concerns about the quality of education provided in Ghana. In Ghana, Sam (2012) argued that the nature of assessment items administered to students contribute to examination malpractice. Sam further explained that because most of the items have one specific procedure and answer, it makes collusion easier and undetected. This, in addition to other factors might explain the high level of examination malpractice especially in mathematics as depicted in Table 2. The researcher's personal experience as a mathematics teacher and WAEC examiner confirms this assertion of Sam (2012). Brennan (2006) and Sam (2012) stated that examination malpractice could be reduced through the use of the test itself. Both Brennan (2006) and Sam (2012) suggested the use of PBA which is individualistic in nature.

Etsey and Gyamfi (2017) and Gyamfi (2017a) argued that Ghana places premium on traditional assessment. Traditional assessment places emphasis on recall of facts rather than application of knowledge. Most of the items used in mathematics are those that require student to use already memorised rule to solve an abstract mathematical problem. However, mathematics is not about just solving problem but using mathematical knowledge to solve real life problems.

Globally, traditional assessment dominates performance assessment (Jiraro, Sujiva, & Wongwanich, 2014). It is typical in Africa. In Nigeria, Agu, Onyekuba and Anyichie (2013) confirmed in their study that most classroom teachers are familiar with traditional forms of assessment. These forms of assessment place emphasis on knowledge rather than the application of the knowledge. In Ghana, Ankomah (2020) reported that most classroom teachers use items under the knowledge and comprehension levels of the cognitive domain. Ghana's performance on the Trends in International Mathematics and Science Study (TIMSS) has continued to be poor (Anamuah-Mensah, Mereku & Asabere-Ameyaw, 2004; Burt, 2017; Butakor, 2016). The items of TIMSS are mostly on application of mathematical and scientific knowledge. This is because there is little room for application of knowledge in assessment in Ghanaian schools. There is therefore the need to consider alternate forms of assessment in mathematics against the backdrop of the weakness of the traditional assessments.

To strengthen the use of PBA for students, there should be PBA in mathematics developed and available for use. Wanner (2004) argued that most instrument available for use in the classrooms are of the traditional type. Only few PBA instruments have been developed for different subject areas and purposes. For example, Pishghadam, Baghdei and Shayesteh (2012) used Rasch model and item response theory to validate a developed instrument for rating students' performance in English. Burdis (2014) also validated a PBA for language development. Further, Pineda (2012) developed a PBA in communication which was scored with a rating scale, Estacio (2015) developed an instrument for

measuring performance in physics while Manning (2015) and Wyatt (2016) developed a questionnaire for family life skills using CTT for the validation. Wanner (2004) developed a PBA instrument for counselling. Rosaroso and Rosaroro (2015) developed a PBA test in science for higher institutions in Philippines. White (2017) also developed an observational instrument in PBA. Surprisingly, none of these PBA was in mathematics.

Further, Sun-Geun and Eun-Hui (2015), Kone (2015) and Sung-Eun (2015) evaluated the impact of PBA on students learning. While Sun-Geun and Eun-Hui's (2015) PBA was in science, Kone (2015) PBA was in oral presentation. Sung-Eun (2015) used a meta-analysis for the study. None of these studies developed a PBA in mathematics for students learning. All these studies were conducted outside Africa.

In Ghana, the only study which developed PBA is that of Arhin (2015). Arhin (2015) developed a PBA in mathematics for Form 2 science students in Ghana National College, Cape Coast in an experimental study. His study was to find out the effect of performance-based driven instruction on students' mathematics performance. The PBA was developed on some selected topics in mathematics. The psychometric properties of the items of the developed PBA were not estimated to establish the validity and reliability of the items.

The last stage in organising classroom assessment is evaluation (appraisal) of the assessment (Asamoah-Gyimah & Anane, 2018). The appraisal is done to ascertain whether the items functioned as intended; if the items were difficult for the students who took the test. Also, the test is evaluated for fairness, efficiency,

and practicality (Nitko, 2001). Nitko (2001) further stated that an assessment is evaluated to ascertain the difficulty and discrimination indices as well as the presence of bias or Differential Item Functioning (DIF).

Aside, the psychometric properties such as difficulty and discrimination indices, at the Ottawa Conference (2010), it was suggested that evaluation of assessment should go beyond the psychometric properties. The conference suggested educational effect, catalytic effect, feasibility and credibility as the additional parameters to be looked out for. The educational effect characteristic means that the instrument should facilitate the teaching and learning process (Ottawa Conference, 2010; Boursicot, Kemp, Wilkinson, Findyartini, Canning, Cilliers, & Fuller, 2020). This characteristic is what PBA does better than the traditional assessment. Catalytic effect of a good assessment suggests that the instrument provides feedback that stimulates learning. Researches (Arhin, 2015; Brennan, 2000; Burkhardt & Swan, 2008) have shown that PBA provide immediate feedback that stimulate students' learning better. The acceptability (credibility) feature of an assessment means that different stakeholders find the examination process and the results credible. Because PBA reveals students' true performance on an assessment, it passes this characteristic. Finally, feasibility means that the examination procedure is practical and realistic. They can be elicited from stakeholders of WAEC examination (students, teachers and examiners).

Performance-based assessment must be capable of meeting the criteria that define quality assessments (validity, reliability, fairness, transfer and

generalizability, content quality and coverage, meaningfulness, and cost/efficiency) and sound psychometrics (Werner, Denner, Campe & Kawamoto, 2012). These characteristics can be scaled down to three categories: those established by expert, those by psychometric statistical procedures and those by established stakeholders. Its therefore, means that if PBA is developed, administered, scored and interpreted to mimic these characteristics, it will become a good alternative to the traditional assessment used in SHSs in Ghana. It is therefore prudent to develop and validate a PBA instrument that could encourage students apply the knowledge in mathematics in real life situation.

Arhin (2015), Sun-Geun and Eun-Hui (2015), Kone (2015) and Sung-Eun (2015) in their studies evaluated the effect of their instrument on students' variable such as performance, motivation and attitude. This confirms the suggestion made at the Ottawa Conference in 2010 that, aside the acceptable level of psychometric indices, an assessment instrument should have a good educational and catalytic effects as well as the feasibility and credibility level. This means validation of assessment instrument comes at two level; 1) those that done through survey such as feasibility, credibility, educational and catalytic effects and 2) those that are via statistical procedures such reliability, difficulty index, discrimination index and bias or DIF.

Globally, with the exception of Sun-Geun and Eun-Hui (2015) in Korea, Kone (2015) in US, Sung-Eun (2015) in Korea and Arhin (2015) in Ghana, no other PBA instrument has been evaluated to check the feasibility, credibility, educational and catalytic effects of PBA in addition to its psychometric

properties. Even, with studies that looked beyond psychometric properties, none evaluated all the four parameters suggested. For instance, while Arhin (2015) and Kone (2015) looked at only catalytic effect, Sun-Geun and Eun-Hui (2015) and Sung-Eun (2015) looked at the educational effect. It is important that, the stakeholders of education, in this case, mathematics teachers and examiners evaluate the feasibility, credibility, educational and catalytic effects of assessment instruments they use in the classroom.

This study sought to develop and validate an assessment instrument in mathematics for SHSs. Developing an assessment instrument such the PBA requires much thinking and thought. In addition it calls for the collaboration of experts in terms of content, assessment to scrutinize both content and structure of the assessment procedure (Estacio, 2015). Development and validation of an assessment instrument using the Benson and Clark (1982) approach follows a four-phase stage; planning, construction, qualitative evaluation and validation (Manning, 2015).

There is no found research on a developed and validated PBA in mathematics for SHSs in Ghana. This study happens to be the first in validating PBA the educational and catalytic effects as well as the feasibility and credibility in addition to the psychometric properties. It is in view of this, that the researcher wants to validate a PBA instrument in mathematics tasks (developed by the author) for SHSs.

Purpose of the Study

The purpose of the study was to develop and validate a PBA items (developed by the researcher) for SHSs. The study also sought to find out the validation of instrument by the mathematics teachers and examiner differ by gender, experience and school category.

Research Objectives

The following objectives were formulated to guide the study. The study sought to find out the/if

1. feasibility of the developed PBA.
2. credibility of the developed PBA.
3. educational effects of the developed PBA on students.
4. catalytic effects of the developed PBA on students.
5. reliability of the developed PBA
6. validity of the developed PBA
7. there is a statistically significant difference in the feasibility of the PBA between examiners and teachers due to gender, school category and experience.
8. there is a statistically significant difference in the credibility of the PBA between examiners and teachers due to gender, school category and experience.
9. there is a statistically significant difference in the educational effects of the PBA on students between teachers and examiners due to gender, school category and experience.

10. there is a statistically significant difference in the catalytic effects of the PBA on students between teachers and examiners due to gender, school category and experience.

Research Questions

The following research questions were formulated to guide the study:

1. What is feasibility of the developed PBA?
2. What is credibility of the developed PBA?
3. What are the educational effects of the developed PBA on students?
4. What are the catalytic effects of the developed PBA on students?
5. What is the reliability of the PBA?
6. What is the validity of the PBA?

Research Hypotheses

The following research hypotheses were formulated to guide the study:

1. H_0 : There is no statistically significant difference in perceived feasibility of the PBA between examiners and teachers due to gender, school category and experience.
 H_1 : There is a statistically significant difference in perceived feasibility of the PBA between teachers and examiners due to gender, school category and experience.
2. H_0 : There is no statistically significant difference in perceived credibility of the PBA between examiners and teachers due gender, school category and experience.

H₁: There is a statistically significant difference in perceived credibility of the PBA between teachers and examiners due to gender, school category and experience.

3. H₀: There is no statistically significant difference in perceived educational effect of the PBA items on students between teachers and examiners due to gender, school category and experience.

H₁: There is a statistically significant difference in perceived educational effects of the PBA items on students between teachers and examiners due to gender, school category and experience.

4. H₀: There is no statistically significant difference in perceived catalytic effects of the PBA items on students between teachers and examiners due to gender, school category and experience.

H₁: There is a statistically significant difference in perceived catalytic effects of the PBA items on students between teachers and examiners due to gender, school category and experience.

Definition of Terms

Catalytic effect: The effect of an assessment instrument to provide feedback that stimulates learning.

Educational effect: The effect of an assessment instrument to facilitate the teaching and learning process.

Feasibility: The effect of an assessment instrument to procedure is practical and realistic.

Credibility: feature of an assessment instrument that different stakeholders find the examination process and the results credible.

Polytomous items: Items that are scored on an interval level. Responses to the items are scored from zero to the maximum score.

Basic Assumptions

The study was conducted based on the following assumptions:

1. students score on an item is independent of the scores on other items (local independence).
2. the PBA items measure a single trait (unidimensionality).
3. responses to the items on the questionnaire are normally distributed.
4. responses to the items on the questionnaires have equal variance.

Significance of the Study

A good understanding of a validated PBA will help education stakeholders to make informed decisions. For instance, Ghana Education Service and others who use test results to make decision in Ghana, such as test developers in mathematics must ensure that assessment tasks encourage students to apply knowledge to real life situations. There is the need for alternative assessment that would motivate students to learn. Performance-based assessment tasks may be used to achieve this.

It is also believed that the findings from this study would help testing institutions such as WAEC to validate graded response items. That is the study would serve as a guide in validating the validity of items which are polytomous.

Again, it is anticipated that the results of this study would help the classroom teacher develop and validate PBA in mathematics for SHSs. The results of this study would serve as a guide to the development and validation of PBA in Ghana.

Again, this study would help close the research gap in validating PBA which is made up of graded response items. This is because most research on PBA considered only the psychometric properties of an assessment instrument with the educational and catalytic effects as well as the feasibility and credibility.

Furthermore, when the instrument is moderated and piloted, it would be a model on PBA for a large-scale examination such as that of the WAEC which is also high stake. Many research on PBA have reported PBA as a formative assessment.

Delimitation

The study was confined to the development and validation of PBA. The use of PBA in the SHSs was not addressed by this study.

Again, the study considered only the on-demand type of PBA. The items were responded to at a sitting under the invigilation by the teacher. Performance-based assessment could also be extended but the study focused only on the on-demand type.

The study limited the application of on-demand task of PBA to mathematics at the SHS level. However, it could be applied to other subject areas at different levels of education.

The evaluation of the psychometric properties of the designed instrument was delimited to reliability (inter-rater). The difficulty and discrimination indices as well as biases were not estimated. This is because, the research design and the nature of the instrument could be estimated with the classical test theory.

Also, student population for the study was limited to only public SHS in the Western Region of Ghana. Private SHS were not part of the population.

Limitations

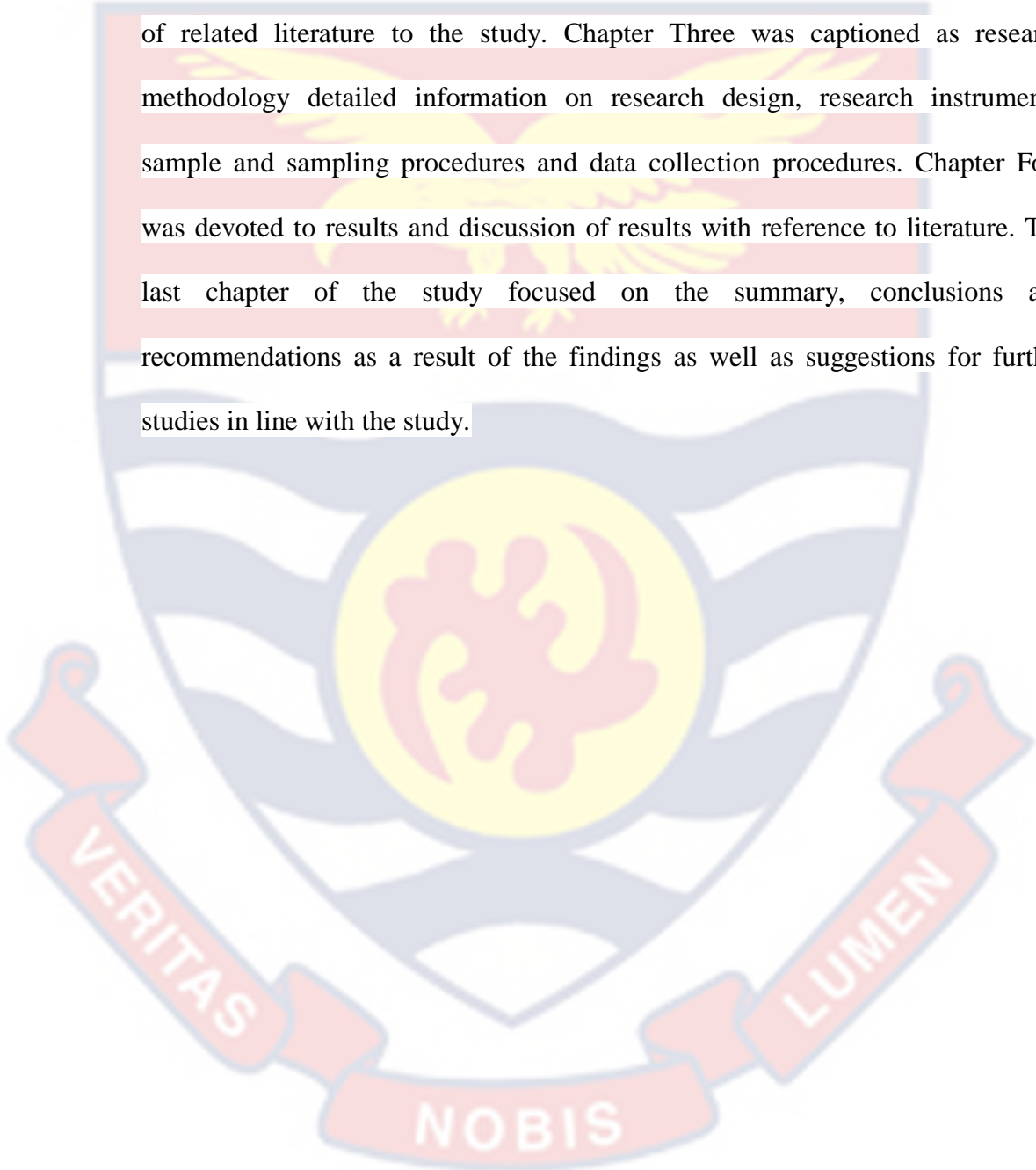
The study was delimited to only public SHS in the western region. This would limit the degree of generalization of the results. The findings cannot be generalized beyond public SHS and western region.

Difficulty and discrimination indices as well as biases were not estimated for the instrument. Difficulty and discrimination indices are essential psychometric properties needed to be validated on an assessment instrument. This is because the xcalibrre software that could perform analysis of the graded response (polytomous) items used in this study does not accept data beyond a threshold of 10. The data collected in this study has more than 10 thresholds. As a result, full validation of the instrument was not done.

Organisation of the Study

This thesis is structured into five chapters; Chapter One which was devoted for introduction focused on the background to the study, problem

statement of the study, purpose, objectives, research questions and hypotheses of the study, significance, delimitations and limitations of the study. Chapter Two which is the literature review details conceptual, theoretical and empirical review of related literature to the study. Chapter Three was captioned as research methodology detailed information on research design, research instruments, sample and sampling procedures and data collection procedures. Chapter Four was devoted to results and discussion of results with reference to literature. The last chapter of the study focused on the summary, conclusions and recommendations as a result of the findings as well as suggestions for further studies in line with the study.



CHAPTER TWO

LITERATURE REVIEW

The literature review is in four parts; conceptual flow chart, conceptual review, theoretical review and empirical studies. The conceptual framework made use of a flow chart to link up the concepts in the study based on the objectives outlined for the study. The conceptual review is in three sections. The first part focused on assessment with much emphasis on PBA, the second section looked into assessment in Mathematics and the last section looked at instrument development and validation. The theoretical framework also focused on validity and reliability theories. The last part of the literature review was on empirical studies on educational effect, catalytic effect, feasibility, credibility, reliability and validity of an assessment instrument.

Conceptual Flow Chart

This study sought to develop and validate PBA items in mathematics for SHSs. The conceptual flowchart seeks to draw the link of the various areas of validation of the PBA based on the research objectives for SHSs. It does not seek to establish any effect of any dimension of validation of the instrument. The conceptual framework of this study is presented in Figure 1.

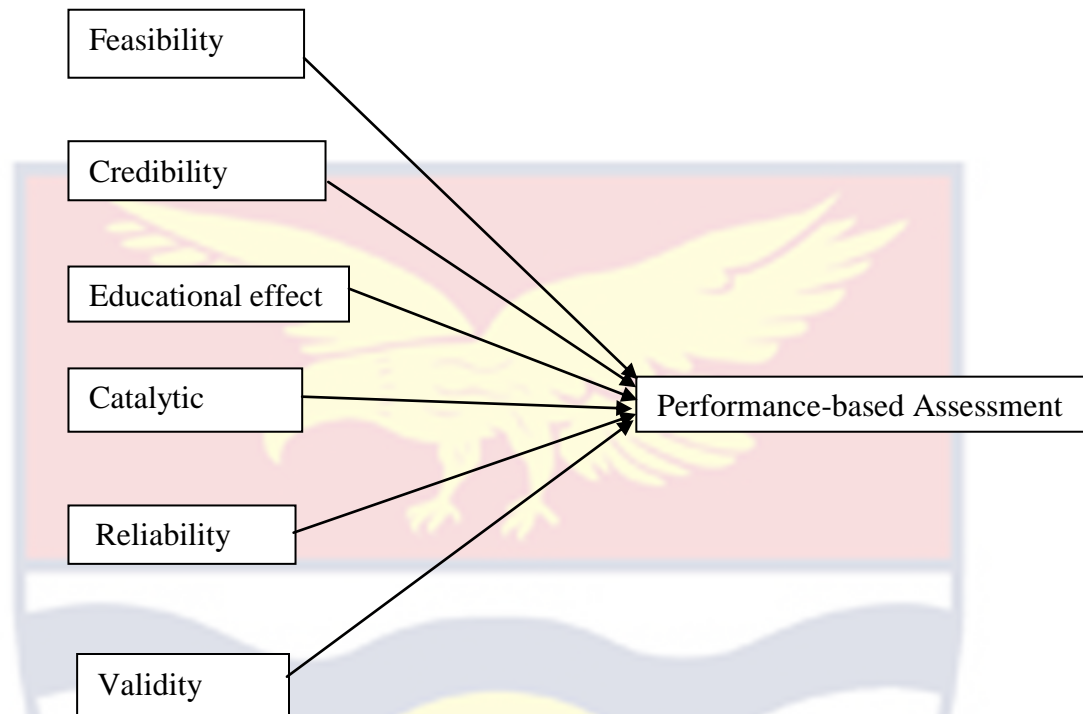


Figure 1- Flow chart for the validation of the PBA
Source: Authors' own design (2019)

Figure 1 shows the flow chart of the concepts of validation outlined in the objectives of the study. Validation of the instrument precedes the development of the instrument. The validation focuses on feasibility, credibility, educational effect, catalytic effect, reliability and validity of the instrument. The degree to which stakeholders (mathematics teachers and examiners) perceive the flexibility and convenience of use of the developed PBA will be measured as the feasibility of the instrument. The degree to which the stakeholders perceive the trustworthiness of results of the instrument is measured as credibility of the instrument.

Stakeholders' expression of the influence of the instrument on the teaching and learning process is measured as educational effect while the expression of

the ability of the instrument to provide immediate feedback to stimulate students' learning explains the catalytic effect of the instrument. Reliability measures the degree of consistency of the results of the instrument. The inter-rater reliability was used for the evaluation of consistency of results with the instrument. Validity is a measure of the appropriateness of the interpretations and uses of results of the instrument (Nitko, 2014). Both the convergent and divergent construct validity as well as the content validity of the PBA are considered in this study. Hence the focus of this research was to develop and validate PBA task for SHSs.

Concept of Assessment

The term assessment is used in every institution in recent times. Every organisation now seeks to examine the worth of either policy, product, staff, students in the case of education and many more. According to Heale and Twycross (2015) and Etsey (2012), assessment is a process of obtaining information for decisions making. This explains why practically every organisation uses the idea of assessment. Assessment occurs anywhere decisions are made based on information gathered. Assessment in schools refers to the process of gathering data to guide decisions on students, programmes, policies, and curriculum. Nitko (2012) therefore defined assessment as a “systematic process of gathering information that is educationally relevant to make legal and instructional decisions about the provision of special services” (p. 99). The definition focuses on education. Nitko continued that, assessment has stages, activity and outcome. The stages are the processes the assessor goes through, the activity is the gathering of the information and the outcome concerns the decision

made as a result of the gathered information on the phenomena. To Heale and Twycross (2015) and Etsey (2012), assessment covers a range of procedures which could either be formal as in pencil and paper test or informal as in observation, interview and the likes for obtaining information about students' learning as in the education setting. Assessment is systematically collecting information about students' performances and serves as an indispensable part of classroom teaching and learning (Dhindsa, Omar & Waldrip, 2007; García-López, González-Víllora, Gutiérrez, & Serra, 2013).

Educational assessment is a “systematic process of gathering and making use of data on the knowledge, skill, attitudes, and beliefs to strengthen programmes and improve students' learning” (Allen, 2004, p.23). Assessment information can be obtained empirically or already existing data from which one can make inferences. Based on its usefulness, assessment has undergone many changes for the purpose of obtaining educationally-relevant information for making decisions about the student and educational programmes. For example, the first form of assessment was for one person at a time but with time, research and development into assessment; saw the birth of mass examination and assessment for school pupils as it pertains in Ghana now (Allen, 2004; Torrance, 2005).

Relationship between Assessment and Teaching and Learning

According to Hodges (2014), facilitating students' learning is the principal goal of any educational programme. In educational programmes, assessment and student learning and performance are like the Siamese twin. Struyven, Dochy, and

Janssens (2005) were of the view that the effect of classroom assessment could significantly be inferred from the performance of students. The approach of students to learning is a determiner of their perception about classroom assessment. Pellegrino and Goldman (2008) and Shepard (2000) believed that improvement in classroom assessment to ensure validity and reliability would lead to improvement in learning. It is reported that significantly, assessment affects the approach students adopt to learn. As a result, assessment patterns are being shifted from assessment of students' learning to assessment for students' learning (Nitko, 2012; Resnick & Resnick, 2001).

According to Goodrum, Hackling, and Renni (2001), assessment enhances learning, provides feedback on student progress, helps students build self-confidence and self-esteem, and offers them evaluative skills. Additionally, when instruction, assessment, and results are interconnected, successful learning is assured. As a result of its direct connection to teaching and learning outcomes, assessment is crucial to learning.

According to Goodrum, Hackling, and Rennie (2001), assessment is a crucial part of the teaching and learning process. This suggests that efficient evaluation in the classroom is necessary for efficient instruction and learning. The issue is that teachers only employ a small number of assessment methods. However, in actuality, teachers rarely employ formative assessment to guide their planning and instruction (Goodrum, Hackling & Renni, 2001).

Hodges (2014) indicated that providing feedback and guidance on students' learning are basically the purpose of assessment. The purpose of the

assessment influences the direction and magnitude of the guidance and feedback that are provided to the students. Assessment thus prepares students for life. The premise is that, aside formal education, learning occurs throughout life (Etsey & Gyamfi, 2017). Based on the effect of classroom assessment on learning, Etsey and Gyamfi pointed out that assessment should assist the students to have a better perspective of their metacognition by providing feedback to them and discouraging them to depend on others for knowledge of their own level of learning.

Even though students' involvement and participation in learning activities have improved as a result of the increase in student-centred approaches to learning, it has not been able to cause the same change in curriculum and assessment practices which can contribute to the desirable outcomes required for lifelong learning (Taras, 2002). Based on the impact of assessment on learning, Taras (2002) suggested that assessment employed by teachers should be capable of producing confident, independent and autonomous learners. Boud and Falchikov (2006) argued that designing the assessment practices for current learning related to the curricula without considering the place of assessment in learning beyond the classroom has been a major challenge in assessment of mathematics.

Boud and Falchikov (2006) further stated that much attention is needed to ensure learning beyond the classroom aside the traditional purposes of assessment for certification and instructional management decisions. Modern assessment should move away from the traditional approaches of assessment in which the

teacher has the sole responsibility of determining what is to be learned, the assessment tasks and its criteria, how the task is to be performed by the student, and the grade that would be awarded to performance. The traditional assessment approaches make the student passive, rather than being active in the assessment instead of assessment practices which are sustainable and that can help prepare (Brennan, 2006) students for lifelong learning beyond the classroom. The student should be dressed up to engage in metacognition (make their own judgments about themselves, their performance as well as their learning) (The Duke Endowment, 2002). Sustainable assessment is one that considers how it contributes to the preparation of students for the future by helping the students to develop self-regulation and development as its promote active student participation.

Forms of Assessment

Due to the usefulness of assessment, it is used in different situations and purposes. Assessment is used for by organization on a day to select suitable applicants for jobs, it used by the same organization over a period of time to obtain information for promotions. In the classroom situations, assessments are equally used for different purposes. As a result, assessment is seen differently by different people based on purpose or scope or format. This places assessment into different forms. Assessment collects information on the learner and the learning community, a course of an academic programme, institution, or the entire educational system. This means that assessment covers all areas of the school environment.

According to Gordon (2008), the use of assessment has advanced the process of teaching and learning. Gyimah, Ntim and Deku, (2012) listed two types of assessment; informal assessment and formal assessment. This is classification of assessment by structure or form. Informal assessment according to Gyimah, Ntim, and Deku, is the form of assessment without any formality. It is very flexible and done without any strict rules or regular form. It can be used at any time without interfering in the instructional time. Smith, Polloway, Patton and Dowdy (1995) also ascertained that informal assessment are usually loosely structured techniques which are more closely tied to teaching. Its purpose is to direct instruction and therefore a process. According to Gyimah, Ntim, and Deku, there are two forms of informal assessment: (i) Those that utilize test items such as teacher- made test, curriculum-based assessment, portfolio assessment and others; (ii) Those that do not utilize test items such as ecological assessment, observation, interview, checklist, rating scales and others.

Formal assessment procedures, on the other hand, are the assessments that are more structured with specific rules for item construction, administration, scoring and interpretation of the results (Nitko, 2012). This means that formal assessment unlike informal assessment has specific time for administering well developed test manual that specifies the test and item specifications. Also, it has its accompanying scoring rubrics. Formal assessment procedures include achievement tests and standardized tests (Gyimah, Ntim, & Deku, 2012).

Formative assessment

According to Asamoah-Gyimah and Anane (2018), formative assessment is the form of assessment that is done continuously throughout the lesson. This means that formative assessment occurs before, during and after the instruction. It can, therefore, be said that formative assessment is tied to the classroom instruction. The aim of formative assessment is to find out how the lesson is going (Asamoah-Gyimah & Anane, 2018). This implies that the idea behind formative assessment is not to grade students' performances but to improve their learning. Formative assessments in the classroom include classroom questions and answers, class exercises, homework, observations, quizzes and class tests. Airasian (2001) also defined formative assessments as the forms of interactive assessment primarily used to alter a process or activity which is ongoing. Formative assessment is concerned with enhancing students' motivation to learn with the purpose of producing work of high quality or thinking (Wang, French & Clay, 2015).

Edmund (as cited in Cullinane, 2011) stated that the teacher and the student are considered as the two different players in formative assessment. Many teachers who are concerned about formative assessment use the method to check for students' understanding. This is done by asking questions and or by observing students in the classroom. In formative assessment, teachers informally collect information that will enable them to determine next line of action in teaching. The teachers are thus the data users of formative assessment. On the part of the students, formative assessment helps students to know what would stimulate their

responses to teachers' questions. The prime aim of formative assessment is about providing immediate feedback to students about what they have learnt. The feedback provided to students, if effective, can significantly increase students' achievement (Marzano, Pickering & Pollock, 2001).

According to Suurtamm et al., (2016), formative assessment could be done from the beginning to the end of instruction. For any course or programme, formative assessment could be used as the tool to provide immediate evidence of what student have learnt. In the classroom, formative assessment happens to be one of the most popular forms of assessment that teachers use and the aim is to enhance the quality of students' learning (Suurtamm et al., 2016). As an important component of teaching and learning, classroom formative assessment can influence the modification of the course or circular when a particular course has not met the students' learning outcomes (Suurtamm et al., 2016). According to States, Detrich and Keyworth (2018), in the classroom, formative assessment also provides important information on programmes that need to be examined if the learning goals and objectives of those programmes have been met in all sections of the course (Bardes & Denton, 2001).

Strengths of formative assessment

Formative assessment information contributes to an overall plan of assessment by helping to identify particular areas in a programme for assessing learning and monitor the progress made with regard to the learning outcomes (Bardes & Denton, 2001).

According to Nitko (2014), formative assessment provides an excellent picture of students' performance over a period of time. Because formative assessment is continuous, several other previous performances of the students are available. Analyses of this information give a clear picture about the performance. Based on that there is enough evidence to say a particular student is good or weak.

Another strength of formative assessment as a classroom assessment is that that it encourages the students to constantly study throughout the period of instruction (Asamoah-Gyimah & Anane, 2018). The oral questions and answers, homework, class exercises and observation that are constantly and continuously used in the classroom means that students should always be on alert and this compels students to study and pay attention in classroom throughout the periods of instruction.

Mussawy (2009) also stated that formative assessment enables the teacher to identify the weakness of individual students. The continual assessment of the students periodically on a particular content helps the teacher to identify students who after the entire lessons still have weakness in grasping the concept. With this, the teacher can plan a remedial and individualized teaching for such students. The students are therefore helped to progress hence improving on their performance.

Weakness of formative assessment

According to Mussawy (2009), one of the weaknesses of formative assessment is the increase of workload on the classroom teacher. In order to obtain clearer picture on the performance of students, continuous and comprehensive assessments have to be done. This means that almost every day,

the teacher has to give assignments, home works, or class exercise and mark to provide students with immediate feedback on their performance so as to know what to do with regard to their performance. This is challenging especially in Ghanaian schools where the classroom teacher has a large class size to handle.

Asamoah-Gyimah and Anane (2018) also stated that effective and efficient formative assessment requires some professional skills which many classroom teachers lack. Effective and efficient assessment requires that the assessor adhere to all principles and practices of assessment. Deficiency in adhering to the principles and practices of assessment means deficiency in the assessment carried out in the classroom and this implies that any decision made based on the assessment is also deficient in a way.

Another problem with formative assessment is of record maintenance. Collection and storage of records are crucial in formative assessment to understand the progress of a student performance (Amedahe, 2012). In most schools in Ghana, adequate storage facilities are not available. There are not adequate cabinets and computers in the schools for storing formative assessment data. This makes handling and retrieval of formative assessment data for use very difficult.

Summative assessment

Summative assessment attempts to find out if a student has mastered the desired goals of learning or achieved the prescribed criteria (Edmunds, 2006). That is, it seeks to measure how much knowledge, skills and attitude students' have achieved at the end of a course of study. Usually, summative assessments

occur at the end of the course and details students' level of learning. As a result, grades are assigned to students' performance as a reflection of how well a student has reached the key instructional goals or outcomes. Basically, the aim of summative assessment is determining the level students are in terms of the content and thinking. Therefore, scores and grades assigned correlate the level of mastery of knowledge, skills and attitude of the student. This makes summative assessment judgemental.

Summative assessment in the classroom according to States, Detrich and Keyworth (2018) is comprehensive and is used to determine the level of students' learning at the end of the programme. The goals and objectives of classroom summative assessment usually shows the cumulative nature of the learning that takes place in a programme (Suurtamm et al., 2016). It is therefore relevant to have summative assessment at the end of the programme to find out if students have acquired the programme goals and objectives. Bardes and Denton (2001) articulated that for thorough information, the use of various methods and measures via summative assessment is key. In Ghana, Basic Education Certificate Examination (BECE), General Business Certificate Examination (GBCE), Advanced Business Certificate Examination (ABCE) and West African Senior Secondary Certificate Examination (WASSCE) are the known summative assessment used. However, other international summative assessment such as Scholastic Aptitude Tests (SAT), General Records Examinations (GRE) and Test of English as a Foreign Language (TOEFL) are also available. Also, end of

semester examinations of tertiary institutions are also example of summative assessment.

Strengths of summative assessment

According to Asamoah-Gyimah and Anane (2018), a major strength of summative assessment is the measurement of students on a larger sample of content. Summative assessment attempts to obtain information on students' overall gains at the end of a course. Comparing to formative assessment which is limited to only the instructional goals, summative assessment covers almost all concepts learned and this gives students opportunity to at least provide an answer to a question.

Additionally, Brennan (2006) ascertained that summative assessment has the advantage of providing enough evidence for placing students into advanced courses. That is, summative assessment enables placement decisions to be made. Summative assessment provides information on students' overall mastery in the course and thus information obtained through summative assessment is enough to decide whether a student is equipped to take up an advanced course or needs a remedial assistance.

Nitko (2001) stated that summative assessment enables the classroom teacher to evaluate his/her own teaching. Students' performance on summative assessment which tends to cover content of the entire course or programme gives information on how well the teaching and learning process has been. If students genuinely perform well on the summative assessment, it is an indication that the

teaching was successful else the teacher would have to modify the teaching strategies and methods.

Weakness of summative assessment

Asamoah-Gyimah and Anane (2018) noted that, summative assessment is less directed to providing suggestions for improvement in students' learning. That is summative assessment unlike formative assessment that provide immediate feedback to students on their performance in order to understand the errors or weakness in students learning, summative assessment only give final score or grades to students. Details of students' performance are not provided on strength and weakness in the performance. Also, it takes a long time for students to get information on their performance which students are given opportunity to discuss the performance. This is because, after summative assessment the concepts are not revisited. The grades are used to judge students rather than helping them improve performance on those concepts.

Brennan (2006) stated that, summative assessments are mostly associated with examination malpractice than formative assessment. Because summative assessment tends to be one shot examination used mostly used for critical decisions such certification and selections. For that reason, students are always poised to passing the examinations. Therefore, all means including foul ones are used by students. Because formative assessment is not judgemental but only to improve learning, students are not minded with cheating in those examinations.

According to Nitko (2001), a principal disadvantage of summative assessment is that a great of time is required in developing the assessment

instrument. Summative assessment covers a large range of content domain; therefore, the test developer has to construct a test that will cover a representative sample of content domain and also all level of cognitive domain. This requires a more time to do. Table of specification has to be constructed to ensure this and objectives that are mostly included to ensure content representativeness. Objective test also comes with its challenges as far as time is concerned.

Performance-based Assessment

Performance-based assessment (PBA) as a contemporary form of assessment is perceived to address many of the challenges associated with the traditional assessment. The focus of PBA has to do with application of knowledge. According to Nitko (2014), PBA is a form of assessment that presents a hand on task which requires students to perform activity that requires application of knowledge and skills from several learning. It allows students to show how well they have learnt. Basically, a PBA is one that students are required to show that they have acquired specific skills and competencies which are evident in what they perform or produce. Ainsworth and Viegut (2006) defined PBA as an “activity that requires students to construct a response, create a product, or perform a demonstration” (p.57). Performance-based assessment deals with the overall experience of a student in performing a learning target by applying their knowledge and skills from several areas. Performance-based assessment also lends itself to multiple procedures to a task therefore resulting in multiple correct responses (Topping, 2015; Arias-Estero & Castejón, 2014).

Performance-based assessment could be used as a summative assessment

procedure to document not only students' knowledge on a topic, but their ability to apply the knowledge in a "real-world" situation (Brennan, 2006; Adib, Rusilowati & Hidayah, 2018). Performance-based assessment becomes authentic assessment when it reflects real life situations and meaningful to students learning. This means that all authentic assessment tasks are performance-based tasks. By asking students to produce an end product, PBA causes students to reorganize their knowledge and use their skills to apply the knowledge in a new set of situations capable of occurring outside the normal classroom (Palm, 2008; Shavelson, Baxter & Pine, 2009). Performance-based assessment includes designing and constructing a model and developing, solving a mathematical problem that mimic real life situation by applying knowledge and skills. Also, students can undertake and report on a survey, conduct a science experiment, write a letter and create and test a computer programme (Darling-Hammond & Pecheone, 2019; Wren, 2009).

Whatever the type of performance, performing an authentic task that excite a real-life experience and imitate real world challenges is the common factor in all PBAs (Wiggins & McTighe, 2015). Performance-based assessment is used in numerous countries and has numerous advantages which are not offered by traditional tasks. Wiggins and McTighe (2015) asserted that, in fact, authentic assessments go beyond just testing to teaching students and their teachers what goes into performing of a subject (Falk, Ort & Moirs, 2007; Shepard, 2009).

Performance-based assessment as a formative assessment provides timely feedbacks than traditional classroom large-scale standardized tests (VanTassel-

Baska, 2013). This is because standardized tests could last for months to produce feedback, but PBA permits teachers to make significant modification while their current students are being taught (Darling-Hammond & Pecheone, 2019). In addition to the impacts of PBAs on student outcomes, the implementation of PBAs procedures could also inform classroom instructional strategies. Though it could be challenging to effect change in the patterns of general teaching and learning under some circumstances such as large class size, PBAs could change particular behaviours and activities in the classroom such as motivation and participation (Topping, 2015).

Assessment policies and practices at all levels are seeing rapid transformation. Complex performances of the traditional assessment are being used as the foundation that is guiding current wheel to change assessment. Examples include the recommendation to use more of essays, open-ended problems, computer simulations of real-world problems, hands on science problems, and students' portfolio. Collectively, these assessment forms are called "authentic or performance" assessments (Werner, Denner, Campe, & Kawamoto, 2012). The term suggests performance of tasks considered to be of importance. In contrast, paper-and-pencil, multiple-choice tests and some essays and computational problems are difficult to mimic real life situations. Being able to transfer classroom learning to real life situations is an indicator and goal of learning. The worse aspect is that, the procedures that may help in achieving the goal become distorted. The lack of correspondence between classroom learning and real-life situation has become an increasingly important concern in

assessments. The resultant is an increased in significant motivation for the recent calls for “authentic” assessment.

Although authentic assessment seems new, standard guidelines from some measurement specialists have been there for a long time. For example, Erzoah, Gyamfi, Yeboah and Langee (2022) argued that “it should always be the fundamental goal of the achievement test constructor to make the element of his test series as nearly equivalent to, or as much like, the elements of the criterion series as consequences of efficiency, comparability, economy, and expediency will permit” (p. 2). With regard to the construction of items for measuring critical reasoning skills and higher-order thinking, Erzoah, Gyamfi, Yeboah and Langee went on to note that “the most important consideration is that the test questions require the examinee to do the same things, however complex, that he is required to do in the criterion situations” (p. 4).

Performance-based assessment tasks

Performance-based assessment assesses either the process or product or both (Brennan, 2006). When it is difficult to assess the processes, only the product is assessed, and when the product is embedded in the process, the focus is placed on the process. It is also possible to assess both process and product (Stone & Lane, 2006). Stone and Lane further stated that PBA could be task-centred when the knowledge and skills that contribute to the proficiency of the task is not specified in advance but specified when preparing the scoring rubrics. Performance-based assessment could also be construct-centred when the set of knowledge and skills to be assessed are valued in the instruction of the task.

Brennan (2006) stated that the strength of PBA lies in ability to have good psychometric values.

Aside being either task-centred or construct-centred Stone and Lane (2006) and Nitko (2014) also stated that PBA could be on-demand task or restricted responses task which requires students to create responses within a short period of time. Performance-based assessment could also be an extended task which lasts for a longer time undertaken by students on an assigned topic like thesis or project work.

Performance assessment has multiple correct procedures to a task therefore has multiple correct responses (Stone & Lane, 2006). This characteristic tends to reduce copying from colleagues or teachers copying answer to students since they cannot have the procedures written for each student. Also, performance assessment requires students to perform the tasks which cannot be done by a third party. Some of the performance assessment task are limited to an individual student therefore, leakages and copying and their source could easily be detected. Students are required to report on the procedures that were used in completing the task (Stone & Lane, 2006).

Scoring Performance Assessment

There are three methods of scoring performance assessment- analytical, holistic and primary trait (Stone & Lane, 2006). The type of a scoring procedure depends on the purpose of the assessment, the constructs being measured and nature of the intended interpretation of the scores. With a holistic scoring, the rater makes a single (one) score to judge the performance. Holistic scoring

describes the overall effect of the characteristics. With the analytic scoring, the task is divided in parts and weights are part on each part. Students are scored on each part according to the weight of the parts. The sum of the scores of the parts of the task gives the overall standing of the student on the task (Office of Educational Research and Improvement, 2009). For the primary trait scoring, one or more relevant trait of the task is identified. Relevant construct of the task is identified and scored in the primary trait scoring. This scoring procedure allows for a general criterion to be tailored to the task allowing for more consistency in raters' application of the rubric.

Nitko (2014) also suggested the top-down approach to scoring performance assessment, where a conceptual framework of the achievement is developed, a detailed outcome of the performance is also identified then a general or specific scoring rubric is prepared for scoring. He also mentioned the bottom-up approach to scoring performance assessment where samples of students' works of degree of quality are used as standard for scoring

How mathematics should be taught and assessed

According to National Council of Teachers of Mathematics (NCTM) (2010), assessment that improves learning of mathematics should be a usual part of on-going classroom activity rather than a hiatus. Assessment is a means to an end and “does not simply mark the end of the learning cycle” (Nitko, 2006, pg. 134). Rather, assessment should be fused into the teaching and learning to encourage and support further learning. Naturally, in every lesson, there are opportunities for informal assessment (Ankomah, 2020; Kamaldeen, Buhari &

Parakoyi, 2012) They include listening to students, observing and making sense of what students say and do in the class. For young children in particular, the observation of students' work brings to bear the qualities of thinking which written or oral activities cannot reveal (Schoenfeld, 2000). Teachers should look out for different assessment opportunities when planning instructions and making decisions about instructions (National Council of Teachers of Mathematics [NCTM], 2010). Questions such as the following should constantly be part of the teachers' planning: "What questions will I ask?" "What will I observe?" "What activities are likely to provide me with information about students' learning?" Gao (2012) stated that "preparation for a formal assessment does not mean regular instruction should pause and resort to teaching to the test" (p. 9). On-going teaching and learning is the best preparation for assessment for students. Similarly, for teachers, the foundation of the best teaching is on-going assessment. This is the way to go with mathematics.

According to Gyamfi (2017a), mathematics is not all about doing, solving problems, performing algorithms but includes an element of appreciation. Appreciation of mathematics involves having a qualitative comprehension of some of the key concepts of mathematics such as proof and structure. The instructional process of mathematics should not be restricted to only the cognitive and psychomotor domains of learning but to the affective domain as well. That is students should be made to understand the principles of the subject in order for them to have a rational understanding of the concepts.

According to the California Department of Education (2013), United States' schools and schools in other parts of the world have now prepared different reforms that detail what students should learn and demonstrate in mathematics as students move through the levels. For example, the California Mathematics Framework, the California Mathematics Standards, and National Council for Teachers of Mathematics have detailed the guidelines and Standards for School Mathematics (California Department of Education, 2013). These documents rally support for assessments that gives attention to students' ability to understand as well as their procedural skills. As detailed in the standards, assessment should measure:

1. Computational skills as well as the application of these skills in familiar and unfamiliar contexts;
2. The use of mathematical processes in context;
3. The use of mathematics to make sense of complex situations;
4. How well students formulate hypotheses, collect and organize information, and draw conclusions and
5. How well students communicate their mathematical reasoning both verbally and in writing (California Department of Education, 2013.)

Assessments that improve learning of mathematics alongside activities that are consistent to teaching are useful. For example, when students learn by communicating their mathematical ideas through writing, the assessment of their knowledge on that particular concept of mathematics should be done by having them write about their mathematical ideas. If the students learn the concept in

groups, the assessment should as well be done in groups. If graphs and calculators are used in teaching, they are to be available for use during assessment. These guidelines are in the domain of PBA.

Mathematics achievement as a psychological construct makes it difficult to be assessed with only one method. According to Crocker and Algina (2008), assessment of psychological construct is associated with problems such as:

1. Inability of a single approach to measure it,
2. usually based on limited sample of the behaviour,
3. lack a well-defined units of measurement scales
4. constructs cannot be defined in terms of operational definition but must also show relationship to other constructs.

The National Council of Teachers of Mathematics (1995) posited a number of classroom activities that are indicators of mathematics learning: oral comments, drawings, models, and other means of representing knowledge. These evidences are useful to the teacher and student, in addition to information from more formal assessment activities, to determine next steps in learning. Activities ranging from scrabbling through to estimating the length of wire for fencing are all evidence of mathematics learning. Continuously assessing the work of the students facilitates their learning, understanding and communication. Moreover, external assessments provide support to the classroom instruction. For classroom work, the teacher's judgments, and students' reflections are considered to be parts of an external assessment. This external assessment enhances students' learning of mathematics. The instructional goals and the assessment are levelled.

Theoretical Review

The purpose of the study was to develop and validate the PBA items in mathematics for SHSs. The theories that support the study are

1. Reliability
2. Validity

Reliability

Reliability is defined as the “degree of consistency between two measures of the same thing” (Yeboah, 2017, p. 35). It is the degree to which assessment results would be similar under the same or slightly different measurement conditions (Feldt & Brennan, 2001). For instance, if one assesses a student twice using the same or similar instrument, it is hoped that almost the same score would be obtained if one assesses the student one day later. Here, if one measures a person’s level of achievement with similar but not identical items, similar scores are anticipated even if under different administrators, using different scorers.

Reliability Theory

Reliability theory emanated from the works of Edward Lee Thorndike in the 1904. Therefore, his works became the foundation of the classical test theory (Crocker & Algina, 2008). Further expansions were made to the works of Thorndike by Spearman. Mathematically and based on logic, Spearman argued that test scores are imperfect measures of human traits. Spearman explained that the test score which is the correlation between the imperfect test scores (error) and the true value is low (Spearman cited in Ankomah, 2020). This argument raised the issue of error which is foundation of the reliability theory of the classical test

theory which attracted much attention and study. Any observed test score, according to traditional test theory, is a function of two hypothetical components: a true score and a random error. It is stated mathematically as $X = T + E$, where X is the observed test score, T is the individual's true score, and E is the random error. The observed score is the one that appears on the exam paper.

When the construct is measured repeatedly, the real score is the predicted value of the actual value of the observed score. The error score is the discrepancy between the observed and true score of an individual. This therefore means that it is the error that distort the equalization of the true score and observed score. When the error is neutralized, individual's score true score and observed will be the same when measured repeatedly. "Reliability is theoretically defined as the ratio of the variance of the true score to the variance of the observed score" (Amedahe & Asamoah-Gyimah, 2015, p. 78). Mathematically, it is expressed as

$$p_{xx}^2 = \frac{\sigma_T^2}{\sigma_X^2}$$

This implies that reliability tells the extent to which the observed score variance is close to true variance. A perfect reliable test is one with zero error score and that observed score and true score are equal. The reliability coefficient of perfect reliable test is 1.0. As the error increase, the reliability reduces. The strength of reliability thus lies in the ability to control error in the test. The American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (cited in Annan-Brew, 2020) define reliability as "the degree to which test scores are

consistent, dependable, repeatable, that is, the degree to which they are free from errors of measurement” (p. 93). When assessment instruments have the following characteristics: mismatch between learning objectives and test items, test with few items and irrelevant items (Amedahe & Asamoah-Gyimah, 2015), error scores are introduced to students' scores and thus into the reliability of the test scores (Amedahe & Asamoah-Gyimah, 2015). Some of these problems, such as a mismatch between learning objectives and test items, tests with few items, and irrelevant items, make it difficult for students to understand what is being measured. As a result, students have more difficulties and their grades suffer as a result. When these inaccuracies are introduced to students' results, they can no longer be used to make informed decisions (Ankomah, 2020).

Correlation Coefficient

Correlation coefficient is an indication of the nature and magnitude of the relationship between two variables measured. It ranges between -0.1 and 1.0 . The coefficient is either 1 or -1 for variables that are in perfect linear relationship. The direction of relationship as positive or negative depicts the operational sign. A zero-correlation coefficient implies no linear relationship between the variables. Several types of correlation coefficients are available for use depending on the nature of the variable. These include: Pearson product moment correlation coefficient, Spearman rank correlation coefficient, Interrater correlation coefficient, Phi correlation coefficients, Biserial correlation coefficients and Point biserial correlation coefficients. The most widely used is the Pearson product-moment correlation coefficient.

Methods of Estimating Reliability

The source of error under consideration gives the different methods for estimating reliability (Liaquat, Asif, Siraji & Maroof, 2012). A number of methods are available for estimating reliability, but the most commonly used ones are:

Test-retest method

The test-retest technique is a measure of stability that takes into account student scores over time. The same test is administered to a group of students twice over a period of time ranging from minutes to years. The results of the two administrations are correlated, yielding an indication of the test's scores stability (Etsey, 2012). The Pearson moment and Spearman correlation coefficient are the suggested statistical procedures for estimating test-retest reliability. These procedures are applicable when the scores are continuous and ordinal respectively.

Equivalent forms method

Different from the test-retest method, the equivalent-form is used to estimate reliability by giving two forms (with equal content, means, and variances) of a test to the same group either on the same day or a later day and correlating the results (Brennan, 2006). With this method, one determines how confident an examinee scores could be generalized to what the examinee would receive if the examinee took a test made up of similar but different items. In this case, it is the changes due to the specificity of knowledge that is measured and not changes from one time to another. The Pearson moment and Spearman correlation

coefficient are the suggested statistical procedures for estimating alternate form reliability. These procedures are applicable when the scores are continuous and ordinal respectively.

Inter-rater

When more than one observer captures the behaviour of respondents at the same time using the same instrument, inter-rater reliability is a notion that looks at whether scores from one sample are consistent. (Creswell, 2002). In statistics, inter-rater reliability (also called by various similar names, such as inter-rater agreement, inter-rater concordance, inter-observer reliability, inter-coder reliability, and so on) is the degree of agreement among independent observers who rate, code, or assess the same phenomenon. Assessment tools that rely on ratings must exhibit good inter-rater reliability, otherwise they are not valid tests.

There are a number of statistical procedures that can be used to determine inter-rater reliability. Different statistics procedures are appropriate for different types of measurement (Gwet, 2014). Some options are joint-probability of agreement, such as Cohen's kappa, Scott's pi and Fleiss' kappa; or inter-rater correlation, concordance correlation coefficient, intra-class correlation, and Krippendorff's alpha (Gwet, 2014). Gwet further stated that the joint-probability methods are used for nominal data. The inter-class correlation works on item response theory. The Cohen's kappa also works on categorical variable. The score in this study is continuous by nature. In this study, the Pearson Product

Moment correlation was used to estimate item and test level inter-rater reliability. This is because, it allows for interval and ratio data.

Standard Error of Measurement (SEM)

Standard error of measurement is the standard deviation of error of measurement in a test or experiment. It is closely related to the error variance, which represents the degree of variability produced by measurement error in a test given to a group. Standard error of measurement is defined by AERA, APA, and NCME (2014) as the standard deviation of measurement errors associated with test scores for a specific group of test-takers. The standard error of measurement is used to figure out how measurement error affects individual test findings.

The standard error of measurement is determined by the standard deviation of observed scores as well as the test's dependability. The standard error of measurement equals 0 when the test is completely dependable. The standard error of measurement is equal to the standard deviation of the observed results when the test is fully inaccurate. The unit of measurement for the standard error of measurement is the original unit of measurement. Mathematically given:

$$SEM = SD_x \sqrt{1 - r_{xx}}$$

The standard deviation and dependability are used to calculate the standard error of measurement. In addition to the dependability coefficient, the standard error of measurement plays a role. Although the reliability coefficient is useful for determining the amount of error in a test when applied to a group or population, it does not reveal the amount of inaccuracy in a single test score. The

standard error of measurement is frequently calculated using the Pearson product-moment coefficient metric of dependability.

(Ramsenthaler, et al., n.d.) posited that reliability as a characteristic of a good assessment has to do with consistency of assessment results. The consistency of assessment results qualifies to be accepted as reliability if and only if the task is the same or an equivalent administered either at the time or different time. Therefore, PBA becomes good assessment if and only if students' performance on a PBA is consistent when the same task or an equivalent one is administered to the student on the same or at different times. This is the essence of the use of inter-rater reliability.

Validity

Validity is the bedding rock of all assessment theories and principles, in that, it underpins all assessment theories. It is the focus or the object of concern of every assessment process. The principal aim of validity is to ensure that outcomes of assessment are given their genuine use and interpretation.

According to Nitko (1996, p. 56), "validity is the soundness of the interpretation and use of assessment results". Messick (cited in Reid, 2014., p. 79) also defined "validity as an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment and not the assessment instrument itself". Messick meant that for every interpretation and use made of assessment scores, there should be evidence to support the appropriateness of interpretation and use made and that assessment

instrument cannot be ascribed as valid or invalid. Nitko (1996) also pointed that validity is the appropriateness of the use and interpretation of students' assessment results. That is evidence need to be provided in support of the interpretation and use of the results of the students. The AERA, APA and NCME (1999) stated that validity is the degree to which the interpretations and use of test scores could be supported by evidence and theory. This view is similar to the discussed views of validity. The underlying theme of validity is evidence to support any interpretation and use of assessment results.

The soundness or evidence in support of the interpretation and use of assessment score suggest that, the evidence may adequately or partially support the interpretations and uses of the assessment result. Messick (2001) stated that validity is a matter of degree and not all or none. This means that the evidence may support a particular interpretation or use but not all or none of the uses and interpretations of the assessment results and not the assessment instrument itself as stated earlier.

Gyamfi (2017a) posited that assessment results can be used for instructional purpose; to find out if students have mastered a particular concept or not. It can also be used for certification, placement and selection as well as guidance and counselling. This means that the assessment results could be used as the bases for any of these decisions.

Nitko (2004) stated with regard to interpretation of assessment results, there are two ways; norm-referenced and criterion referenced interpretations. Norm – referencing is interpreting students results based on the norm or the group

within which the students' results lie. In this case, there is nothing like standards. Examples are selection of students for award or position of students which all depend on the students' results within the group. Criterion referenced interpretation is interpreting students' results based on standards. It judges if a students' meets a standard or not. This is used for certification, placement and programme evaluation.

Validity therefore is the appropriateness of the use of students' assessment results for certification, placement or selection (Hamavandy & Kiany, 2014). Can the results be used for the proposed use? Can the results be interpreted using norm-referenced or criterion referenced approach? Is it appropriate to use or interpret assessment results as proposed? These are the questions that come to mind with regard to validity. Assessment results that are considered valid for a particular use may not necessary be a valid for another use. The degree, to which it is appropriate for these interpretations and uses to be made of the assessment results, is what is termed as validity.

According to (Drost, n.d), valid results are not bias. This is because, bias items do not produce results that are good for comparison or predictions or measuring students true standing on a construct. Also, that, it does not discriminate between students matched to the same ability level but to those of discriminate ability level. Also, the items should be of equal difficulty to students matched to the same ability level. Estimation of these parameters lies in the domain of the item response theory.

Performance-based assessment needs to satisfy the conditions of validity before it could be used. Content-related, construct-related and criterion related evidences have to be provided to support the use and interpretation of PBA to make it more valid for consideration for use in schools especially in a high stake large-scale examination like the West African Senior Secondary Certificate Examination.

Validity Theory

Due to the importance of validity and it, being central to assessment, many studies have been done and there are still on-going studies of validity. This is to ascertain the best evidences to support the interpretation and use of assessment results. Messick (cited in Ankomah, 2020) stated that new findings have changed the phases of validity over time for better understanding of the phenomena. This means that the concept validity keeps undergoing metamorphosis. It can therefore be stated that what is a valid result today, may not be a valid results tomorrow. Ankomah (2020) stated that validity is a phenomenon which keeps on changing and validation is a continuing process. This is because, evidence is always not complete, and it is essential to make the most current reasonable use of the assessment results which is guaranteed in advanced research. Theoretically and gradually, the concept of validity has changed over the years (Anastasi, cited in Chalhoub-Deville, 2016; Ankomah, 2020).

One or another of these forms of evidence, or combination of them gave birth to the status of types of validity in the past (Messick cited in Ankomah, 2020). Scholars based on the sources of evidence as considered as types of

validity. However, because all the sources of evidence depend on the valid interpretation and use of assessment scores, there cannot be types of validity. According to Nitko (2004), validity is a unitary concept. This means that it has been established that there cannot be types of validity. All the evidences support that unitary concept, validity.

One major evident of validity that was neglected in early views of validity is the consequential use and interpretation of assessment scores. Chalhoub-Deville (2016) noted consequential basis of assessment validity has received little attention since the 1950s because validity has been conceptualized in terms of the functional worthiness of the assessment, that is, in terms of how well the assessment does the proposed purpose. Guilford (1946) claimed that an assessment result is valid for anything with which it correlates. Recent studies have underscored the continuing need for validation practice to address the realities of potential and actual assessment consequences on society. Emphasis is being placed on social values implied by the interpretation and use of the assessment results. The social consequences of assessment results are also seen to be subsumed as aspect of construct validity.

The 1954 technical recommendations (AERA, APA, and NCME cited in Chalhoub-Deville, 2016) listed three types of validity-namely, content, predictive, and construct validities. However, the AERA, APA, and NCME (cited in Sireci, 2013) reduced the types to three, namely, content, criterion –related and construct validities. These validity types were based on a particular aim of assessment. These aims include 1) determining how an individual is currently performing in a

collection of content, 2) forecasting an individual's future standing or to estimating the individuals present standing on some important trait other than the assessment, and 3) inferring the degree to which an individual possesses some construct acclaimed to be reflected in performance of the assessment task (Royal, 2017). The American Psychological Association (cited in Sireci, 2013) further pointed out that the three types of validity are by concept, independent, and seldom, one is important than other in a particular situation. All the types of validity are needed for a thorough study of assessment. The study is incomplete without the others.

Further clarification on the concept of validity was detailed in the AERA, APA, and NCME (cited in Ankomah, 2020). Behaviour was replaced with content. Content validity was described as how well the behaviours demonstrated in assessment constitute a representative sample of domain of behaviours. The shift from content to behaviour means content validity cannot be evaluated by a mere professional judgement of content relevance and representativeness. Thus, content validity requires evidence of reliable response which are consistent on the assessment and that the assessment, and the domain of assessment are similar or from same response (Messick, cited in Sireci, 2013). This has placed the evaluation of content validity beyond mere professional judgement.

The 1985 standards AERA, APA, and NCME (cited in Ankomah, 2020) also showed more light on the conceptualization of validity. The standards stressed on the unitary nature of validity, referring to the appropriateness, usefulness and meaningfulness of the specific inferences made from the

assessment scores. This notion nullifies the notion of “types of validity” to “categories of validity evidences” as content-related, criterion-related and construct –related evidence of validity. Evidence from the related areas should be provided to support the interpretation, use and social consequences of the assessment results before it is deemed valid.

Works of Anastasi and Cronbach (as cited in Royal, 2017) portray some evolution of validity. Anastasi, in his work in 1954 organised validity in terms of face validity, content validity, factorial validity, and empirical validity. Face validity has been phased out in recent validity analysis because face validity refers to what an assessment appears to measure to the layperson. Validity has come to be understood as not about the assessment itself but the results. Empirical validity has been established to an aspect of construct validity and therefore no more in operation. Empirical validity is about the procedures used to check content validity, which, construct validity measures by evaluating how well the content measures the behaviour. Factorial validity also in the work of Anastasi has been phased out. Factorial validity refers to the correlation between the assessment scores and a factor common to a group of assessment or other measures of behaviour. Contemporary construct validity is established by finding the correlation of the assessment results with other measures (Amedahe, 2000). This suggests that the Anastasi’s factorial validity is an aspect of contemporary construct-related evidence of validity.

In the work of Cronbach in 1949, Cronbach, organised his work on validity in terms of logical validity and empirical validity, as in the work of

Anastasi. Cronbach's logical validity was based on judgement of precisely what the assessment results measures. It was evaluated by making a careful study of the assessment itself. On the bases that validity is about the assessment result and not the assessment itself, this logical validity of Cronbach has been phased out. The empirical validity of Cronbach has been phased out on the justification for the phasing out of Anastasi's empirical validity.

The works of Mehrens and Lehmann (cited in Ankomah, 2020), Plake, Impara and Buckendahl (2004) and Smisko, Twing and Denny (2000) on validation process gave birth to a so-called type of validity known as curricular validity. Curricular validity is evaluated by comparing the assessment instrument to the curriculum that was dictated for the assessment. The so-called curricular validity has been phased out from contemporary validation process on the bases that curriculum is reflected in the content of the assessment. Therefore, curriculum validity perfectly subsumes under content validity.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014) again highlighted some concerns with regard to validity of assessment results 1) construct underrepresentation or constructs deficiency and 2) construct irrelevant variance or constructs contamination. "Construct underrepresentation refers to the degree to which assessment results fails to capture the important aspect of the construct" (AERA, APA, and NCME, 2014, p. 27). Construct validity is said to be underrepresented when there is no evidence to support a complete representation of essential constructs to be measured by the assessment

instrument. This results in a narrow meaning ascribed to the assessment results. Construct –irrelevant variance also refers to the degree to which the assessment results are affected by extraneous variable. Constructs are possibly influenced by factors that are not intended, for example measuring students' mathematical ability may be influenced by vocabulary, or anxiety. The ability of the construct to be devoid of such extraneous variables ensures construct validity. According to AERA, APA, and NCME (1999, p. 19) “sources of validity evidence are classified under content, response processes, internal structure, consequences of assessment and relation to other variables.” Each source perfectly comes under one of the three related evidences of validity.

Other concepts that have emerged in validity analysis are convergent and discriminant evidence of validity. The “convergent and discriminant validity” have not been popular in previous studies (Bollen, 2011). For convergent evidence measures that in reality correlates should perfectly correlate in evaluating construct validity and for discriminant validity, those that in reality do not correlate should not in any way correlate (AERA, APA & NCME, 1999). The AERA, APA, and NCME (2014) however separated all the sources of validity evidence under three main themes, 1) “establishing intended uses and interpretations, 2) issues regarding samples and 3) settings used in validation and specific forms of validity” (p. 21). All the sources of validity suggested in AERA, APA, and NCME (1999) are clustered under specific forms of validity in AERA, APA, and NCME (2014).

Contemporary view of validity

Validity refers to the appropriateness of the interpretation of the results. It does not refer to the instrument. With regard to use and interpretation, references are made to the results (scores) of the assessment. Decisions are made based on the results of the assessment. No information is obtained from the face of the assessment instrument say class test, however, the scores on the class test are used for decision or are interpreted. An assessment instrument cannot be said to be valid because the assessment instrument can be used for any decision or interpreted in any form.

Validity is a matter of degree. Assessment result is valid for different purposes and situation and ranges from low to high. The soundness or evidence in support of the interpretation and use of assessment score suggest that, the evidence may adequately or partially support the use and interpretation of the assessment scores. Chalhoub-Deville (2016) stated that validity is a matter of degree and not all or none. This means that the evidence may support a particular interpretation or use but not all or none of the uses and interpretations of the assessment results and not the assessment instrument itself as stated earlier.

Validity is for a specific interpretation or use. Results cannot be valid for all purposes. There are diverse uses and interpretations of a single assessment results. It is therefore, difficult (almost impossible) to put a single result to all the available interpretations and uses. The results will be useful in one instance but not the other or well interpreted in one way and not the other.

Validity is a unitary concept based on various kinds of evidence that support the concept. One or another of these forms of evidence, or combination of them gave birth to the status of types of validity in the past (Erzoah, Gyamfi, Yeboah & Langee, 2022). Scholars based on the sources of evidence as considered as types of validity. However, because all these sources of evidence bear on the valid interpretation and use of assessment scores, there cannot be types of validity. According to Nitko (2004), validity is a unitary concept. This means that it has been established that there cannot be types of validity. All the evidences support that unitary concept, validity.

Validity involves an overall judgement. Several types of evidence have to be studied and combined. An assessment result can only be concluded as valid when evidences are checked from different sources. No single source provides enough information to conclude on the validity of the assessment results. It is a comprehensive judgement about the assessment results ranging from authenticity through content representation to biasness.

Principles of validation

Interpretation (meaning) are valid to the degree evidences can be produced to support their appropriateness. Nitko (2004) stated with regards to interpretation of assessment results, there are two ways; norm-referenced and criterion-referenced interpretation. Norm-referencing is interpreting students results based on the norm or the group within which the students' results lies. In this case, there is nothing like standards. Examples are selection of students for award or position of students which all depend on the students' results within the group. Criterion-

referenced interpretation is interpreting students' results based on standards. It judges if student meets a standard or not. This is used for certification, placement, and programme evaluation. Any of interpretations made to an assessment result can be proven sound or appropriate.

Uses are valid to the degree evidences can be produced to support their appropriateness. Asamoah-Gyimah and Anane (2018) and Gyamfi (2017b) posited that assessment results can be used for instructional purpose; to find out if students have mastered a particular concept or not. It can also be used for certification, placement and selection as well as guidance and counselling. This means that the assessment results could be used as the bases for any of these decisions. A decision to put an assessment result to any of these uses should be proven appropriate.

Interpretations and uses are valid when educational and social values implied are appropriate. The interpretation and use made of assessment results arise from educational and social values. The interpretation made of the results is implied by the educational and social values. That is the interpretations and uses should be deemed appropriate from the educational and social lens.

Interpretations and uses are valid when consequences of these of these interpretations and uses are consistent with appropriate values. Every action has its corresponding consequences and so are the interpretation and use of assessment result. The intended and unintended consequences should be consistent. For example, if from the interpretation and use of an assessment results and students are placed in a remedial class with the intention of improving

the students learning and ends up that the student gets frustrated or the remedial does not improve the students leaving, then that consequence of the interpretation and use of the results has not been consistent and that the result's validity is hanging.

Categories of validity evidence

Messick (cited in Chalhoub-Deville, 2016) stated that new findings have the existing evidence of validity evidence. This means that the concept of validity keeps undergoing metamorphosis. It can therefore, be stated that what is a valid result today, may not be a valid result tomorrow. Messick (cited in Sireci, 2013) again stated that validity is an always-changing property and validation is a continuing process. This is because, one source of evidence is always not complete and it is essential to make the most current reasonable use of the assessment results which is guaranteed in advanced research. The theoretical conception of validity has gradually changed over the years (Ankomah, 2020; Nitko, 2014).

According to Messick (cited in Sireci, 2013), “since the early 1950s, validity has been broken into three or four different types. Specifically, validity has been divided into three types, of which one comprises two subtypes” (p. 232). These are content validity, criterion-related validity comprising predictive and concurrent validity, and construct validity. These are what AERA, APA, and NCME (cited in Garrison, Chandler & Ehringhaus, 2020) ascribed as traditional validity types. Research has proven that these perceived types are rather sources of evidence that support the unitary concept, validity.

Content related evidence

This evidence is about the content representativeness and relevance of the assessment results. Content validity is defined as “the degree to which items on an instrument reflect the content universe to which the instrument will be generalized” (Chan & Malim, 2017). Content-related evidence of validity is assessed by showing the degree to which the content of assessment results represents the content about which conclusions are to be drawn. The judgement on content relevance focuses on whether tasks included in the assessment are in the test domain definition. The relevance of the assessment results is the extent to which the assessment matches the school’s curriculum target (Azwar, 2012; Retnawati, 2017). There should be an overlap between the assessment domain and the curriculum. The weight given to each content area should be appropriate to the local curriculum (Nitko, 2004). According to Nitko, to ensure content validity, the items should have the following characteristics: (1) reflect current thinking of the subject matter of what is essential to teach and assess (2) accurately represent the subject matter (3) keyed correctly and (4) contain meaningful and relevant content.

To judge whether as assessment the content has related evidence to support the interpretation and uses of the assessment results, table of specification is prepared and used (Nitko, 2004). The table of specification is a means of defining the domain for standardized position on achievement test. It contains the major content areas and skills to be assessed and the percentage of tasks content-

skills. In recent times, statistical procedures have been developed to estimate the content validity of an instrument.

It is highly recommended to apply content validity while the new instrument is developed. In this study, content validity was applied by expert indicating relevant/not relevant to each of the items. In general, content validity involves evaluation of a new survey instrument in order to ensure that it includes all the items that are essential and eliminates undesirable items to a particular construct domain (Chan & Malim, 2017). The judgmental approach to establish content validity involves literature reviews and then follow-ups with the evaluation by expert judges or panels. The procedure of judgemental approach of content validity requires researchers to be present with experts in order to facilitate validation. When it is not possible to have many experts of a particular research topic at one location, a quantitative approach may allow researchers to send content validity questionnaires to experts working at different locations (Choudrie & Dwivedi, 2005). In order to apply content validity, the following steps are followed:

1. An exhaustive literature reviews to extract the related items.
2. A content validity survey is generated (each item is assessed using three-point scale (not necessary, useful but not essential and essential)).
3. The survey administered to the experts in the same field of the research.
4. A suitable approach is selected to analysis
5. Decision on status of items

Several statistical tools have been developed for content validity (Choudrie & Dwivedi, 2005). Popular among them are the Lawshe (1975) method content validity ratio (CVR) and modified kappa statistic (K).

Lawshe (1975) method

According to Davis (cited in Polit, Beck & Owen, 2017; Nugroho & Tomoliyus, 2019), the CVR (content validity ratio) introduced by Lawshe is a linear transformation of a proportional level of agreement on how many "experts" within a panel evaluate an item "important." Mathematically, it is expressed as:

$$CVR = \frac{N_e - \frac{n}{2}}{\frac{n}{2}}$$

where CVR stands for content validity ratio, N_e is for the number of panel members who indicated "essential," and N stands for the total number of panel members. Tomoliyus, Sumaryanti and Jadmika (2016) suggested that CVR of 0.500 is acceptable level of content validity. The number of panels determines the final decision to keep the item based on the CVR) Table 3 shows the guideline for the valid value of CVR for the evaluated item to be retained suggested by Lawshe (1975).

Table 3- *Minimum value of CVR*

No. of Panellists	Minimum Value
5	0.99
6	0.99
7	0.99
8	0.75
9	0.78
10	0.62
11	0.59
12	0.56
13	0.54
14	0.51
15	0.49
20	0.42
25	0.37
30	0.33
35	0.31
40	0.29

Source: Source: (Lawshe, 1975)

Table 3 shows the acceptable CVR per the number of experts evaluating the instrument. The table indicates that the acceptable CVR ranges from 0.29 to 0.99 for 40 to 5 experts respectively as suggested by Lawshe (1975) There is an inverse relationship between the number of experts and the acceptable level of CVR. If a small number of experts are involved, a higher CVR is expected. In this study, 250 experts were involved thus lesser CVR is still acceptable to judge the content validity of the instrument.

Modified kappa statistic (K) method

To obtain content validity index for relevancy and clarity of each item (I-CVIs), the number of those judging the item as relevant or clear was divided by the number of content experts but for relevancy, content validity index can be calculated both for item level (I-CVIs) and the scale-level (S-CVI) (Polit, Beck &

Owen, 2017). In item level, I-CVI is computed as the number of experts giving a relevant to the relevancy of each item, divided by the total number of experts. The universal agreement among experts (S-CVI/UA) or averages the item-level CVIs (S-CVI/Ave) could be used (Polit, Beck & Owen, 2017). The S-CVI/UA estimates reliability for each item whiles S-CVI/Ave estimate reliability for the entire test.

Wynd, Schmidt and Scheafer (2013) proposed both content validity index and multi-rater kappa statistic in content validity study because, unlike the CVI, it adjusts for chance agreement. The following steps were suggested for using the modified kappa statistic:

1. estimate the item content validity index using the following

$$I - CVR = \frac{A}{N}$$

In this formula, N= number of experts in a panel and A= number of expert who agree that the item is relevant

2. the probability of chance agreement was first calculated for each item by following formula:

$$P_c = \left[\frac{N!}{A!} (N - A)! \right] \times 0.5^N$$

3. kappa was computed is using the formula

$$K = \frac{(I - CVR) - P_c}{1 - P_c}$$

Evaluation criteria for kappa is the values above 0.74, between 0.60 and 0.74, and the ones between 0.40 and 0.59 are considered as excellent, good, and fair, respectively (Cicchetti & Sparrow, cited in Wynd, Schmidt & Scheafer,

2013; Tomoliyus, Sumaryanti & Jadmika, 2016). In this study, the modified kappa statistics was used for the evaluation of the content validity.

Criterion related evidence

Criterion-related evidence of validity measures how well an assessment results can predict a future performance on similar content. It is established by comparing the assessment results with scores of one or more external variables (called criteria) which is considered to provide a direct measure of the trait of interest. There are two types of criterion related evidence of validity. The predictive validity of the criterion validity gives an indication of the extent to which an individual's future performance on the criterion is predicated from a previous performance. The purpose is to predict the future performance of a criterion variable. The concurrent validity also gives an indication of the degree to which the assessment results estimate individuals' present standing on the criterion. The purpose is to substitute the assessment results for the score of a related variable. The line of difference between the predictive and the concurrent validity then becomes the time of measure of the future (criterion) and present (predictor) standing on the criterion.

To assess this evidence, the correlation coefficient of the criteria and predictor is estimated. The coefficient gives an indication as whether there is a relationship between the scores and how well the predictor predicts or relate with the criteria. Another approach to check predictive validity is by the use of the expectancy table. It is a two-way table that allows criteria to be predicted from a score.

Criterion validity is not related to this study. This is because the instrument was administered to the student once. Two sets of scores are required to be able to estimate at least the concurrent validity of the instrument (Nitko, 2014).

Construct related evidence

The construct validity is established by studying what qualities the assessment measures, that is, finding the degree to which the constructs account for the performance on the assessment. Taherdoost (2016) referred to construct validity as how well a concept, idea, or behaviour that is a construct is translated or transformed into a functioning and operating reality. Asamoah-Gyimah and Anane (2018) stated that this evidence refers to how the assessment results can be interpreted as reflection of an individual's achievement on what is being measured. Convergent and discriminant validity are two aspects of construct validity.

Discriminant Validity

The degree to which a latent variable discriminates from other latent variables is known as discriminant validity (Fornell & Larcker cited in Taherdoost, 2016). Discriminant validity refers to a latent variable's ability to explain for more variance in the observed variables than a) measurement error or other unmeasured external influences; or b) other constructs within the conceptual framework. If this isn't the case, the individual indicators' and the construct's validity is in doubt (Fornell & Larcker, cited in Wynd, et al, 2013). In a nutshell,

discriminant validity (also known as divergent validity) verifies that constructs that should not be related are, in fact, unrelated.

Convergent Validity

Convergent validity is a term used in sociology, psychology, and other behavioural sciences to describe the degree to which two measures of constructs that should be connected theoretically are really related (Wynd, Schmidt & Scheafer, 2013; Taherdoost, 2016). Convergent validity, in a nutshell, ensures that structures that should be connected are indeed related and those which should not be related, in reality, are not related.

Loevinger (as cited in Taherdoost, 2016) pointed out that, for all the so-called validity types, construct validity is the whole and all the others are ad hoc. Thus, construct validity encompasses almost all the forms of validity evidences. Content validity, “professional judgement is the bases of the “so-called relevance and representativeness of the assessment content” of a particular domain of interest. So, the so-called content validity addresses the assessment instrument representativeness and not the scores. In that sense, the so-called validity cannot be validity.

Methods of estimating construct validity

Principal component analysis: for discriminant and convergent validity, a factor analysis can be conducted utilizing principal component analysis (PCA) with varimax rotation method (Koh & Nam, 2005; Wee & Quazi, 2005). Items loaded above 0.40, which is the minimum recommended value in research are considered for further analysis. Also, items cross loading above 0.40 should be deleted.

Therefore, the factor analysis results satisfy the criteria of construct validity including both the discriminant validity (loading of at least 0.40, no cross-loading of items above 0.40) and convergent validity (eigenvalues of 1, loading of at least 0.40, items that load on posited constructs) (Straub, Boudreau & Gefen, 2004; Strube, 2002). There are also other methods to test the convergent and discriminant validity.

In the essay format of mathematics items, knowledge from different concepts is pulled together to respond to an item. In a single item, various constructs such as reading ability and mathematical skills are consolidated. The various constructs in the item are not separated. Thus, the observed score reflects the student's mathematical knowledge. Mathematics essay items are thus measure for unidimensionality rather than multidimensionality. In this sense the items are checked for convergent validity and not divergent validity. The items must measure a single construct. The Principal Component Analysis is used to estimate the unidimensionality using the eigenvalue of 1 for decision making on the items.

Factors that affect validity

Asamoah-Gyimah and Anane (2018) listed the following as the factors that affect validity of assessment result:

1. Unclear directions
2. Too difficult vocabulary
3. Test being too short
4. Improper arrangement of items
5. Cheating

6. Unreliable scoring
7. Group characteristics

Nitko (2014) and Asamoah-Gyimah and Anane (2018) stated that too short test has questionable validity. The test should cover a representative sample of content studied. Again, the test should represent relevance. It is inaccurate to generalize by use or interpretation of a students' score on limited sample or irrelevant content. It is therefore important to note that content validity is a key dimension in the validity of the assessment result. Again, Nitko (2014) cited unreliable scoring as a factor that affects the validity of assessment results. Inconsistency in the scores of a student performance on the same or similar construct distorts the interpretations and uses of students scores as it is difficult to speak on students' true scores for decision making. Reliability parameter, thus, is another important domain to be considered in validating assessment instruments for classroom use.

Test Theories and Test Development

Test development is the “process of producing a measure of some aspect of an individual's knowledge, skill, ability, interests, attitudes, or other characteristics by developing items and combining them to form a test, according to a specified plan” (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014, p. 75). This means test development encompasses before and after test construction.

Testing has become the popular form of assessment in schools and so the term test is also popular in school environment. According to Nitko (2004) and Asamoah-Gyimah and Anane (2018), achievement testing are the testing processes and activities that measure the degree of present knowledge and skills. By inference, achievement testing seeks to measure students' current standing on learned knowledge and skills. That is, it is about finding how much knowledge and skills have students achieved on learned concepts. Standardized and teacher made testing are the two forms of achievement testing. Whiles standardized testing proud itself with uniformity of the testing processes and use of expert in every stage of the testing process, teacher made testing lacks the uniformity and expert involvement in the testing (Nitko, 2004). Achievement testing is broadly for providing information for public accountability and instructional decision (Brennan, 2006). For accountability purposes, the testing is done to provide information to authorities on the performances of teachers. The instructional decision purposes of achievement testing, focuses on finding the weakness and strength of students' achievement of concepts and skills learned (Gyamfi, 2022b). The prime goal of both accountability and instructional decision purposes of achievement testing is to improve student achievement. According to Hambleton and Jones (as cited in Nitko, 2012), achievement testing goes through the following stages:

1. Determining the purpose of the test
2. Preparation of test specifications
3. Preparation of the test item pool

4. Field testing of the items
5. Revision of the test items
6. Test development
7. Pilot testing
8. Final test development
9. Test administration
10. Technical analyses
11. Preparation of administrative instruction
12. Printing and distribution of the test and manual.

It could be seen that, there are variations in the stages depending on the test theory and the type of achievement test being used. Teacher made testing and the classical test theory do not follow vividly all the listed stages of testing. For instance, there is no pilot testing in teacher made test and the technical analyses of the test differ from one test theory to the other.

Achievement testing lends itself to several testing theories. Among the test theories are classical test theory, item response theory and generalizability theory. All these theories in relation to achievement test focus on test content and test characteristics. The test theories are applied at the stages of test development and technical analysis of achievement testing and so the variations in roles of the test theories in achievement testing are found at the stages where they are used. The test development covers issues such as construction of items, pooling of items to form a test and field testing and ability estimation for appropriate items to be administered. The technical analysis covers analysis of item difficulty, item

discrimination and test reliability. Some of the theories are limited to some particular technical analyses. For instance, generalizability is limited to test reliability (dependability).

Test development involves pooling items from a pool that consists of more questions than are needed to populate the test forms to be built. This allows items that meet particular specifications to be selected. Establishment of these specifications is done through item review and item try-out. Items are reviewed for content quality, clarity, and construct-irrelevant variances of content that influences an examinee response. The item try-out help determine some of the psychometric properties of the items. These psychometric properties are item difficulty and item discriminations (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014). The psychometric properties are determined by applying test theories such as Classical test theory, items response theory, and generalizability theory to the test scores (Royal & Gonzalez, 2016).

Development and Validation of Instrument

There are quite a number of procedures that have been suggested for development and validation of instrument. Among the procedures are the one proposed by Benson and Clark (as cited in Pineda, 2012), Newman and McNeil's (as cited in El-sehrawy, 2020) and Onwuegbuzie, Bustamante and Nelson's (2010). The Benson and Clark (1982) is a not as current as the others. However, the newer approaches took inspiration from the former.

Benson and Clark in 1982 proposed a four-phase process. This four-phase model was somewhat tweaked to serve as the conceptual foundation for this study's instrument development and validation. (a) planning, (b) construction, (c) quantitative evaluation, and (d) validation are the four steps (Pineda, 2012).

Phase one: Planning

Phase one of instrument creation, according to The American Association for Public Opinion Research (2016, p. 56), is the planning phase, “which is considered the most significant stage in the development.” This phase is used to fully flesh out the study questions, aims, and objectives. This phase also includes a study of previous research as well as an assessment of similar or related instruments. A thorough comprehension of the constructs being measured, as well as a clear understanding of the instrument's purpose and priority for possible future usage, considerably boosts the chances of a successful final form (Dillman, Smyth & Christian, 2014).

The planning step is also an excellent time to identify the target population and determine the sample frame. In order to reduce and quantify sampling error and bias, it is necessary to have a well-defined target population and a sample frame that closely resembles the population (Dillman, et al, 2014; Fowler, 2008). “Identifying the underlying assumptions, both in construct and methodology, including the process of conducting a survey at all, during the planning phase assures a successful instrument development.” Apart from surveying, a number of different methods for gathering data “that may provide more accurate or comprehensive data are available” (Draugalis, Coons & Plaza, 2008, p. 66).

The instrument's success depends on a well-defined aim, plan, and technique. Researchers that skip the planning stage have poor results (Gable & Wolf, 2012). In many of these circumstances, the survey's methodology or items fail to assess the targeted construct. The researcher's evaluation is bases of decision making throughout the instrument creation process; hence, a thorough grasp of the constructs being tested is necessary to avoid bias (Dillman et al., 2014). As a result, item selection or development is harmed by a lack of knowledge with the literature or the lack of established contextual frameworks (Kelley, Clark, Brown & Sitzia, 2003).

Phase two: Construction

The second phase entails the creation and examination of a vast item pool. Traditionally, “a test or instrument developer will conceive one or more domains based on the understanding of the constructions and try to come up with items or behaviours that represent or exhibit the construct in question.” However, this methodology introduces a layer of subjectivity to the instrument in creation, as well as the possibility of omission of relevant domains and an unquantifiable bias (Crocker & Algina, cited in Dillman et al., 2014). As a result, the goal of Phase Two is to develop a more methodical approach to item development in order to limit potential researcher bias.

It is impossible to completely eliminate this bias, according to Pineda (2012). Instrument development is a delicate balance of art and science, as content experts' wisdom, experience, and subjectivity must be balanced with scientific and statistical approaches, which are, incidentally, also open to interpretation

(Schmeiser & Welch, 2006). The item pool contains more questions or tasks than are required to populate the instrument, as per the Standards for Educational and Psychological Testing (American Educational Research Association, 2014).

As previously stated, the researcher's experience is required for the construction of the item pool. Crocker and Algina (as cited in Pineda 2012, p. 92) recommended “engaging in one or more of the following actions to widen, refine, or validate the researcher's view of the construct: content analysis, assessment of the study, critical incidents, direct observations, expert judgment, and instruction objectives.” Each of these steps is described in the following list:

1. “Content analysis is a qualitative approach that entails asking open-ended questions about the concept of interest to participants in the target demographic. These responses are then organized into relevant groups and used to create new items.”
2. A review of the research comprises looking at how other researchers have viewed the construct in the past. Gable and Wolf (2012) remark on the importance of this exercise, saying, "A well-done literature review will be a rich source of content" (p. 33).
3. “Critical episodes are a compilation of tales or behaviours related to the construct that are produced by subjects or the researcher and are useful in identifying extremes on the construct's continuum.”
4. “The researcher's direct observations of the individuals or environment aid in the identification of prospective construct domains.”

5. Expert judgment is obtained when the researcher collects further information on the concept by interviewing people who have firsthand experience with it.

6. “Instruction objectives are created when a researcher delivers material to field experts and requests that objectives be derived from the information provided. This method is better suited to assessing skill or knowledge growth than a perception survey instrument” (Armah, 2018, p. 99).

After the pool of questions has been created, it is improved in Phase Three through a content validity assessment process and subsequently pruned using statistical approaches. The researcher must decide on the proper answer format and scale size before testing and refining the item pool in Phase Three. For example, an instrument designed to assess how effectively pupils understand and apply a concept in a real-world setting is usually concerned with locating individuals at various locations along the constructs' continuum; as a result, a subject-centered approach is suitable (Crocker & Algina cited in Pineda, 2012).

Phase three: Validation

“The overall evaluative judgment of how well experimental data and theoretical frameworks support the appropriateness of instrument results interpretations is known as validity.” Validity is a quality of the meaning of the test scores, not of the test or assessment itself (Messick as cited in Gable & Wolf, 2012). In the broad idea of validity, Messick (as cited in Gable & Wolf, 2012, p. 67) considered “not only the meaning of the test scores, but also the interpretation, usage, and potential repercussions (both intended and unforeseen)

of the instrument as evidence for or against validity.” However, both conceptually and practically, the relevance of interpretation and consequences in the research of validation is debatable (Kane, 2006).

When it comes to determining validity, ignoring purpose is akin to defining validity for a useless instrument. The current definition of validity, as stated in the 2014 edition of the Standards for Educational and Psychological Testing, encompasses both interpretations and uses, and serves as a good beginning point for validation (Sireci, 2015). While Benson and Clark (cited in Pineda, 2012) refer to this phase as validation, this is misleading because the entire validation process is interwoven in all four phases. “Validation appears in all phases of instrument development and validation process since the instrument is not validated independently of the purpose—for example, setting purpose in phase one is part of the validation process—validation appears in all phases of the validation process” (Kane, 2012, p. 18).

Validity was first devised as a “correlational statistic between the test result and subsequent performance of the criterion being assessed” (Nunnally & Bernstein, cited in Tabachnick & Fidell, 2013 p. 76). Concurrent correlational statistics, “in addition to predictive criterion correlations, were utilized as a measure of how accurate an instrument was relative to similar instruments as instruments became more extensively deployed” (Lissitz & Samuelson, 2007, p. 89).

The validation of an assessment instrument’s stage according to Benson and Clarke (as cited in Morrell & Carroll, 2010) covers content

representativeness, coherence, feasibility or practicability, credibility of the assessment results of the instrument, educational and catalytic effects of the assessment instrument. Validation of coherence, feasibility or practicability, credibility of the assessment results of the instrument, educational and catalytic effects of the assessment instrument are established by expert and practitioners in assessment and the subject area.

Content validity in educational assessment refers to “how well a test measures the content that was taught” (Morrell & Carroll, 2010, p. 98). Construct validity was introduced by Cronbach and Meehl in 1955 as a fourth type of validity, alongside predictive, contemporaneous, and content validity. When no specific criterion exists, “construct validity is defined as how successfully the assessment instrument aligned and measured the domains and nomological networks of the targeted construct” (Lissitz & Samuelson, 2007)

Feasibility establishes the practicality of usage of the instrument for the intended group. Materials, time and cost of usage of the instrument are the bases for the feasibility validation (Brennan, 2000). Credibility of the assessment instrument is established by validating the trustworthiness of the assessment results. It answers questions like “how close is the observed score to the true score, can the assessment results of the instrument be trusted, does the instrument reduce malpractices”.

The validation of educational effect of an assessment instrument establishes the impact of the assessment instrument on students’ learning. It measures the ability of the instrument to motivate students to learn. The

assessment procedure should encourage students' participation and learning. The validation of the catalytic effect of an assessment instrument establishes whether the instrument could provide immediate feedback to students' learning (Newton, 2014). Classroom assessment must serve the need of the learner (Nitko, 2002). Therefore, constant and immediate feedback of students' learning should be provided to students. Any good assessment procedure should be able to give immediate feedback to stimulate students' learning.

Phase four: Quantitative evaluation

Phase four entails administering the item pool to a large representative sample in a first pilot, followed by item and factor analyses to refine item selection and inform construct domains. Fitting data to a common model is done in a number of ways. The assumptions and the methodology of these processes, such as Maximum Likelihood, Principle Component Analysis (PCA), and Principal Axis Factoring (PAF), differ slightly (Bichi, Embong, Mamat, & Maiwada, 2015).

The quantitative evaluation in the model of Benson and Clarke (as cited in Morrell & Carroll, 2010) focuses on the estimation of the item parameters. These include the reliability, difficulty and discrimination indices of the instrument as well as the presence of DIF. Different theories with their varying statistical procedures are available depending on the type of data the assessment instrument produces. The most popular theories that are applied in the parameter estimations are "the Classical Test Theory (CTT), Item Response Theory (IRT) and

Generalizability Theory (GT).” For dependability, generalizability theory is used to estimate the error component of each factor in the study.

For dichotomously scored items, the classical test theory could be used to estimate the reliability using either test-retest, equivalent method or split half. The difficulty index is estimated by calculating the proportion of examinees who correctly responded to each item. The discrimination index is estimated by calculating the differences in the proportions of higher and lower achievers who correctly responded to the item. The CTT assumes a constant ability level for all examinees hence do not consider DIF. The CTT is basically for dichotomously scored items. Item response theory considers the ability level of examinees in the estimation of the difficulty and discrimination indices of the items. It also estimates DIF. The IRT could estimate the item parameters for both dichotomously and polytomously scored items. The generalizability theory estimates the dependability (reliability) for both dichotomously and polytomously scored items. The G-theory does not estimate difficulty and discrimination indices and DIF.

Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), according to Truxillo (2003) and Williams, Brown and Onsmann (2012) are powerful statistical procedures for the construction of measurement instruments. According to Truxillo (2003), in order to apply EFA and CFA, the factor structure, or what is assumed of it, is identified by the blueprint. The EFA is necessary to uncover the underlying constructs for a group of measured variables. The CFA, on the other hand, allows for the testing of the similarities of

traits being measured by the items on the test. For CFA and EFA, sample sizes of 100-200 are adequate if the components are properly specified (Tabachnick & Fidell, 2013). Sample sizes of at least 300 are necessary when there are low commonalities (shared item variance), a small number of factors, or three to four indicators for each component. Sample sizes of considerably over 500 are required in the most unfavourable circumstances with low commonality and a large number of weakly determined components. If there are continuously high similarities, sample size can be lowered (Tabachnick & Fidell, 2013).

The Newman and McNeil (as cited in El-sehrawy, 2020) approach was written in a comprehensive question and answer format. A list of general guidelines in sequential steps was presented to follow in instrument development. This approach was easily adapted because it was seen simple but thorough. Newman and McNeil (as cited in El-sehrawy, 2020) did not categorically state that the approach is a mixed methodology, just like Benson and Clark. This is because an observation of the approach reveals that they could adopt either qualitative or quantitative methods. However, the quantitative dominates the approach. For example, as in the case of Benson and Clark (1982), “the initial stages highlighted the crucial role that stakeholders hold in planning for and creating the instrument, which was a qualitative-type of decision. Consecutive stages included quantitative instrument psychometric considerations”.

In contrast to the Newman and McNeil approach, Onwuegbuzie, Bustamante and Nelson’s (2010) approach was specifically designed for mixed methodology. Onwuegbuzie, Bustamante and Nelsons (2010) named their

approach to instrument development and validation as the Instrument Development and Construct Validation (IDCV) model. The approach sought to develop and validate quantitative instruments using mixed methods research. The IDCV is, “designed to help instrument developers undergo a rigorous and comprehensive process during instrument development/construct validation” (p. 60). “Different frameworks and models such as (a) Onwuegbuzie and Combs’ (2010) crossover analyses, (b) Green, Caracelli, and Graham’s (1989) rationale for combining qualitative and quantitative data, (c) Onwuegbuzie, Daniel, and Collins’ (2009) meta-validation framework, and (d) Onwuegbuzie, Leech, and Collins’ (2008) framework for debriefing the researcher were combined to develop and validate the IDCV”. The IDCV approach was made up of ten detailed stages and many procedures for estimating validity and reliability of the instrument. However, this approach was not different from the approach of Benson and Clark. The only difference was that Onwuegbuzie, Bustamante and Nelson (2010) elaborated the four-phases of Benson and Clark stage to get the ten stages.

It can therefore be stated that Benson and Clark approach to instrument development and validation is the mother of other latter approaches for development and validation of an assessment instrument. The researcher employed the Benson and Clark approach to developing and validating the instrument used in this study.

It can be deduced that the quantitative or mixed methodologies are the ideal approaches for the development and validation of instrument. It should be

stated that in each of the approach, the quantitative is superior to the qualitative which is buried in the quantitative. Reviewed literature on instrument development and validation, categorically or not lent itself to the mixed methodology. Since the quantitative dominates the qualitative, a quantitative approach could be adopted with the Benson and Clark approach.

Criteria for Evaluating Assessment Instrument

Evidence must back up the interpretations and show that the “authentic” assessments are technically sound. Skeptics will dispute the usefulness of new assessments and demand proof that they are worthwhile, both financially and in terms of effort (Wilkins, Norton & Boyce, 2013). But what kind of proof is required, and how should these alternatives to present standardized examinations be evaluated? Many concerns surrounding the evaluation of newly established kinds of assessment have yet to be adequately addressed (Kulas & Stachowski, 2009). Certain criteria for judging such assessments should be addressed in order to build technically valid performance assessments, portfolios, and simulations.

Of fact, psychometric criteria for determining the technical adequacy of measurements are widely established. Criteria derived born out of the fundamental ideas of reliability and validity are particularly important, but given the benefits of many new techniques to evaluation, expanding on their traditional conceptions seems reasonable (Oslin, Mitchell & Griffin, 2016). Validation, according to El-sehrawy (2020) entails the establishment of a consequential basis for interpretation of the test score and usage in addition to the more traditional evidential basis. If an assessment programme encourages teachers to devote more

time to concepts and information covered on the test and less time to content not covered on the test, repercussions must be considered when evaluating the validity of the assessment results' interpretations and applications (DeMars, 2018). Similarly, if an assessment results in extensive use of practice resources that closely resemble the format of the assessment, this is a consequence that must be considered.

Despite the importance of consequential basis of validity as expressed by theorists such as Messick and Cronbach (cited in Garrison, Chandler & Ehringhaus, 2020), “consequences were rarely listed among the major criteria by which the technical adequacy of an assessment was evaluated prior to the recent pleas for authentic assessment.” If performance-based evaluations are to achieve the potential that its proponents hope for, the consequential basis of validity must be given far more importance among the criteria used to judge assessments. The collecting of evidence about the intended and unintentional consequences of assessments on how teachers and students use their time and think about educational goals should be given top priority.

It can be assumed that a more "authentic assessment" will result in more "learning-friendly" classroom activities. For instance, in the area of mathematics assessment, a question like *“James knows that half of the students from his school are accepted at the public university nearby. Also, half are accepted at the local private college. James thinks that this adds up to 100%, so he will surely be accepted at one or the other institution. Explain why James may be wrong. If possible, use a diagram in your explanation.”* Items like the above are consistent

with mathematical conceptualizations like those articulated in Curriculum and Evaluation Standards for School Mathematics (National Council of Teachers of Mathematics, as cited in Gyamfi, 2017a). These “items conceptualize problem solving, the idea that there are many ways to solve problems rather than a single right answer or algorithm to memorize, and mathematics communication” (National Council of Teachers of Mathematics, 1995, pg. 21). When validity is considered in terms of consequences, focus is placed on parts of the evaluation process that the instrument designers may not have planned or anticipated.

The results of standardized tests can be tampered with. New modalities of assessment should not be presumed to be immune to such pressures. The consequences criteria is pertinent to the concepts of directness and transparency given by Johnson (2011). Because of the expected consequences on teaching and learning, directness and transparency are regarded to be crucial aspects of an assessment. It could be claimed, for example, that directness is vital since emphasis on indirect indicator measurements could cause training to be distorted. “This idea is illustrated by the comparison of multiple-choice questions concerning writing with direct writing samples” (Johnson, 2011, p. 68).

Similarly, transparency is valued because knowing how one's performance will be rated makes it easier to improve one's own performance. In other words, both directness and transparency are assumed to be ways to a better educational outcome (Nugroho & Tomoliyus, 2019). However, evidence is required that these apparent properties of an evaluation have the expected effects while avoiding unforeseen consequences.

Fairness. Any assessment should be judged on the basis of fairness. However, decisions about the fairness of an assessment are likely to be highly influenced by how the assessment results are interpreted and used. Due to differences in familiarity, exposure, and motivation on the tasks of interest, there are gaps in performance among groups (Nugroho & Tomolius, 2019). To adequately prepare students for complex, time-consuming, open-ended assessments, “significant changes in instructional strategy and resource allocation are required” (Hild, Gut & Brückmann, 2018, p. 4). It is critical to provide teachers with the necessary training and support to help them move in towards student centred assessment. However, teaching for success on these assessments is an identified challenge in and of itself, and it tests our understanding of teaching and learning. Because there is no ready-to-use technology to help with instruction, performance gaps may persist.

The “important point is unaffected by the degree of group differences. Regardless of the proportional extent of the performance gap between specific pairs of groups, problems of fairness will loom as big for PBA as they do for traditional examinations. Fairness issues arise not only in the assignment of performance assignments, but also in the scoring of responses” (Taut & Rakoczy, 2016, p. 48). As Schreiber, Theyßen and Schecker (2016) have stated, “it is critical that the scoring procedures are designed to assure that performance ratings reflect the examinee's true capabilities and are not a function of the perceptions and biases of the persons evaluating the performance” (p. 33). In this case, rater training and calibration are crucial. These procedures have “utility for flagging

items that may need to be eliminated or, at the very least, submitted for additional review before being used". The CTT item bias and DIF procedures are recommended.

Differential item functioning (DIF) is distinct from bias (Annan-Brew, 2020). As Dorans and Schmitt (as cited in Johnson, 2011) recently pointed out, DIF methods rely on the availability of performance on a large number of things that can be used as the matching criterion in grading each individual item. As a result, DIF techniques are unlikely to be directly relevant to performance evaluations with a small number of independent tasks. A technique devised by Welch and Hoover (as cited in National Association of Testing Authorities, 2012, p. 58) "that can be employed for DIF analysis of polychotomously scored items when an independent matching variable" is also available is one promising breakthrough in this field. DIF analysis may thus be applicable to assessments that combine performance-based measures and fixed-response items; nonetheless, it appears that a greater dependence on judging appraisals of performance tasks is unavoidable. For both performance assessments and traditional test questions, procedures for detecting materials that may be offensive to particular groups of students or that are the cause of irrelevant difficulty for a student must be used. Prior knowledge, for example, is proven to have a significant impact on pupils' capacity to grasp what they read (Schreiber, et al., 2016; Hild, et al., 2018).

An item is said to bias if different groups of examinees on the test have different probability of correctly responding to the item (Gyamfi, 2022b). It is an estimation of difference in the difficulty indices of item between two groups of

examinees on the test. If there is a difference in the proportion of students correctly responding to the item, then the item is said to be bias. An unbiased item has same proportion of examinees of identified groups correctly responding to the item. It uses the procedure for item difficulty in CTT. In the CTT, item bias is susceptible to multiple choice item which is dichotomously scored. Unlike DIF, item bias does not take into consideration, the ability level of the groups.

In standardized exams, one strategy to dealing with diversity of backgrounds and experiences is to sample a variety of topics, sometimes involving distinct cultural contexts. Because the time-consuming nature of the issues restricts their number, this strategy is more difficult to apply with performance-based examinations. Miller-Jones (as cited in Schreiber, et al., 2016) proposed an alternate technique including the use of "functionally similar" activities that are tailored to the culture and instructional context of the individual being evaluated. However, establishing task equivalence is "very challenging" (Taut & Rakoczy, 2016, p. 3). To say the least, developing differential tasks that are suited to the persons being assessed and can be utilized to produce fair, functionally equivalent performance assessments poses a significant challenge for those interested in developing them. In this study, PBA was not validated for item fairness. This is because, the items are graded response type which are not suitable for CTT item bias analysis. The only appropriate procedure for item fairness is DIF. However, the research design does not warrant DIF analysis.

Transfer and Generalizability. The transfer and generalizability criteria of validation seeks to find out the dependability or the degree to which results of the

instrument could be inferred (Pineda, 2012). An assessment instrument ought to be checked for consistency of the results. This is achieved through the CTT reliability or the G-theory. Within CTT, different reliability methods are available depending on factors such as item format, and number of occasions of examination. For instance, if items are administered on two different occasions, the test-retest method is appropriate. If a test is administered on a single occasion, a measure of internal consistency is appropriate. This is possible when items are dichotomously scored. If the items are graded response type, the inter-rater reliability is appropriate.

Theory of generalizability (Quansah, 2020) also provides a framework for determining how generalizable performance assessment outcomes are. At the very least, data on the extent of variability attributable to raters and task sampling are required. Task variability is likely to be significant in performance evaluations in other contexts, such as the military (Shavelson, Mayberry & Rowley, 2012) or medical licensing exams (National Association of Testing Authorities, 2012). When designing an assessment programme, the “limited degree of generalizability across tasks should be taken into account, either by increasing the number of performance assessments given to each student or by using a matrix-sampling design in which different performance assessment tasks are given to separate samples of students” (National Association of Testing Authorities, 2012, p. 5).

According to Hild, et al. (2018), consistency from one component of a test to the next, or from one form to another similar (parallel) form, is not sufficient. There must be justification of the generalization from specific assessment tasks to

the larger realm of achievement, whether it is based on scores on “fixed-response” exams or judgments of scores on PBA such as written essays, laboratory experiments, or student work portfolios. When judging outcomes from classic standardized tests, confirmation of how well the skills and knowledge that lead to strong performance on the traditional assessment questions transfer to other tasks should be demanded (Wilkins, Norton & Boyce, 2013). Evidence of “near and far transfer” is required, such as the ability to apply abilities learned on multiple-choice tests to real-world challenges. Multiple-choice tests are not the only test format that should be concerned about transfer and generalization, however. It is critical in other forms of assessment as well. In this study, the CTT reliability was used to validate the PBA for generalizability by the use of the inter-rater reliability coefficient.

Cognitive Complexity/Difficulty. An effective assessment instrument is one that meet the cognitive level of the examinees (Johnson, 2011, p. 78). It is therefore important to assess every assessment instrument on the basis of the difficulty level of the items for the group. Performance-based evaluations promise to place a stronger emphasis on problem solving, comprehension, critical thinking, reasoning, and metacognitive processes, among other things. These are admirable objectives, but they will necessitate that criterion for grading all forms of assessment pay attention to the procedures that students must perform. For example, it should not be assumed that a hands-on scientific task promote the development of problem-solving abilities, reasoning capacity, or more sophisticated mental models of the scientific phenomenon (Nugroho &

Tomoliyus, 2019). It should also not be assumed that pupils will need to engage in more advanced cognitive processes to solve seemingly more complex, open-ended mathematical issues. The National Academy of Education, (as cited in Schreiber, et al., 2016) stated that the examinee should not be equally confused about the overall goals of both activities” (p. 4).

Depending on the novelty of the challenge and the learner's prior experience, constructing an open-ended proof of a theorem in geometry might be a cognitively challenging activity or just the display of a memorized series of responses to a specific situation. Decisions on an assessment's cognitive difficulty should begin with an analysis of the task, but they should also consider student familiarity with the difficulties and how they seek to solve them (National Association of Testing Authorities, 2012). Analyses of open-ended replies that go beyond general quality evaluations can be very useful in assessing cognitive difficulty of the assessment. It is therefore essential that the criteria for cognitive complexity analysis of the tasks and the nature of the responses that they engender should be included in every assessment.

Content Quality. The subject of content quality should be included among the clear criteria for grading any assessment. The content should reflect what is considered to be aspects of quality that will stand the test of time while also being consistent with the best current understanding of the area (Polit & Yang, 2016; Metzger, Gut, Hild & Tardent, 2014). More importantly, the activities chosen to assess a given content domain should be worthy of students' and raters' time and effort. These factors are especially significant in light of the limited sampling that

performance-based measurements are likely to have (Schreiber, et al., 2016). Poor quality assessments, as well as poor quality instructional material can develop misconceptions regardless of format. Retnawati (2017) stated that subject matter specialists are needed to make systematic assessments on the task quality, similar to the content reviews of items secured by many commercial test publishers. It would also be beneficial to offer evidence of how pupils comprehend the information presented (Nugroho & Tomoliyus, 2019).

One way to ensure the content quality of newer exams is to include subject matter experts not just in task reviews but also in task design (Nugroho & Tomoliyus, 2019). For instance, in America, aside using “award-winning teachers for the selection of primary source material in American History for a newly designed assessment, the scoring criteria for student performance were developed using contrasts between essays composed by active historians and those produced by teachers and students” (National Association of Testing Authorities, 2012, p. 31). This method concentrates on the level of content understanding demonstrated in student responses. In this study, practitioners both mathematics teachers and examiner with enough experience validated the content quality of the items.

Content Coverage: Another potential criterion of interest is the comprehensiveness, which Frederiksen and Collins (as cited in Johnson, 2011) referred to as the scope of content covering. Process sampling takes precedence over traditional content sampling in performance evaluations. The breadth of coverage, on the other hand, should not be underestimated. House (as cited in National Association of Testing Authorities, 2012) and Johnson (2011) found

significant disagreement among historians regarding what should be taught (and graded) in history, implying that a diverse group of content experts may be beneficial in defining content coverage. “If there are gaps in coverage, teachers and students are likely to underemphasize those elements of the content domain that are not assessed”, pg. 23, as Wilkins, Norton and Boyce (2013) pointed out. In this case, the absence of proper material covering definitely resulted in not only misleadingly high ratings, but also a distortion of the instruction delivered. It is possible that the breadth of material coverage and some of the other mentioned criteria will have to be traded off. It is possible that this is one of the reasons by which traditional tests appear to have an edge over more comprehensive performance evaluations. Regardless, it is one of the factors that must be applied to every evaluation.

Educational effect. Educational effect of an assessment instrument means that the instrument facilitates teaching and learning (Sung-Geun & Eun-Hui, 2015). The assessment instrument must make teaching and learning easier. A validation of the educational effect of a classroom assessment instrument is key because all classroom activities including assessment gears towards teaching and learning. According to Wilkins, Norton and Boyce (2013), one of the arguments for more contextualized examination is that they allow students to work on real-world challenges that provide valuable learning opportunities. Analyses of the tasks may yield some useful data for this criterion. Investigations into student and instructor perceptions of performance tests and their reactions to them, on the other hand, would give more systematic data relevant to this criterion (Schreiber, et al., 2016).

Furthermore, research of motivation during large-scale tests, such as the National Assessment of Educational Progress (NAEP), and how it relates to meaningfulness may shed light on why people perform poorly. Like the study of Sung-Geun and Eun-Hui (2015), this study evaluated the educational effect of the instrument by seeking the views of mathematics and examiners; those who matter, on the subject.

Practicality/feasibility. For effective use of any assessment instrument, the instrument must be feasible (Ottawa Conference, 2010). Feasibility means that the instrument should be easy for use, low cost but effective, require less skills, less labour intensive and not too much sophisticated. To be feasible, especially for large-scale assessments, methods for keeping costs low must be devised. Wilkins, et al. (2013) and Nitko (2014) stated that one of the most appealing aspects of paper-and-pencil multiple-choice tests is that they are incredibly efficient and relatively inexpensive when compared to other options. More focus will need to be paid to the development of efficient data gathering designs and scoring techniques as labour-intensive performance assessments become more common (National Association of Testing Authorities, 2012). As a result, this study validated the feasibility of the assessment instrument by the mathematics teachers and examiner with considerable experience.

Credibility. Nitko (2014) stated that the interpretation and use of assessment results is valid only when the educational and social values implied by them are appropriate. It is therefore important to evaluate the degree to which an assessment instrument produces results that reflect students' true performance. By

credibility, the results reflected by the assessment instrument should be trusted and accepted by all stakeholders (Ottawa Conference, 2010). This is to say that stakeholder should have confidence in the instrument to produce reliable results.

In this study, the credibility of the instrument was validated by the mathematics teachers and examiner with considerable experience.

Catalytic effect. One of the principles of assessment states that assessment must serve the need of the learner (Nitko, 2014). Thus, constant feedback must be provided to students to stimulate their learning. At the Ottawa Conference in 2010, it was suggested that every assessment instrument should be validated for catalytic effect, among other things. The catalytic effect is concerned with the effect of the instrument directly on students learning. It seeks to find out if immediate feedback could be given using the instrument, and if the instrument motivates students to learn (National Association of Testing Authorities, 2012). Catalytic effect measures the direct effect of the instrument on students' performance. As was done in the study of Arhin (2015), this study also validated the catalytic effect of the assessment instrument by the mathematics teachers and examiner with considerable experience.

Empirical Studies

Educational effect

Sun-Geun and Eun-Hui (2015) used a quasi-experimental design to evaluate the educational value of performance assessment in science. A simple random sampling technique was used to select two classes each for the experimental and control groups. The experimental group was made up of 79

students while the control was made up of 77 students. A performance assessment was used for the experimental group for nine weeks after which a post test was used to assess students' performance in science. It was found that performance assessment has positive effect on the educational value in the teaching and learning of science.

Kone (2015) evaluated the motivational effect of performance assessment of university learners. A descriptive design was used for the study. A simple random sampling technique was used to select 21 international English for Speakers of other Languages (ESL) students taking an intensive course in oral for non-native speakers. A questionnaire was administered to the students to indicate their level of motivation taking the oral presentation. It was found that the performance assessment has positive effect on the motivation of the students. However, the motivation of the students varied across time and experience.

Catalytic effect

Sung-Eun (2015) investigated into the consequences of implementing performance assessment by conducting a meta-analysis using Hierarchical Linear Modelling (HLM). Most of the studies analysed used quasi-experimental design in the studies. The results indicated that performance assessment improves students learning in the subjects they were used.

Arhin (2015) conducted a quasi-experimental study to find out the effect of PBA driven instruction on SHS students' mathematics performance and attitude at Ghana National College of the Cape Coast Municipality. A simple random sampling technique was used to select two Form One sciences were

selected as the control and experimental groups for the study. An open-ended test and questionnaire were used to collect data on performance and attitude respectively for the study. The results of the independent t test analysis revealed that a significant difference exist in performance and attitude between the experimental and control groups with the post intervention test and questionnaire. The experimental group performed better and showed a positive attitude than the control group.

Feasibility

Metin (2013) conducted a study to determine the difficulties of teachers in preparing and implementing performance assessment task. A case study method was used for the study. Artvin elementary school in Turkey was used as the case of the study. A simple random sampling approach was used to select 25 of the teachers of the school for the study. The sample was made up of 5 science and technology, primary, mathematics, social science and Turkish teachers each. Interview, observation and documentary analysis were employed as the data collection tools and were analysed with content analysis. It was found that teachers lack the requisite skills in developing performance tasks and their rubrics. Teachers also lack knowledge on which topics to use in developing performance task. The study further found that crowded classroom, insufficient time for assessment and learning environment are the challenges teachers face in developing and using performance assessment in schools.

Reliability and Validity

Chan and Malim (2017) conducted a study to gather empirical data on the Teaching Framework for Mathematics (TF@Maths) questionnaire's reliability and validity. A survey of 436 students from the Mathematics Education was done in one public university and one teacher education institution in Malaysia's Northern Zone. Using the Statistical Package for the Social Sciences (SPSS) software version 23, the TF@Maths questionnaire's reliability and validity were assessed using Cronbach's alpha and Exploratory Factor Analysis (EFA). The items were then subjected to EFA utilizing principal component analysis extraction and Varimax rotation to determine validity. There were 62 items that retained factor loadings greater than 0.4. The TF@Maths yielded six factors, according to the factor analysis: mathematical content knowledge, mathematical pedagogical knowledge, general pedagogical knowledge, classroom management skill, mathematics disposition and quality mathematics teacher. The overall score of the Cronbach's alpha test was 0.939, indicating that the items in the instrument are very reliable. Unfortunately, content validity was not evaluated. Only construct validity was established.

Reid (2014) validated a developed instrument for measuring interest. The instrument was validated for validity and reliability. A total of 15 items on interest was administered to 53 students on two occasions. Cronbach alpha and the confirmatory factor analysis were used to evaluate the reliability and validity respectively. Reliability with Cronbach were 0.851 and 0.822 respectively. Confirmatory factor analysis factor ranged from 0.453 to 0.859 high loading on

attitude factor. The gap in the validity evaluation was that only construct validity was evaluated. There was no evaluation of the content validity. Content validity is useful source of information to validate an assessment instrument (Nitko, 2014).

Once the instrument was administered at two different occasions, the test-retest method would be a better method for estimating the reliability of the instrument. This is because, the Cronbach alpha estimate the internal consistency of the instrument.

Hasnida and Ghazali (2016) created and tested an instrument to assess the validity and reliability of teachers' perceptions on SBA implementation in schools. The CIPP (context, input, process, and product) Evaluation Model, established by Daniel Stufflebeam, serves as the foundation for the instrument. A total of 120 primary and secondary school instructors were given the instrument in the form of a questionnaire. The response rate was set at 80%. Internal consistency reliability, which is determined by alpha coefficient reliability or Cronbach Alpha, was used to assess the instrument's reliability. The results of this pilot investigation indicated that the instrument was reliable. The reliability coefficient was 0.867. Experts analysed the content validity, whereas Exploratory Factor Analysis was used to assess the construct validity (EFA). Finally, depending on the loadings, 69 of the 72 pieces were kept. The results of this pilot investigation demonstrated that the instrument is reliable.

Through a two-step approach, Zamanzadeh, Ghahramanian, Rassouli, Abbaszadeh, Alavi-Majd, and Nikanfar (2015) investigated the content validity of the patient-centered communication instrument (development and judgment).

Domain determination, sampling (item creation), and instrument formation were performed in the first stage, while content validity ratio, content validity index, and modified kappa statistic were performed in the second step. The instrument face validity was tested using expert panel suggestions and item impact scores. Trust building (eight items), informational support (seven items), emotional support (five items), problem solving (seven items), patient activation (10 items), intimacy/friendship (six items), and spirituality strengthening are among the seven dimensions identified by the content validity process from a set of 188 items (14 items). The instrument has an appropriate level of content validity, according to the content validity study. The instrument's overall content validity index using the universal agreement technique was low; however, given the large number of content experts who make consensus difficult and the high value of the S-CVI with the average approach, which was equivalent to 0.93, it can be justified. Surprisingly, an important component of validity, construct validity, was not evaluated.

Experience and educational effect

In Ado-Odo/Otaand Ifo Local Government Areas in Ogun State, Ewetan and Ewetan (2015) evaluated the impact of teachers' teaching experience on the academic performance of public secondary school students in Mathematics and English Language. The research was conducted using a descriptive research approach. The study included all 31 Senior Secondary Schools in the two local government regions chosen. A total of 20 schools were selected from the population using a basic random selection technique, including 14 schools in the

Ado-Odo/Ota Local Government Area and 6 schools in the Ifo Local Government Area. The instrument for data collection was an inventory schedule. A total of 388 out of 400 surveys, or 97 percent, were returned, with 20 preschool questionnaires administered. Content analysis was used to examine their responses. At 0.05 alpha level, the regression analysis and t-test were employed to examine hypotheses produced for the study. Instructors' teaching experience has a considerable impact on teachers' assessments of students' academic performance in Mathematics and English Language, according to the findings. Schools with more teachers with more than 10 years of teaching experience outperformed schools with less teachers with 10 years of teaching experience.

Experience and feasibility

Secondary school mathematics instructors' categorisation of concepts along the axes of difficulty level was explored by Iji and Omenka (2014). The goal was also to see how mathematics teachers' cognitive beliefs and conceptualizations influenced their perceptions of whether or not mathematics learning elements were difficult. This research enlisted the help of 95 secondary school mathematics teachers. Algebra, number and numeration, geometry, trigonometry, and statistics were all covered by the device. The instrument's principles were also taken from the West African Examinations Council's O' level syllabus and the Nigerian Educational Research and Development Council's (NERDC) secondary school mathematics curriculum. The results of the analysis of the responses of the individuals in the study revealed that there was little agreement in the classifications of mathematics concepts. Only five issues had a

moderate level of agreement. The majority of the items, it appears, are significant but simple to understand and teach, according to the maths teachers.

Chapter Summary

Generally, literature has revealed that PBA is the appropriate assessment method which allows both content and procedural skills to be measured. As far as assessment in mathematics is concerned, PBA provides a better option for the assessment procedures.

The nature of PBAs necessitates the exploration of a slew of difficult conceptual, measuring, and statistical difficulties. The ultimate purpose of such research is to give proof that inferences made about people's scores are correct.

The cost of a good assessment instrument must be justified. This means that the assessment procedure must be cost effective. Aside cost effectiveness, the assessment procedures should motivate students to learn (educational effect) and that it should be possible to provide immediate feedback to students to stimulate their learning (catalytic effect). Furthermore, the instrument must be usable and the results with credibility. Performance-based assessment is found to have the educational and catalytic effects as well as feasibility and credibility. The validation of any assessment instrument must cover these criteria as well as good psychometric properties. Again, the literature reviewed revealed that the methodology for the development and validation of instrument could lend itself to either quantitative or mixed methods approach.

CHAPTER THREE

RESEARCH METHODS

Introduction

The purpose of this study is to develop and validate a PBA instrument in mathematics for SHSs. The methodology chapter was in line with the stages for development and the validation of instrument. For this study, the instrument was a PBA items in mathematics for SHSs.

Research Philosophy

The destination of focus for the positivist and interpretivist is the same; knowing the reality of the world (Fraenkel & Wallen, 2000). The difference comes with how each philosophy sees reality, how to know what is reality, perception about human being and the role of the enquirer in the process of knowing. The study therefore adopted the positivist philosophy of research. Positivism claims that their approach of knowing or gaining knowledge is more certain and objective than knowledge which originated from other paradigms. Positivists are of the view that “reality is stable and can be observed and described from an objective viewpoint without interfering with the phenomena being studied” (Fraenkel & Wallen, 2000, p. 98). Thus, the study employed the quantitative approach of research.

Research Design

This study employed quantitative instrumentation research design, used for developing and/or testing instruments by Benson and Clark (1982) and Onwuegbuzie, Daniel and Collins (2009). The design is made up of a four-phase

instrument development and validation process: planning, construction, qualitative evaluations, and quantitative validation. The ‘minority-qualitative’ aspect of the Benson and Clark (1982) was changed to quantitative in this study.

It is specifically advantageous for researchers developing a new instrument (Creswell, 2009). It is specifically advantageous for researchers developing a new instrument (Creswell, 2009). “Quantitative research allows researchers to be independent in exploring their ideas on developing proper guidelines for their studies, and it seeks a reality that is objective, singular, and that can clarify existing theories” (Creswell, 2013, p. 123). Park, Bahrudin and Han (2020) asserted that quantitative research lends itself to deductive reasoning. It begins the idea of research with general concepts and moves to specific concepts.

For instrument development, Wyatt (2016, p. 35) suggested to first “obtain themes and specific statements.” In this study, the researcher formulated the themes for the instrument. The experience of the researcher as a mathematics examiner and teacher was used in formulating the themes. These data became the basis for developing the PBA items and the rubric. Based on the information, table of specifications, was prepared to guide the development of the instruments to ensure content validity and relevance (Newman, Lim & Pineda., 2013).

In order to get comprehensive data for the development and validation of the proposed PBA instrument, data were also gathered from literature on similar issues (Pineda, 2012). The information from the literature was used in establishing the criteria for validating the instrument as feasibility, credibility, educational and catalytic effects as well as the psychometric properties. Then,

data were collected from mathematics teachers and examiners to validate the instrument in terms of feasibility, credibility, educational and catalytic effects. Also, quantitative data were collected from students on the PBA items to estimate the psychometric values of the items.

Study Area

The study was conducted in the Western Region of Ghana. The region is bordered by Central Region to the east, Western North to the North, Cote d'Ivoire to the West and the Atlantic Ocean to the South. It is the host to almost all the resources (gold, oil, cocoa, rubber (latex), fish, and timber) of the country. The predominant occupations in the region are primary occupation such as farming, fishing, and mining. However, some towns like Bogoso and Agona Nkwanta have notable market centres. Sekondi-Takoradi, the regional capital, and its environs is the only industrial area in the region.

The nature of occupation in the region attracts a lot of the students thus affecting school enrolment, punctuality, and regularity, which altogether affect academic performance. The region has only five Category A schools (according to GES classification) with majority of the schools in Category C. The students in the SHSs in the region are mostly from communities within the region. Only few students from other regions attend SHSs in the region. Thus, the background of the students reflects the predominant characteristics of the region. Almost all the schools (with the exception of the five Category A schools in the regional capital) have students with similar entry characteristics. Only Sekondi-Takoradi, the regional capital and Tarkwa, a mining town, are cosmopolitan.

Population

The population for the study made up of WAEC mathematics examiners and teachers, and public SHS Three students in the Western Region of Ghana.

The Western Region of Ghana was selected because it where the problem was conceptualised by the researcher. The researcher is a mathematics teacher and mathematics examiner in the Western Region. There are 275 mathematics examiners in the region (WAEC, 2019), 321 mathematics teachers and 7498 SHS 3 from 35 SHSs in the region as at 2019 (GES, 2019). Out of the 35 schools, five (5) are in category A, 12 in category B and the remaining 18 in category C (GES, 2019).

The accessible population comprised SHS 3 students and mathematics teachers in the 15 SHSs selected for the study. The accessible population also included the mathematics examiners in the region. The distribution of the target and accessible population is presented in Appendix A.

Sampling Procedures

A multistage sampling procedure was used for the selection of respondents for the study. The study made use of stratified, simple random, census and purposive sampling techniques.

For the validation phase, all mathematics teachers in the selected schools who satisfied the inclusion condition were selected via census. The mathematics teachers should not be mathematics WAEC examiner and have taught for not less than one year. This was to avoid one person responding as an examiner and as a teacher. Also, census method was used to select all WAEC mathematics

examiners in the region with the exception of first-time examiners. First-time examiners and teachers with less one year experience were excluded because they have not had any significant experience in marking WAEC examinations. For content validity, WAEC mathematics examiners who are team leaders or Heads of mathematics Departments (HODs) of the selected SHSs and are not mathematics examiners were purposely selected to indicate relevant/not relevant for each item based on their expertise. They were 35 in all. Sample of 2-40 experts is enough for content validity (Lawshe, 1975)

In the first stage of the second phase of quantitative evaluation (psychometric), a stratified sampling technique (Neuman, 2003) was used to select 15 SHSs. The Ghana Education Service's category of schools was used as the strata. Five schools from each categories A, B and C were selected. In the next phase, a simple random sampling technique was used to select two SHS 3 classes from each school selected. The number of SHS 3 classes in the selected schools ranges from 7-19. Each individual in the population of interest had an equal likelihood of selection (Creswell, 2013; Cooper & Schindler, 2009). Each unit in the population was identified, and each unit had an equal chance of being in the sample. Selection of one unit did not affect the chances of any other unit (Adjei & Tagoe, 2009; Cohen, Manion & Morrison, 2000).

Mugenda and Mugenda (as cited in Ankomah, 2015) recommend that, for descriptive studies, 10% or above of the target population is enough for the entire study. The researcher wanted the sample selected from each group to be the same for easy analysis. This supported the stance to select five schools from each

stratum and two SHS 3 classes from each school as an ideal sample size for the study. By census, all students in each class were selected for the study.

In all, sample of 240 mathematics examiners, 150 mathematics teachers and 750 SHS Three students in the Western Region were used for the validation phase of the instrument development.

Data Collection Instrument

The instruments for the data collection of the study were the on-demand Performance-based Items in Mathematics (odPIM) developed by the researcher (Appendix B2) and questionnaire (Appendix C). The researcher named the developed instrument odPIM for easy reference. The researcher with his experience as a mathematics teacher and examiner, analysed the content of the mathematics syllabus to consider the most popular concepts on which the items would be developed. Five major topics which draw knowledge from other topics were considered: Rigid Motion as Item 1, statistics as Item 2, mensuration as Item 3, geometric construction as Item 4 and equation as Item 5. Based on the topics and with literature, the researcher developed the five-item test in PBA.

The PBA test items were constructed by the researcher with the help of test specification. There were five items on the test to be responded to in 2 hours. All tasks were on-demand type where the students were expected to complete all at a sitting within the given period of time (Brennan, 2006). The purpose for the test was to obtain scores that were used for the item parameters of the PBA items.

The test was used for the psychometric parameters of the instrument designed in terms of reliability and construct validity while the questionnaire was

used for the evaluation of the instrument in terms of educational and catalytic effects as well as the feasibility and credibility of the instrument. The questionnaire was designed using the Likert type scale format with close ended questions. Osuola (2001) asserted that questionnaire is at its best whenever the sample size is relatively large enough to make it uneconomical for reasons of time or funds or almost impossible to observe or interview every subject. It was used to find out information on the PBA as to whether it is feasible in terms of development, administration, scoring and interpretation, credibility of students' results and impact on students learning. The four-point Likert type scaled questionnaire was mainly used and had various score values. Positive statements were scored as Strongly agree (SA) = 4, Agree (A) = 3, Disagree (D) = 2 and Strongly Agree (SD) = 1.

The questionnaire was of five parts; the first was on the bio-data of respondents, the second part looked at the feasibility of the PBA items. The third part of the questionnaire for the evaluation of the PBA items looked at the credibility of the PBA items, fourth part was meant to find out the educational effect and the final part found out the catalytic effect. An additional section meant to evaluate the content validity was added. The respondents were to indicate relevant/Not relevance for each of the five items.

Initially, 42 items to validate the instrument were crafted. The questionnaire was given to three lecturers to establish the face validity and make any necessary corrections in the wording. Some of the items were revised. For example, an item that read "is practicable for large scale assessment" was revised

to “PBA is practicable is for large scale assessment”. Items that that were found problematic were removed. Three items were removed by this process. The remaining 39 items were administered to 50 teacher and examiners for pilot testing of the instrument.

Data collected for the final study from the pilot testing were analysed in two stages: (1) EFA and (2) CFA. Principal components extraction method and varimax orthogonal rotation were used in the first stage to extract the factors and their factor loadings. Four factors were extracted from EFA and this represented the four dimensions of the questionnaire for the validation: feasibility of the instrument, credibility of the instrument, educational effect, and catalytic. The loadings of the items ranged from .515 to .931, above the cut-off value of .50 as recommended by Hair, Black, Babin, and Anderson (2010). Three items out of the 39 had loadings below .50 so these items were discarded. Finally, 36 items were maintained. These comprises 13 items on feasibility, Cronbach’s alpha = .892; seven items on credibility, Cronbach’s alpha = .704; nine items on educational effect, Cronbach’s alpha = .806; and seven items on catalytic effect, Cronbach’s alpha = .703. The average variance extracted (AVE) estimates for feasibility, credibility, educational effect, and catalytic effect are above the cut-off value of .50. Overall, the results presented good reliability of the scale measured and signified the convergent validity of the questionnaire.

Validity of Instrument

The content validity and construct validity of the questionnaire for the evaluation of the developed instrument was established by submitting the

questionnaire to lecturers of the Department of Education and Psychology whose area of specialisation are Educational Measurement and Evaluation and Research Methods, for their scrutiny and critique. Suggestions that were made by them helped improve the content and construct related evidence of validity of the questionnaire. The questionnaire was subjected to exploratory and confirmatory factor analysis to ascertain the components and factor loading of the items (Appendix D).

The PBA items were also given to mathematics teachers and examiners for content validity, administered to students to check construct validity and reliability as part of the objectives of the study. A sample of the questionnaire which would be used for the validation was attached to each set of the instrument. The mathematics teachers and examiners were to respond to the questionnaire after their assessment of the instrument. Items that were ambiguous were reframe. The data from the responses of the mathematics teachers through the pilot testing were analysed using exploratory and confirmatory factor analysis. Four factors were explored and items with acceptable level of factor loadings were included in questionnaire.

Pilot-testing of the instrument

The questionnaire for the evaluation of the developed PBA items was pilot-tested on fifty (50) examiners and mathematics teachers selected from the Ahanta-West Municipality in the Western Region. Specifically, Sankor High School was used for the pilot testing of odPIM. The sample for the pilot-testing was randomly selected from the Ahanta –West. The selection of Ahanta West

Municipality was on the basis that students in Sankor Senior High have same characteristics as those at Baidoo Senior High/Technical school selected for the study. The result thus represented the sample for the study. The results of the pilot-testing of the instruments helped improve the quality of the items.

The developed PBA items were administered to two Form Three classes of one of the schools not selected for the study. This was also to ensure construct-related evidence through ‘think aloud’ (Sarantakos, 2000). The

Reliability of Instrument

The pilot-testing results were used to determine the reliability of the questionnaire with the Cronbach’s Alpha (α) measure of internal consistency. The IBM Statistical Product and Service Solution (SPSS) Version 21.0 was used for the computations. The result of the Cronbach alpha of the scales of the questionnaire is presented in Table 4.

Table 4- *Cronbach alpha of the Questionnaire*

Scale	Number of items	Cronbach alpha	Remarks
Feasibility	13	0.892	Acceptable
Credibility	7	0.704	Acceptable
Educational effect	9	0.806	Acceptable
Catalytic effect	7	0.703	Acceptable
Overall instrument	36	0.843	Acceptable

Source: Field Data (2020)

Ethical Consideration

Ethical clearance was acquired from the Institutional Review Board in the University of Cape Coast. The clearance spelt out the purpose of the study, the need for individual participation, anonymity as well as confidentiality of respondents' responses. With the ethical clearance, informed consent was sought from participants (students above 18 years, mathematics teachers and examiners, assessment experts and the WAEC mathematics chief examiner) by explaining the purpose of the study to them. For students who were less than 18 years, the consent was sought from the senior house master of the school.

Anonymity of respondents was given a priority. Neither names nor any identifiable information from respondents were recorded or taken. This was done to ensure that participants' identities were hidden. In order to ensure confidentiality, participants were assured that their responses would be kept secretly, and that no individual known to them would have access to the information that they would provide.

Data Collection Procedures

An ethical clearance form and an introductory letter (Appendix E) were taken from the Institutional Review Board in the University of Cape Coast and Department of Education and Psychology respectively to seek permission from the various schools where the study was carried out. With the permit from the University, the consent of the various headmasters and regional head of WAEC

was sought to conduct the study in the schools and centre respectively. Also, students' consent was sought for support and collaboration.

In the first phase, the questionnaire was administered to the 240 mathematics examiners at the marking centre at the time of conference marking and coordination with a sample of the PBA test items attached to the questionnaire. This was to provide quantitative information on the instrument in terms of feasibility, credibility, educational and catalytic effects. To strengthen the quantitative data from the examiners, the questionnaires with a sample of the PBA test items was administered to the 150 selected mathematics teachers in their respective schools. The purpose of the questionnaire was to elicit information on feasibility, credibility, educational and catalytic effects. In the final phase, selected class of students sat for the PBA test. This test was supervised under external examination conditions. The purpose of the test was to estimate the psychometrics properties of the test (inter-rater reliability and construct validity). Two assistants were trained and deployed for the data collection. They were trained for the administration of the questionnaire and odPIM. Specifically, their training included explanation of the questionnaire, the purpose of the study and retrieving of the questionnaire and test papers from respondents.

Stages of the development and validation of the PBA

The development and validation of the assessment instrument followed the four-phase instrument development and validation by (Benson & Clark, 1982) as presented in Figure 1.

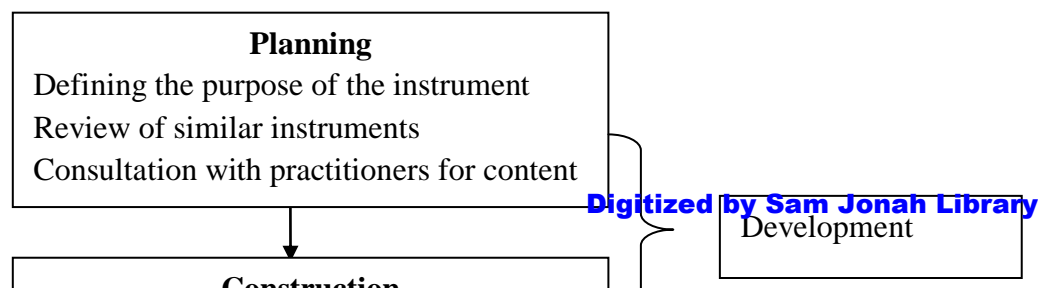




Figure 2- Conceptual Framework for the development and validation of the PBA
Source: Adapted (Benson & Clarke, 1982)

The development stage of the instrument involves the planning and construction of the instrument. The evaluation and validation stages constitute the actual validation of the instrument.

Planning

The purpose of this phase was to determine the instrument's goals and objectives. There are two stages to the planning process. The first section includes a review of previous research as well as an assessment of similar or related survey instruments. The goal of this method was to comprehend and define the existing state of research, as well as to identify research gaps that the instrument may potentially fill (Gable & Wolf, 2012). This study's literature review comprised a

research review on PBA, content and construct validity of graded response, item reliability estimates for graded response items, and instruments developed to measure a mathematical construct.

The second part of the planning phase was a further review to examine the traditional items and how the PBA items in mathematics could be developed to encourage students to apply knowledge to real life situations (Gable & Wolf, 2012). The findings laid the foundation for the item construction and the categories of themes: 1) feasibility, 2) educational effect, 3) catalytic effect, and 4) acceptability. The themes constituted the foundation for the study's findings. These themes were then compared to established instruments and themes that came from the literature. The instrument's goals and objectives were written based on the themes that surfaced, and a list of possible uses for the instrument was compiled.

Construction

The goal of the construction phase was to create an item pool and compare it to previous research to see if it was construct valid. Consistent with the AERA, APA, and NCME (2014), an item pool of five items were constructed based on the concepts in the SHS core mathematics syllabus. The first source was questions that arose from a review of the existing instruments surrounding an appropriate assessment instrument that can replace the traditional assessment instrument being used in SHS. The second source for developing the item pool was content analysis. To ensure that content validity of the instrument, a table of specification

was constructed for representation of the contents learnt (Crocker & Algina, cited in Armah, 2018).

Test specification was designed to guide the construction of the instrument (Appendix B1). Every test construction begins with defining the target construct to be assessed and translating that into test specification. The test specification allows alternate forms for the tasks to be constructed (Nitko, 2004). With the test specifications, the PBA items were constructed. Dillman, Smyth and Christian (2014) proposed guidelines for developing good assessment items. Each of the item was to be checked to ensure relevance, language simplicity, technical accuracy, and proper sentence structure. The items were read by three people for clarity and face validity (Dillman, Smyth & Christian, 2014). These readers were chosen since they were neither practitioners nor specialists in the subject, and they were given the task of proofreading for grammatical errors, any unclear sentence structure, and new or undefined vocabulary. The goal of this preliminary proofing was to reduce construct-irrelevant bias and ensure that the language burden was suitable (American Educational Research, 2014). The item pool was then typed and saved on a computer once it had been modified and revised. Specific care was given to layout.

Validation of the instrument

The validation section of the study covered the last two phases of the instrument development phases by Benson and Clark (1982) which are the validation and quantitative evaluations. Figure 3 shows the stages for the development and validation of the PBA items.

At the validation phase which preceded the construction (item writing) phase, a questionnaire with a sample of the developed PBA items was administered to a larger sample of practicing teachers and WAEC examiners to also validate the PBA items with regard to 1) feasibility, (2) educational effect, (3) catalytic effect, and (4) acceptability. This is to get a broad picture and quantitative information on the validity of the instrument.

The psychometric properties: reliability and validity indices were estimated. The summary of the stages of activities in the development and validation of the PBA items is presented in Figure 2.

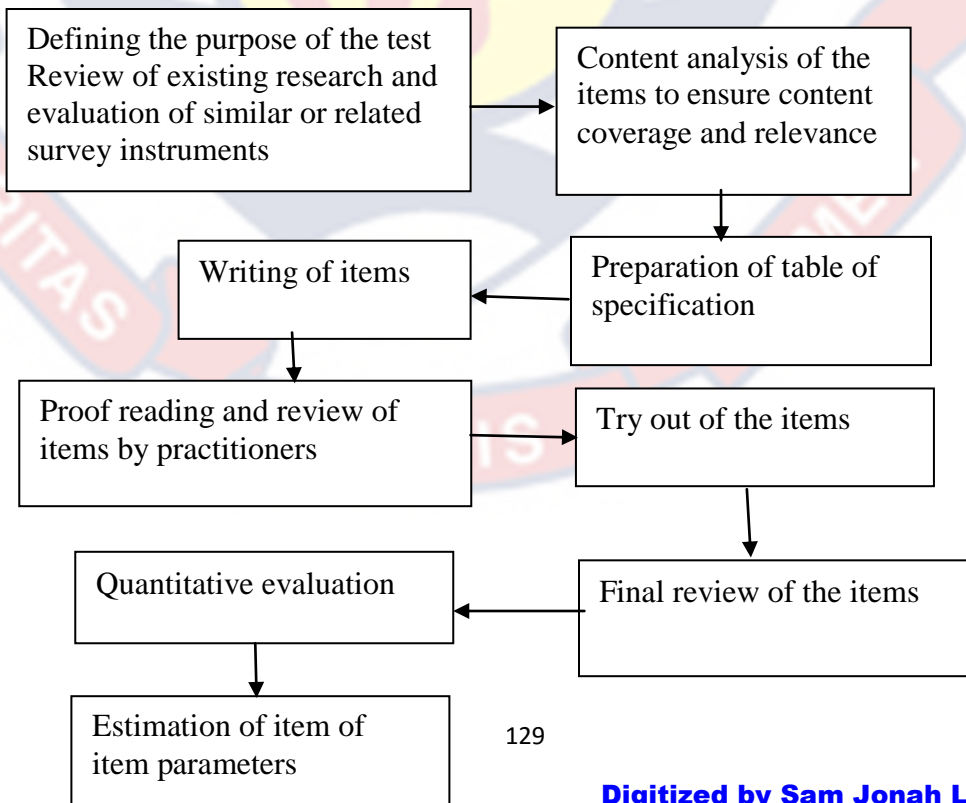


Figure 3- Stages of validation of the instrument

Source: Author's own (2019)

Data Processing and Analysis Procedure

For the questionnaire, the scoring was reversed for the negative statements as (SA) = 1, Agree (A) = 2, Disagree (D) = 3 and Strongly Agree (SD) =4 for items on the feasibility, credibility, educational and catalytic effects. This was done to leave the results in the positive form to make interpretations and discussion easy. The test was marked by mathematics teachers who are also WAEC mathematics examiners. The scripts were distributed among 3 examiners (including the researcher). Sixty scripts were randomly selected from each category of school, photocopied for each examiner to mark. This was done to evaluate the inter-rater reliability of the instrument. For the purpose of the tasks and the constructs, the primary trait scoring procedure was used in preparing the scoring rubric and the top-down approach by Nitko (2004) was used (Appendix B3). Mathematics involves series of processes, methods and procedures for completing tasks which need to be assessed. The various constructs for each task were identified, weighted and scored for each examinee. Quantitative data on the feasibility, credibility, educational and catalytic effects for research question one to four were collected with questionnaire.

For research questions one to four, data were analysed with means and standard deviations. This is because the research questions sought to examine

participant perception of the quality of the instrument in terms of educational effect, catalytic effect, feasibility and credibility.

Data on research question five were analysed using the Pearson Moment correlation coefficient. The research question sought to evaluate the reliability of the instruments both at the item and test level. The data is interval and administered once. The inter-rater reliability was considered the best approach for the estimate of reliability. Data on research question six was analysed with confirmatory factor analysis and principal component analysis. These were used to evaluate the convergent construct validity of the instrument. The modified kappa statistic was used to evaluate the content validity of the instrument.

Data on all research hypotheses were analysed with four-way ANOVA. There were four independent variables and one dependent variable in each case. The independent variables were status and experience. The dependent variables were 1) feasibility of the developed PBA, 2) credibility of the instrument, 3) educational effect, and 4) catalytic effect.

Chapter Summary

The study sought to develop and validate the PBA for SHSs. The chapter elaborated the research methods employed. A four-phase instrument development and validation was used. The data were analysed using descriptive statistics (means and standard deviations), four-way ANOVA, Pearson Product Moment correlation coefficient, confirmatory factor analysis and principal component analysis and modified kappa statistic.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

The purpose of the study was to develop and validate the PBA items in mathematics for SHSs. This chapter dealt with the analysis and presentation of the data collected from WAEC mathematics chief examiner, mathematics examiners and students that participated in the study. Quantitative data were collected. The data were analysed and discussed based on the research questions and hypotheses.

The data were analysed using frequency and percentages, means and standard deviations, four-way ANOVA, Pearson Product Moment correlation coefficient, confirmatory factor analysis and principal component analysis and modified kappa statistic. The first part of chapter described the demographic characteristics of respondents and checking of assumptions for the ANOVA analysis. In the second part, the research findings are presented based on the research questions and the hypotheses. The last part detailed the discussion of the findings of the study.

Analysis of Bio-Data of Respondents

The study was carried out in the Western Region, with a sample size of 750 students, 240 mathematics examiners, and 150 mathematics teachers. Frequencies and percentages were used for the analysis. The distribution of students, mathematics examiners and teachers by gender is presented in Table 5.

Table 5-Distribution of Respondents by Gender

Gender	Students		Examiners		Teachers	
	N	%	N	%	N	%
Male	388	51.73	199	82.92	115	76.67
Female	362	48.27	41	17.08	35	23.33
Total	750	100	240	100	150	100

Source: Field data (2020)

Table 5 shows the distribution of the students, examiners and teachers of the study by gender. The Table shows that for students, 388 of the respondents representing 51.73% are males while 362 representing 48.27% are females. The study, therefore, revealed that majority of the SHS 3 students are males. The Table again shows that for examiners, 199 out of the 240 representing 82.92% are males while the remaining 41 representing 17.08% are females. The study, therefore, revealed that majority of the WAEC mathematics examiners are males. Also, the Table shows that for teachers, 115 of the respondents representing 76.67% are males while the remaining 35 out of the 150 representing 23.33% are females. The study, therefore, revealed that majority of the mathematics are males. The study revealed that generally, males dominate females in all areas ranging from learning to practice of mathematics. The distribution of teachers and examiners by experience is presented in Table 6.

Table 6- *Distribution of Examiners and Teachers by Years of Experience*

Experience	Examiners		Teachers	
	N	%	N	%
1-5	87	36.25	60	40.00
6-10	70	29.17	55	36.67
11-15	46	19.17	27	18.00
16-20	13	5.42	8	5.33
Above 20	24	10.00	0	0.00
Total	240	100	150	100

Source: Field data (2020)

Table 6 shows the distribution of examiners and teachers by experience. The Table shows that for examiners, 87 out of the 240 mathematics examiners representing 36.25% have 1-5 years of experience, 70 representing 29.17% have 6-10 years of experience, 46 representing 19.17% have 11-15 years of experience, 13 representing 5.42% have 16-20 years while 24 representing 10.00% have more than 20 years of experience. The study, therefore, revealed that majority of the examiners have more than five years of experience which is enough to give reliable information with respect to the feasibility, credibility, educational effect and catalytic effect of the PBA on students.

Table 6 further showed that for teachers, 60 out of the 150 mathematics teachers representing 40.00% have 1-5 years of experience, 55 representing 36.67% have 6-10 years of experience, 27 representing 18.00% have 11-15 years of experience while eight representing 5.33% have more than 20 years of experience. The study, therefore, revealed that majority of the examiners have more than five years of experience which is enough to give reliable information

with respect to the feasibility, credibility, educational effect and catalytic effect of the PBA on students.

Analysis of Data on Research Questions

A questionnaire was admitted to the 1) mathematics teachers, 2) mathematics examiners on the same themes; feasibility, credibility, educational and catalytic effect that were evaluated by the assessment experts and the WAEC officer. A list of items that measures the feasibility, credibility educational and catalytic effects were given to respondents to indicate their extent of agreement to the statements. For the questionnaire, on a four-point Likert-type scale 4 = strongly agree, 3 = agree, 2 = disagree, and 1 = strongly disagree, teachers and examiners were asked to indicate their levels of agreement or disagreement with statements posed on the feasibility, credibility, educational and catalytic effect of the PBA items for SHSs. The data were analysed using means and standard deviation. The total value of the scores is 10 (4 + 3 + 2 + 1). This gives a mean of 2.5 for each of the responses out of the total of 4. That is the total 10 divided by the 4 responses. The 2.5 is also the middle point for the four –point scale. The difference of the minimum of 1 and 2.5 which gives 1.5 is divided into 2 making 0.75. Therefore, the mean cut-off points for the questionnaire for the variables were: 3.25 – 4.00 = strongly agree, 2.50 – 3.25 = agree, 1.75 – 2.50 = disagree and 1.00 – 1.75 = strongly disagree. A mean of 2.50 and above indicates respondents' agreement while a mean of 1.75 and below indicates respondents' disagreement. The means of the items were estimated by adding up all the

responses to each item by each respondent and then dividing by number of respondents who responded to that particular item.

Research Question One

What is the feasibility of the developed PBA?

The descriptive statistics of the results on the feasibility of the PBA items for SHSs is presented in Table 7.

Table 7- *Descriptive statistics of the Results by Mathematics Teachers and Examiners on the Feasibility of PBA items for SHSs (N = 390)*

S/N	Statement	M	S D
1	Marking of script will comparatively be of the same time as the traditional system	3.19	.666
2	Same number of scripts could be marked in the PBA as in the traditional system could be marked by an examiner	3.28	.708
3	Scripts marking of PBA will be of the same difficulty as the traditional	3.07	.707
4	Same number examiners for the traditional system could finish marking the PBA items	3.28	.642
5	Constructions of the PBA items will not be difficult just like the traditional system	3.04	.751
6	Construction of alternate forms of the PBA is feasible	3.25	.678
7	With a well-designed test specification, alternate forms can be created	3.30	.628
8	Item constructions of the PBA will require much time and skills	3.27	.601
9	The PBA will be able to cover all content learned in a single test	3.26	.657
10	Student could be assessed with PBA within the allotted time	3.18	.636
11	Materials for using the PBA for examinations are available	3.12	.645
12	Use of PBA would not produce extra cost to the assessment system	3.35	.557
13	The PBA is practicable for a large number of examinees	3.23	.546
Mean of Means		3.23	
Mean of Standard deviation			.648

Source: Field data (2020)

The results in Table 7 show that generally, the teachers and examiners agree with the statements concerning the feasibility of the PBA items for SHSs. It was realized that the mean of means; $M = 3.23$; $SD = 0.648$ is greater than the cut-off mean of 2.50 indicating that the teachers and examiners agreed with the statement on feasibility of the PBA items for SHSs in the Western Region of Ghana. The results revealed that the teachers and examiners believed that there is feasibility of the PBA instrument for SHSs.

All the 13 items on feasibility of the PBA items for WAEC examination had means greater than the average mean of 2.50 meaning that the examiners and teachers agree to all the statements on the feasibility of the PBA items for SHSs. Out of the 13 items, the teachers and examiners expressed that eight are more feasible because the means were greater than the mean of means of 3.23. The result of the items with more feasibility is presented in Table 8.

Table 8- *Results of More Feasible Statements of the PBA by Mathematics Teachers and Examiners (N = 390)*

S/N	Statement	M	S. D
1	Same number of scripts could be marked in PBA as in the traditional system could be marked by an examiner	3.28	.708
2	Same number examiners for the traditional system could finish marking the PBA items	3.28	.642
3	Construction of alternate forms of the PBA is feasible	3.25	.678
4	With a well-designed test specification, alternate forms can be created	3.30	.628
5	Item constructions of PBA will require much time and skills	3.27	.601
6	PBA will be able to cover all content learned in a single test	3.26	.657
7	PBA would not produce extra cost to the assessment system	3.35	.557
8	PBA is practicable for a large number of examinees	3.23	.546

Source: Field data (2020)

The table shows that the PBA is feasible to be used for SHSs in that same number of examiners and scripts as in the traditional system could be used in the marking, construction of alternate forms is feasible, item constructions will not require much extra time and skills and that representative content could be covered and learned for a single test. The Table further showed that the use of PBA would not produce extra cost to the assessment system and that it is practicable for a large number of examinees.

Research Question Two

What is the credibility of the developed PBA items?

The descriptive statistics of the results on the credibility of the developed PBA items is presented in Table 9.

Table 9-Descriptive Statistics of the Credibility of the Developed PBA items in Mathematics by the Mathematics Teachers and Examiners (N = 390)

S /N	Statement	Mean	S D
1	Results reflect students' true performance	3.43	.612
2	Malpractice associated with examination is reduced	3.49	.521
3	The results from the PBA can be trusted	3.25	.499
4	Differences in students' performance become real	3.36	.617
5	Knowledge level becomes the same as application level	3.28	.620
6	The PBA provides accurate estimation of student performance	3.26	.441
7	Results from the PBA could be generalized	3.26	.618
Mean of means		3.33	
Mean of Standard deviation			.651

Source: Field data (2020)

Table 9 shows that generally, the teachers and examiners strongly agree to the statements on the credibility of the developed PBA items. This is because the mean of means; $M = 3.33$; $SD = 0.651$ which is greater than 2.50 lies in the cut-off point of strongly agree. The results revealed that the teachers and examiners believed that the PBA is credible for SHSs.

All the seven items on the views of teachers and examiners on the credibility of the developed PBA had means greater than the average mean of 2.50 meaning that the examiners and teachers agree to all the statements on the views of teachers and examiners on the credibility of the developed PBA items. Out of the seven items, the teachers and examiners expressed three as more credible characteristics of PBA; results reflect students' true performance ($M = 3.43$, $SD = .612$), malpractice associated with examination is reduced ($M = 3.49$, $SD = .521$) and that differences in students' performance become real ($M = 3.36$, $SD = .617$). This is evidenced by the means which are greater than the mean of means of 3.33.

Research Question Three

What are the educational effects of the developed PBA?

The descriptive statistics of the results on whether the PBA items could stimulate students' learning of mathematics is presented in Table 10.

Table 10-*Descriptive Statistics of the Educational Effect of the PBA items in Mathematics by the Mathematics Teachers and Examiners. (N = 390)*

S /N	Statement	Mean	S. D
1	Students will be compelled to learn	3.55	.518
2	Students are motivated to learn	3.57	.516
3	Encourages students to think differently on an issue	3.43	.516
4	Causes students to think critically on problems	3.22	.529
5	Encourages students to learn extensively	3.40	.705
6	Makes learning easier	3.18	.717
7	Encourages learning every domain	3.06	.590
8	Can be integrated into the teaching and learning processes	3.22	.597
9	Encourages learning of mathematical skills	3.30	.612
Mean of means		3.33	
Mean of Standard deviation			.589

Source: Field Data (2020)

Table 10 shows the results of the views of teachers and examiners on whether the PBA items could stimulate students' learning. The results show that generally, the teachers and examiners strongly agree to the statements on whether the PBA items could stimulate students' learning. This is because the mean of means; $M = 3.33$; $SD = 0.589$ which is greater than 2.50 lies in the cut-off point of strongly agree. The results revealed that the teachers and examiners believed that the PBA items could stimulate students learning of mathematics.

All the nine items on the views of teachers and examiners on whether the PBA items could stimulate students' learning had means greater than the average mean of 2.50 meaning that the examiners and teachers agree to all the statements on the views of teachers and examiners on whether the PBA items could stimulate students' learning. Out of the nine items, the teachers and examiners expressed that four are major means by which PBA could stimulate students' learning; students are motivated to learn ($M = 3.57$, $SD = .516$), students are compelled to

learn ($M = 3.55$, $SD = .518$), the PBA encourages students to think differently on an issue ($M = 3.43$, $SD = .516$) and that the PBA encourages students to learn extensively ($M = 3.40$, $SD = .705$). This is evidenced by the means which are greater than the mean of means of 3.33.

Research Question Four

What are the catalytic effects of the developed PBA?

The descriptive statistics of the results on whether the PBA items could provide feedback that stimulate students learning of mathematics is presented in Table 11.

Table 11-*Descriptive Statistics of the Catalytic Effect of the PBA items in Mathematics by the Mathematics Teachers and Examiners (N = 390)*

S /N	Statement	Mean	S.D
1	Immediate feedback can be given to students	3.59	.513
2	Reveals areas of students' strength and weakness on each aspect of content learned	3.41	.492
3	Students will be able to reflect on their performance	3.47	.581
4	All domains of learning are assessed	3.21	.535
5	Makes learning individualistic	3.39	.607
6	Could be used in the classroom to give prompt feedback to students	3.46	.519
7	Measures diversity of behaviour	3.40	.632
Mean of Means		3.43	
Mean of Standard deviation			.562

Source: Field data (2020)

Table 11 shows the results of the views of teachers and examiners on whether the PBA items could provide feedback that stimulates students' learning. The results show that generally, the teachers and examiners strongly agree to the

statements views of teachers and examiners on whether the PBA items could provide feedback that stimulates students' learning. This is because the mean of means; $M = 3.43$; $SD = 0.562$ which is greater than 2.50 lies in the cut-off point of strongly agree. The results revealed that the teachers and examiners believed that the PBA items could provide feedback that stimulates students' learning.

All the eight items on the views of teachers and examiners on whether PBA item could provide feedback that stimulates students' learning had means greater than the average mean of 2.50 meaning that the examiners and teachers agree to all the statements that the PBA item could provide feedback that stimulates students' learning. Out of the eight items, the teachers and examiners expressed that four provide better feedback that stimulate students' learning: the PBA reveals students' true performance ($M = 3.46$, $SD = .615$), immediate feedback can be given to students ($M = 3.59$, $SD = .513$), students will be able to reflect on their performance ($M = 3.47$, $SD = .581$) and that the PBA could be used in the classroom to give prompt feedback to students ($M = 3.46$, $SD = .519$). This is evidenced by the means which are greater than the mean of means of 3.43.

Research Question Five

What is the reliability of the instrument?

The research question sought to evaluate the reliability of odPIM. the inter-rater reliability method was used to evaluate the reliability. The item was administered once so the test-retest and alternate form method could not be used. Again, because the items are the graded response types, methods for internal consistency were not applicable. Sixty (60) out of the 250 scripts for each rater

were randomly selected, making a total of 180 scripts. The scripts were exchanged for marking. The scores were used for estimating the inter-rater reliability. Because the scores were of the continuous type, the Pearson Product Moment correlation was used to estimate the reliability at the item level and the scale level. A coefficient of 0.70 and above shows the good reliability of results (Gwet, 2014). The result for reliability is presented in Table 12.

Table 12: *Pearson Product Moment Correlation for Inter-rater Reliability*

	Rater pairing		
	A/B	A/C	B/C
Item 1			
Pearson Correlation	0.978	0.895	0.941
Sig. (2-tailed)	0.000	0.000	0.000
Item 2			
Pearson Correlation	0.974	0.958	0.972
Sig. (2-tailed)	0.000	0.000	0.000
Item 3			
Pearson Correlation	0.969	0.972	0.958
Sig. (2-tailed)	0.000	0.000	0.000
Item 4			
Pearson Correlation	0.972	0.995	0.973
Sig. (2-tailed)	0.000	0.000	0.000
Item 5			
Pearson Correlation	0.994	0.992	0.991
Sig. (2-tailed)	0.000	0.000	0.000
Overall test			
Pearson Correlation	0.988	0.970	0.981
Sig. (2-tailed)	0.000	0.000	0.000

Source: Field Data (2020)

Table 12 shows the reliability coefficient of odPIM for each item and the entire instrument. The results show there is a significant high correlation among the raters for each item and the entire instrument. For the items, the coefficient ranges from 0.895 to 0.994. This is an indication that there is good inter-rater reliability for the items. At the scale level, the reliability coefficient ranges from 0.970 to 0.988 indicating a significant high reliability among the raters on the instrument.

Research Question six

What is the validity of the instrument?

The research question sought to evaluate the construct validity of the newly developed PBA items in terms of content validity and construct validity.

Content validity

For the content validity, participants were asked to indicate relevant/not-relevant for each of the five items. Number of agreed relevant was counted. The modified Kappa statistics was then applied to evaluate the content validity of the instrument. The modified Kappa statistics operate on three rules:

$$I - CVR = \frac{A}{N} \quad (1)$$

$$P_c = \left[\frac{N!}{A!} (N - A)! \right] \times 0.5^N \quad (2)$$

$$K = \frac{(I - CVR) - P_c}{1 - P_c} \quad (3)$$

Where N = total of number of experts responding (N = 35) and A = total number indicating relevant. For a sample of 35, kappa statistics of 0.31 and above indicate

acceptable level of content validity. The result of the content validity for item and the scale is presented in Table 13.

Table 13: *Modified Kappa Statistics for Content Validity Ratio for Item and the Scale*

Item	Number agreed relevant (A)	$I - CVR = \frac{A}{N}$	$P_c = \left[\frac{N!}{A!} (N - A)! \right] \times 0.5^N$	Modified Kappa statistic (K) $K = \frac{(I - CVR) - P_c}{1 - P_c}$
1	35	1.000	29.1038×10^{-12}	1.000
2	31	0.886	0.0008777529	0.886
3	33	0.943	0.00000006927	0.943
4	32	0.914	0.00000685744	0.914
5	30	0.857	0.13605169952	0.834
Average	32	0.914	0.00000685744	0.914

Source: Field Data (2020)

Table 13 shows the modified Kappa statistics of content validity. Comparing the computed K-statistics with the acceptable level of 0.31 for N = 35 stated by Lawshe (1975), it could be seen that the content validity ratio for the all the items is far above the acceptable level. The K-statistics of the items ranges from 0.834 to 1.00. The content validity for the entire instrument, which is the average of the K-statistic is 0.914. The result indicate there is a good content validity for odPIM, both for item and the scale.

Construct validity

For the construct validity, the convergent validity was evaluated to check the unidimensionality of the items. That is items should measure a common construct. The items are expected to relate on the construct. The odPIM was administered to the selected students and the scores on the item were used for the analysis. The unidimensionality means that every item should measure a single trait. To assess the convergent construct validity of the PBA, a principal component analysis of the residual by use of the SPSS was used. The KMO and Bartlett's Test was done to check for sampling adequacy. A significant result of $s = 0.544$, $p = 0.000$ showing adequate sampling was obtained (Appendix F). Bro and Smilde (2014) suggested that proportion of variance that the components explain, eigenvalues and the scree plot are combined to check the unidimensionality the test. Bro and Smilde (2014) suggested that the principal components that explain an acceptable level of variance between 80-90% or better should be retained. It is an indication that the components measure a single trait. For the Eigenvalues, only the principal components with eigenvalues that are greater than 1 should be retained. These components, in this case, items measure a single trait. On the Scree plot, the components in the steep curve before the first point that starts the line trend should be retained. The scores on Item 1 to Item 5 were analysed for unidimensionality (convergent construct validity). The results of the construct validity are presented in Figure 3 and Table 14.

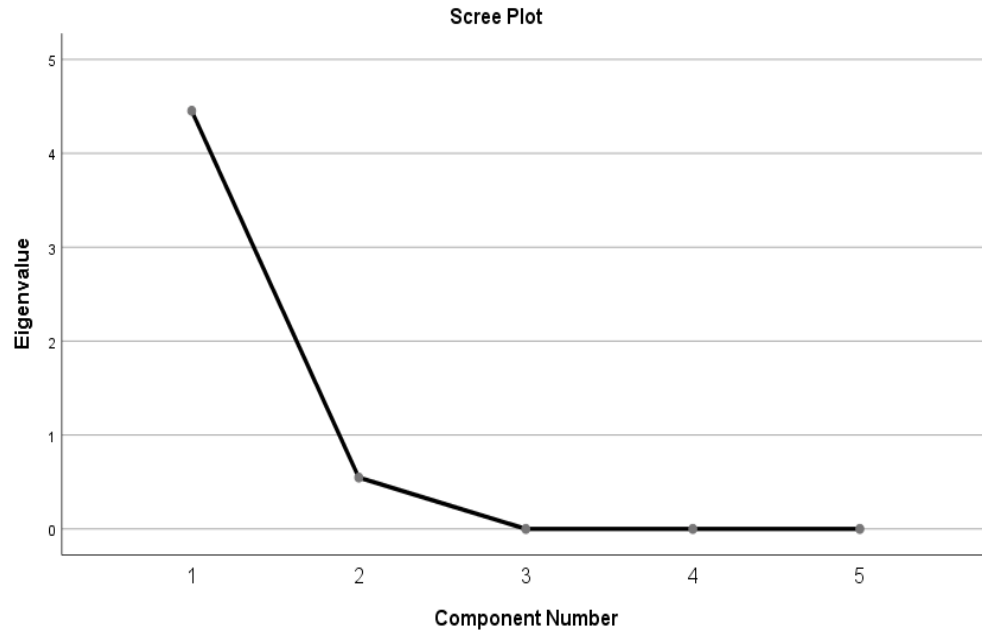


Figure 4-Scree Plot for the Items
Source: field data (2020)

In this result, there is only one principal components with an eigenvalue greater than 1. The scree plot shows that the eigenvalues do not form a straight line even at the first principal component (construct). This means that all the five items measure a single construct of mathematics. The total variance explained by the component is presented in Table 14.

Table 14-Eigenvalues of Total Variance Explained

Comp.	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cum. %	Total	% of Variance	Cum. %
1	4.453	89.058	89.06	4.453	89.058	89.058
2	.547	10.942	100.00			
3	9.890E-16	1.978E-14	100.00			
4	7.618E-16	1.524E-14	100.00			
5	-1.479E-16	-2.958E-15	100.00			

Source: Field Data (2020)

From Table 14, all the items constitute one principal component have eigenvalues greater than 1. The Items explain 89.06% of the variation in the data. The scree plot shows that the eigenvalues do not form a straight line even at the fifth principal component (item). The 89.06% is an adequate amount of variation explained in the data, hence all five items are considered to be measuring one construct (component). The result shows that the component of the test has an eigenvalue greater than 1. All the items measure a single trait. All the items fit the convergent construct validity of the instrument. The results of the loadings of the items are presented in Table 15.

Table 15-*Factor Loadings of the PBA Items*

Item	Theme	Loading
1	Rigid motion	0.973
2	Statistics	0.786
3	Mensuration	0.712
4	Geometric construction	0.699
5	Equations	0.704

Source: Field Data (2020)

The result also shows the all the five items have acceptable level of loading ranging from 0.973 to 0.699. This is indicating that the items have the acceptable level of being related to the dimension measured.

Analysis of Hypotheses

The hypotheses sought to find out if statistically significant difference exist in the feasibility, credibility, educational and catalytic of the newly developed PBA items for SHSs and how the instrument could reduce examination malpractice in SHSs among status (teachers and examiners), School Category

(Category A schools, Category B schools and Category C schools), gender (male and female) and experience. A four-way ANOVA was used to test the hypotheses at 95% confidence interval. The four-way ANOVA was used because, there were four independent variables with at least two sublevels each and one dependent variable being compared among the independent variables in each.

Checking assumptions for ANOVA

For the analysis of the hypotheses, the normality assumption and homogeneity of variance assumption were checked. The normality assumption checks whether the distribution of the scores is skewed or evenly distributed around the mean. For ANOVA analysis, there should be a normal distribution of the scores. The normality assumption was checked for the distribution of scores on the questionnaire for the validation of the developed PBA. The Q-Q plot was used to check the normality of the scores on the feasibility, credibility, educational and catalytic effects. The normality checks how close the scores are to the normality line on the Q-Q plot. The test of homogeneity of variance assumptions was checked with the Levene's test.

Hypothesis 1

The hypothesis sought to find out if the developed PBA had significant difference in feasibility among teachers and examiners due to status, School Category, gender and experience. Information to the hypothesis was provided by mathematics teachers and examiners selected for the study. A list of statements that measures feasibility and credibility were given to indicate their degree of agreement to the statements. Responses were added up. The normality assumption

was checked for each of the levels of the independent variables using the Shapiro-Wilk Test and the overall (total) using the Q-Q plot (Appendix G1).

The results show that with the exception of scores for Category A schools and 1-5yrs of experience, all the scores for all the levels of the independent variables were normally distributed. This is because the Shapiro-Wilk sig values are greater than 0.05. The overall normality of the scores of feasibility are presented in Appendix G1.

The distribution of scores on the feasibility of the developed PBA is normally distributed as shown the Q-Q plot. The homogeneity assumption was checked using the Levene Test (Appendix G1). The result shows the test of homogeneity of variance among gender, School Category, status and experience, $F(37, 352) = 3.964, p = 0.000$. The results of the Levene's test shows the variances of scores on the feasibility of the newly developed PBA items are assumed not equal. This is because the Levene's sig value of 0.070 is greater than 0.05. The descriptive statistics of the results of the four-way ANOVA is presented in Table 16.

Table 16- *Descriptive Statistics of the Results of the Feasibility of the PBA*

status	Sch. Cat.	Gender	Experience	Mean	Std. Deviation	N	
Examiner	Cat A	Male	1-5yrs	40.29	7.156	31	
			6-10yrs	40.04	4.666	23	
			11-15yrs	40.38	6.292	13	
		Female	1-5yrs	43.63	4.658	8	
			6-10yrs	41.00	2.449	5	
			1-5yrs	39.56	4.908	25	
	Cat B	Male	6-10yrs	46.18	1.328	11	
			11-15yrs	42.59	5.821	17	
			16-20yrs	49.00	.000	2	
		Female	>20yrs	40.00	6.753	6	
			1-5yrs	45.40	3.373	10	
			6-10yrs	52.00	NA	1	
	Cat C	Male	11-15yrs	51.00	.000	2	
			>20yrs	45.00	.894	6	
			1-5yrs	39.42	4.852	12	
		Female	6-10yrs	45.40	6.646	25	
			11-15yrs	44.29	3.891	14	
			16-20yrs	40.44	4.851	9	
	Teacher	Cat A	Male	>20yrs	42.18	5.419	11
				1-5yrs	38.00	NA	1
6-10yrs				47.80	6.017	5	
Female			16-20yrs	38.00	.000	2	
			>20yrs	46.00	NA	1	
			1-5yrs	39.91	5.117	22	
Cat B		Male	6-10yrs	46.00	.000	7	
			11-15yrs	42.00	6.226	14	
			1-5yrs	47.00	.000	7	
		Female	1-5yrs	40.16	8.092	19	
			6-10yrs	38.39	3.852	18	
			11-15yrs	38.25	6.551	4	
Cat C	Male	1-5yrs	40.25	4.500	4		
		6-10yrs	41.00	2.449	5		
		1-5yrs	42.71	12.880	7		
	Female	6-10yrs	43.40	7.099	15		
		11-15yrs	42.89	3.333	9		
		1-5yrs	38.00	NA	1		
			6-10yrs	48.50	5.836	10	
			16-20yrs	38.00	.000	8	

Source: Field data (2020)

Table 16 shows the descriptive statistics of the feasibility of the developed PBA among examiners and mathematics teachers due to School Category, gender and experience. The results show that for the male examiners in the Category A schools, the examiners with 11–15 years of experience expressed much feasibility of the developed PBA for use in schools and WAEC ($M = 40.38$, $SD = 6.292$, $N = 13$) and those with 6-10 yrs (40.04 , $SD = 4.666$, $N = 23$) experience expressed the least feasibility. For the female examiners in the Category A schools, examiners with 1-5 years of experience ($M = 43.63$, $SD = 4.658$, $N = 8$) and 6–10 years of experience (41.00 , $SD = 2.449$, $N = 5$) the examiners with 6–10 years of experience expressed that there is much feasibility of the developed PBA for use in schools and WAEC.

The Table also shows that for male examiners in the Category B schools, examiners with 1-5 years of experience ($M = 39.56$, $SD = 4.908$, $N = 25$) and 16–20 years ($M = 49.00$, $SD = .000$, $N = 2$) expressed the least and much feasibility of the developed PBA for use in SHS respectively. For female examiners in the Category B schools, examiners with 6–10 years of experience ($M = 52.00$, $N = 1$) and above 20 years ($M = 45.00$, $SD = .894$, $N = 6$) expressed the least and much feasibility of the developed PBA for use in schools and WAEC respectively

The Table also shows that for male examiners in the Category C schools, the examiners with 6–10 years ($M = 45.40$, $SD = 6.464$, $N = 25$) and 1–5 years ($M = 39.42$, $SD = 4.852$, $N = 12$) experience expressed much and least feasibility of the developed PBA for use in schools respectively. For female examiners in the Category C schools, examiners with 1-5 years of experience ($M = 38.00$, $N = 1$)

and 6-10 years ($M = 47.80$, $SD = 6.017$, $N = 5$) expressed the little and much feasibility of the developed PBA respectively.

The results of Table 16 again show that for the male mathematics teachers in the Category A schools, teachers with 1-5 years of experience ($M = 39.91$, $SD = 5.117$, $N = 22$) and 11–15 years ($M = 42.00$, $SD = 6.226$, $N = 14$), the teachers with 11–15 years of experience expressed the least and much feasibility of the developed PBA respectively. For the female teachers in the Category A schools, all were with 1-5 years of experience ($M = 47.00$, $SD = .000$, $N = 7$).

The Table also shows that for male teachers in the Category B schools, examiners with 1-5 years of experience ($M = 40.16$, $SD = 8.092$, $N = 19$) and 11–15 years ($M = 38.28$, $SD = 6.551$, $N = 4$), the teachers with 1–5 years of experience expressed much and least feasibility of the developed PBA for use in schools respectively. For female teachers in the Category B schools, teachers with 1-5 years of experience ($M = 42.25$, $SD = 4.500$, $N = 4$) and 6–10 years of experience ($M = 41.00$, $SD = 2.449$, $N = 4$), the examiners with 6–10 years of experience expressed much feasibility of the developed PBA for use in schools and WAEC.

The table also shows that for male teachers in the Category C schools the teachers with 6–10 years experience ($M = 43.40$, $SD = 7.099$, $N = 15$) expressed much feasibility of the developed PBA for use in schools with those with 11–15 years ($M = 42.71$, $SD = 12.880$, $N = 7$) expressing that there is little feasibility. For female teachers in the Category C schools, teachers with 1-5 years of experience ($M = 38.00$, $N = 1$) and 6–10 years of experience ($M = 48.50$, $SD =$

5.836, $N = 10$) the teachers with 6–10 years of experience expressed the least and much feasibility respectively of the developed PBA for use in schools. Table 17 shows whether the difference(s) in the means is/are significant.

Table 17- *Four-Way ANOVA Results on Feasibility of the PBA*

Source	Sum of Squares	df	Mean Square	F	Sig.
status	23.718	1	23.718	.754	.386
School Category	1.431	2	.716	.023	.978
Gender	213.747	1	213.747	6.793	.071
Experience	478.199	4	119.550	3.799	.050
status * School Category	568.401	2	284.201	9.032	.000
status * Gender	1.158	1	1.158	.037	.848
status * Experience	29.960	3	9.987	.317	.813
Sch Cat * Gender	53.699	2	26.850	.853	.427
Sch Cat * Experience	255.014	6	42.502	1.351	.234
Gender * Experience	87.126	4	21.781	.692	.598
status * Sch. Cat * Gender	135.001	2	67.500	2.145	.119
status * Sch Cat * Experience	464.365	4	116.091	3.690	.006
status * Gender * Experience	16.487	1	16.487	.524	.470
Sch Cat * Gender * Experience	48.013	3	16.004	.509	.677
Status* Sch Cat* Gender* Exp	2.702	1	2.702	.086	.770
Error	11075.582	352	31.465		
Total	703410.000	390			

Source: Field data, (2020)

The Table shows the ANOVA results of the feasibility of the developed PBA for use in schools and WAEC examinations among status, School Category, gender and experience. The results show that the main effect (status, School Category, gender and experience) had no significant difference in feasibility of the developed PBA, $F_{1, 352} = .754, p = .386$; $F_{1, 352} = .023, p = .978$; $F_{1, 352} = 6.793, p = .071$ and $F_{4, 352} = 3.799, p = .050$, respectively. The teachers and examiners of both gender in the Western Region with different experience levels expressed the same thing on feasibility of the developed PBA.

Again, with the exception of status * School Category , $F_{2, 352} = 9.032, p = .000$ and status * School Category * Experience, $F_{4, 352} = 3.690, p = .006$ interactions that showed significant difference in feasibility of the developed PBA, all the others, status * Gender, $F_{1, 352} = .037, p = .848$, status * Experience, $F_{3, 352} = .317, p = .813$, School Category * Experience, $F_{6, 352} = 1.351, p = .234$, School Category * Gender, $F_{2, 352} = .853, p = .427$, Gender * Experience, $F_{4, 352} = .692, p = .598$, status * School Category * Gender, $F_{2, 352} = 2.145, p = .119$, School Category * Gender * Experience, $F_{3, 352} = .509, p = .677$, status * Gender * Experience, $F_{1, 352} = .524, p = .470$ and status * School Category * Gender * Experience, $F_{1, 352} = .086, p = .770$ showed no significant differences in feasibility of the developed PBA.

This means that teachers and examiners at different locations and teachers and examiners at different categories of schools with different experience showed different feasibility of the developed PBA. However, examiners and teachers of a particular gender, examiners and teachers with a particular experience, male and female at a particular location, male and female examiners and teachers with a particular experience at a particular category of school showed the same feasibility of the developed PBA.

Hypothesis 2

The hypothesis sought to find out if the developed PBA had significant difference in credibility among teachers and examiners due to status, School Category, gender, and experience. The results of the normality assumption of scores within the levels of the independent variables and the overall scores on the

credibility of the PBA are presented in Appendix G2. The results show that with the exception of scores for teachers and 16-20 yrs of experience, all the scores for all the levels of the independent variables for credibility of results of the PBA were normally distributed. This is because the Shapiro-Wilk sig values are greater than 0.05. The overall normality of the scores of credibility is presented in Appendix G2. The distribution of scores on the credibility of the developed PBA is normally distributed as shown in the Q-Q plot.

The homogeneity assumption used check was using the Levene Test. The result is presented in Appendix G2. The result shows the test of homogeneity of variance of the scores of credibility among gender, School Category, status and experience, $F(37, 352) = 5.496, p = 0.000$. The results of the Levene test shows the variances of scores on the credibility of the newly developed PBA items is assumed not equal. This is because the Levene sig value of 0.000 is less than 0.05. Even though the assumption holds that variances should be assumed equal for parametric, the ANOVA is robust hence ANOVA could be performed.

The descriptive statistics of the results of the four-way ANOVA for the credibility of the PBA items for SHSs is presented in Table 18.

Table 18- *Descriptive Statistics of Credibility of the PBA*

status	Sch Cat	Gender	Experience	Mean	Std. Dev	N
Examiner	Cat A	Male	1-5yrs	22.84	1.951	31
			6-10yrs	24.39	1.924	23
			11-15yrs	21.77	3.032	13
		Female	1-5yrs	22.63	2.875	8
			6-10yrs	23.40	3.130	5
			11-15yrs	23.56	.917	25
	Cat B	Male	6-10yrs	24.73	.905	11
			11-15yrs	21.47	2.154	17
			16-20yrs	26.00	.000	2
		Female	above 20yrs	21.67	2.733	6
			1-5yrs	23.80	.422	10
			6-10yrs	28.00	NA	1
	Cat C	Male	11-15yrs	28.00	.000	2
			above 20yrs	22.33	.516	6
			1-5yrs	23.33	1.775	12
		Female	6-10yrs	24.60	2.723	25
			11-15yrs	24.07	1.859	14
			16-20yrs	22.11	2.205	9
			above 20yrs	21.91	1.973	11
			1-5yrs	25.00	NA	1
			6-10yrs	25.80	3.493	5
	16-20yrs	21.00	.000	2		
	above 20yrs	23.00	NA	1		

Table 17 cont'd.

status	Sch Cat	Gender	Experience	Mean	Std. Dev	N
Teacher	Cat A	Male	1-5yrs	23.64	.953	22
			6-10yrs	25.00	.000	7
			11-15yrs	21.00	2.075	14
	Cat B	Female	1-5yrs	24.00	.000	7
			6-10yrs	22.11	2.158	19
			11-15yrs	24.22	2.157	18
	Cat C	Male	1-5yrs	24.00	3.916	4
			6-10yrs	21.25	3.775	4
			11-15yrs	23.40	3.130	5
	Cat C	Female	1-5yrs	23.71	2.138	7
			6-10yrs	23.33	2.350	15
			11-15yrs	23.33	1.000	9
			16-20yrs	25.00	NA	1
				6-10yrs	26.90	2.601
			16-20yrs	21.00	.000	8

Source: Field data (2020)

Table 17 shows the descriptive statistics of the credibility of the developed PBA among examiners and mathematics teachers due to School Category (Category A schools, Category B schools and Category C schools), gender and experience. The results show that for the male examiners in the Category A schools, examiners with 1-5 years of experience ($M = 22.84$, $SD = 1.951$, $N = 31$),) and 11–15 years ($M = 21.77$, $SD = 3.032$, $N = 13$) had the highest and lowest means respectively on the credibility the developed PBA. For the female examiners in the Category A schools, examiners with 6–10 years of experience ($M = 23.40$, $SD = 3.130$, $N = 5$) had the highest mean on the credibility of the developed PBA.

The Table also shows that for male examiners in the Category B schools, examiners with 11–15 years ($M = 21.47$, $SD = 2.154$, $N = 17$) and 16–20 years (M

= 26.00, SD = .000, N = 2) had the lowest and highest means respectively on the credibility the developed PBA. For female examiners in the Category B schools, examiners with 6–10 years of experience (M=28.00, N = 1) and 11 – 15 years (M = 28.00, SD = .000, N = 2) had the highest and lowest means respectively on the credibility of the developed PBA.

The Table also shows that for male examiners in the Category C schools, examiners with 6–10years experience (M = 24.60, SD = 2.723, N = 25) above 20 years (M = 21.91, SD = 1.973, N = 11) had the highest and lowest means respectively on the credibility of the developed PBA. For female examiners in the Category C schools, examiners with 6–10 years of experience (M = 25.80, SD = 3.493, N = 5) and 16–20 years (M = 21.00, SD = .000, N = 2) had the highest and lowest mean respectively on the credibility of the developed PBA.

The results of Table 17 again show that for the male mathematics teachers in the Category A schools, teachers with 6–10 years of experience (M = 25.00, SD = .000, N = 7) and 11–15 years (M = 21.00, SD = 2.075, N =14), had the highest and lowest means respectively on the credibility of the developed PBA. For the female teachers in the Category A schools, all were with 1-5 years of experience (M = 24.00, SD = .000, N = 7).

The Table also shows that for male teachers in the Category B schools, examiners with 1-5 years of experience (M = 22.11, SD = 2.158, N = 19), and 11 – 15 years (M = 24.00, SD = 3.916, N = 4) had lowest and highest means respectively on the credibility of the developed PBA. For female teachers in the

Category B schools, teachers with 6–10 years of experience (23.40, SD = 3.130, N = 4) had the highest means on the credibility of the developed PBA.

The Table also shows that for male teachers in the Category C schools, teachers with 1-5 years of experience (M = 23.71, SD = 2.138, N = 7) and 6–10 years experience (M = 23.34, SD = 2.350, N = 15) had the highest and lowest means respectively on the credibility of the developed PBA. For female teachers in the Category C schools, teachers with 6–10 years of experience (M = 26.90, SD = 2.601, N = 10) and 16–20 years (M = 21.00, SD = .000, N = 8) had the highest and lowest means respectively on the credibility of the developed PBA. Table 19 shows whether the difference(s) in the means is/are significant.

Table 19- *Four-Way ANOVA results of Credibility of the Developed PBA*

Source	Sum Squares	Df	Mean Square	F	Sig.
status	3.334	1	3.334	.799	.372
School Category	25.040	2	12.520	2.999	.051
Gender	24.393	1	24.393	5.842	.160
Experience	190.438	4	47.610	11.403	.000
status * Sch Cat	19.277	2	9.639	2.309	.101
status * Gender	1.239	1	1.239	.297	.586
status * Experience	4.109	3	1.370	.328	.805
Sch Cat * Gender	17.910	2	8.955	2.145	.119
Sch Cat * Experience	100.963	6	16.827	4.030	.001
Gender * Experience	53.215	4	13.304	3.186	.014
status * Sch Cat * Gender	20.808	2	10.404	2.492	.084
status * Sch Cat * Experience	44.109	4	11.027	2.641	.034
status * Gender * Experience	.014	1	.014	.003	.953
Sch Cat * Gender * Experience	4.508	3	1.503	.360	.782
status * Sch Cat * Gender * Exp.	7.470	1	7.470	1.789	.182
Error	1469.695	352	4.175		
Total	214710.000	390			

Source: Field data (2020)

The Table shows the ANOVA results of credibility of the developed PBA among status, School Category, gender and experience. The results show that the status main effect is not significant in the credibility of the developed PBA, $F_{1, 352} = .799, p = .372$. The teachers and examiners expressed the same thing on credibility of the developed PBA. School Category main effect was also not significant, $F_{2, 352} = 2.999, p = .051$ indicating that generally, teachers and examiners in the Western Region expressed the same thing on credibility of the developed PBA. Gender main effect was also not significant, PBA, $F_{1, 352} = 5.843, p = .160$ indicating that generally, male and female teachers and examiners the same thing on credibility of the developed PBA. However, experience main effect was significant on the credibility of the developed $F_{4, 352} = 11.403, p = .000$ respectively. This means that there was difference in credibility of the developed PBA among respondents based on their experience.

Again, with the exception of status * School Category * Experience, $F_{4, 352} = 2.641, p = .034$, Gender * Experience, $F_{4, 352} = 3.186, p = .014$ and School Category * Experience, $F_{6, 352} = 4.030, p = .001$, interactions that were significant on the credibility of the developed PBA, all the others, status * School Category, $F_{2, 352} = 2.309, p = .101$, status * Gender, $F_{1, 352} = .297, p = .586$, status * Experience, $F_{3, 352} = .328, p = .805$, School Category * Gender, $F_{2, 352} = 2.145, p = .119$, status * School Category * Gender, $F_{2, 352} = 2.492, p = .084$, status * Gender * Experience, $F_{1, 352} = .003, p = .953$, status * School Category * Gender * Experience, $F_{1, 352} = 1.789, p = .182$, showed no significant effect on the credibility of the developed PBA.

This means that, respondents with a particular experience at particular category of school and respondents of a particular experience with a particular gender showed different credibility of the developed PBA. However, examiners and teachers of a particular gender, examiners and teachers with a particular experience, male and female at a particular location, male and female examiners and teachers with a particular experience at a particular category of school showed the same credibility of the developed PBA.

A post hoc test was performed for the significant experience main effects to ascertain the source of the significant difference due to experience. Because variances are not assumed equal, the Games-Howell test was used. The result of the post hoc test is presented in Table 20.

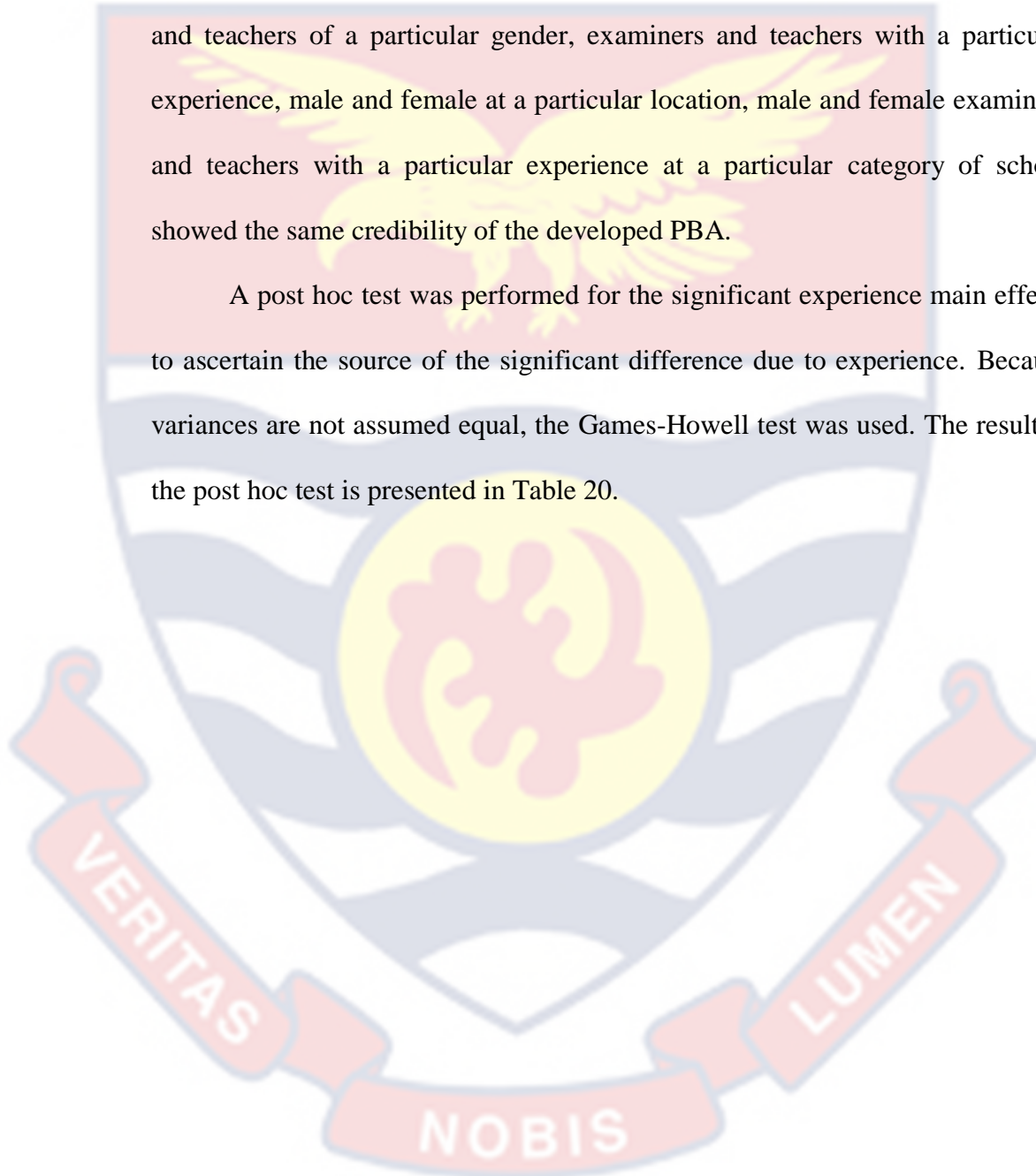


Table 20-Post Hoc Test of Scores of Credibility of Experience Main Effect

	(I) Experience	(J) Experience	Mean Diff (I-J)	Std. Error	Sig.
Games-Howell	1-5yrs	6-10yrs	-1.39*	.262	.000
		11-15yrs	.68	.344	.279
		16-20yrs	1.21	.463	.098
		above 20yrs	1.16	.408	.056
	6-10yrs	1-5yrs	1.39*	.262	.000
		11-15yrs	2.07*	.378	.000
		16-20yrs	2.60*	.490	.000
		above 20yrs	2.55*	.438	.000
	11-15yrs	1-5yrs	-.68	.344	.279
		6-10yrs	-2.07*	.378	.000
		16-20yrs	.53	.538	.863
		above 20yrs	.48	.491	.865
	16-20yrs	1-5yrs	-1.21	.463	.098
		6-10yrs	-2.60*	.490	.000
		11-15yrs	-.53	.538	.863
		above 20yrs	-.05	.581	1.000
	above 20yrs	1-5yrs	-1.16	.408	.056
		6-10yrs	-2.55*	.438	.000
		11-15yrs	-.48	.491	.865
			16-20yrs	.05	.581

Source: Field data (2020)

Table 19 shows the Post results of experience main effect of the scores of the credibility of the PBA. It was shown that there were significant differences in the multiple comparison of scores on the credibility of the instrument due to experience level of both teachers and examiners (sig values less than 0.05). The significant differences were between 1-5 years and 6-10 years with a mean of 1.39 against 1-5 years (6 -10 years expressed high credibility of the instrument than 1-5 years) and between 6 -10 years and 11-15 years with a mean of 2.07 in favour of 6-10 years (6-10 years expressed high credibility of the instrument than 11-15 years). It was also observed between 6-10 years and 16-20 years with

a mean of 2.60 in favour of 6-10 years (6-10 years expressed high credibility of the instrument than 16-20 years) and between 6-10 years and above 20 years with a mean of 2.55 in favour of 6-10 years (6-10 years expressed high credibility of the instrument than those above 20 years)

Hypothesis 3

The hypothesis sought to find out if statistically significant differences exist in the educational effect of the PBA items among teachers and examiners due to status, School Category, gender and experience. Information to the hypothesis was provided by mathematics teachers and examiners selected for the study. A list of statements that measure educational and catalytic effects was given to indicate their degree of agreement to the statements. Responses were added up. The normality assumption was checked for each of the levels of the independent variables using the Shapiro-Wilk Test and the overall (total) using the Q-Q plot for educational effect of the PBA. The result of the Shapiro test is presented in Appendix G3. The results show that all the scores for all the levels of the independent variables for educational effect of the results of the newly developed PBA were normally distributed. This is because the Shapiro-Wilk sig values are greater than 0.05. The overall normality of the scores of educational effect of the newly developed PBA is presented in Appendix G3. The distribution of scores on the educational effect of developed PBA is normally distributed with few skewed scores as shown the Q-Q plot.

The result of the homogeneity of variance of score on the educational effect of PBA is presented in Appendix G3. The result shows the test of

homogeneity of variance of the scores of educational effect among gender, School Category, status and experience, $F(37, 352) = 5.742, p = 0.673$. The results of the Levene test shows the variances of scores on the educational effect of the newly developed PBA items is assumed equal. This is because the Levene sig value of 0.673 is greater than 0.05. This means that the variances of scores of each group are equal. The assumptions for ANOVA have been met. The descriptive statistics of the results of the four-way ANOVA are presented in Table 21.

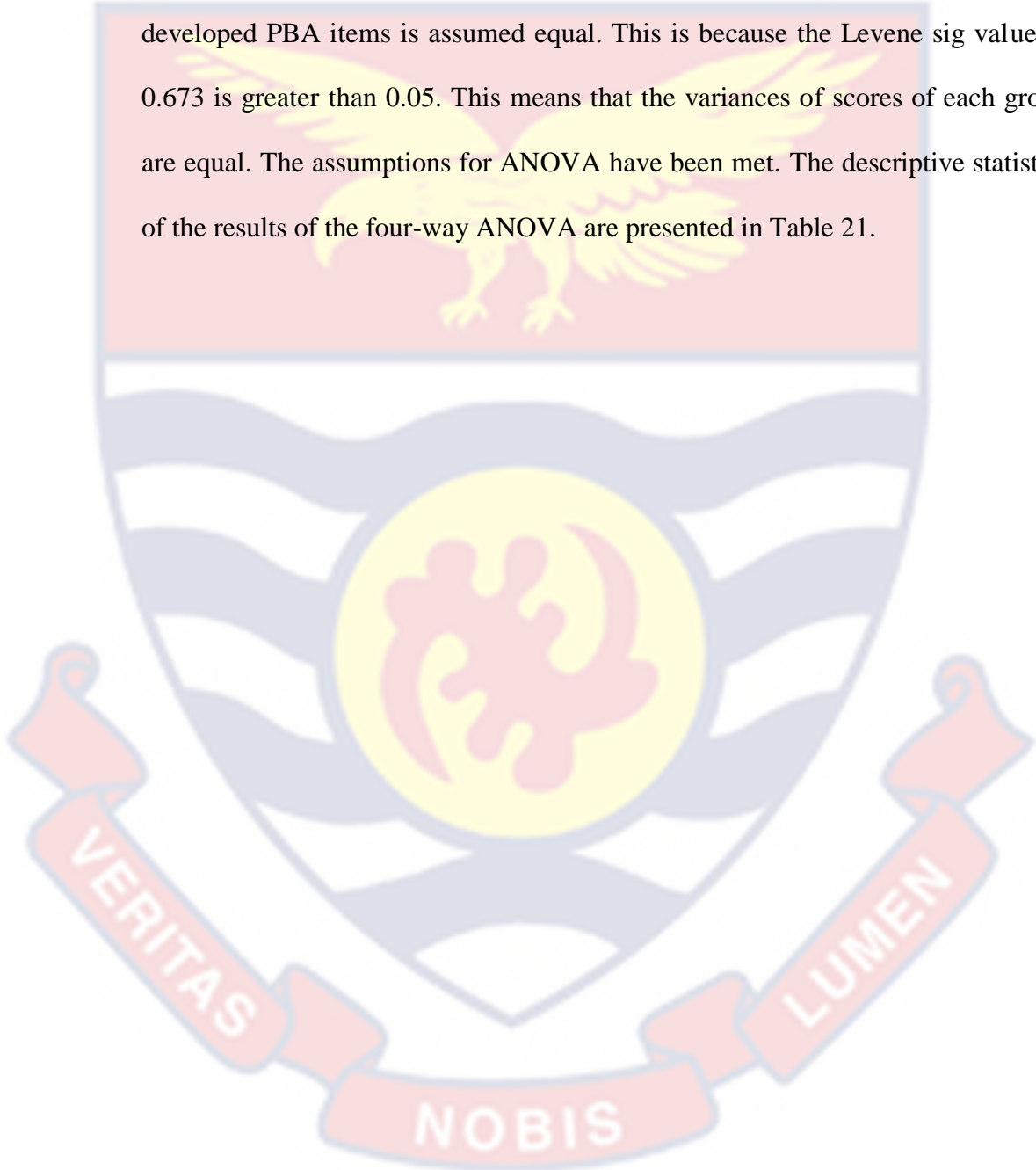


Table 21-Descriptive Statistics of the Results of Educational Effect of the PBA

Status	Sch Cat	Gender	Experience	Mean	Std. Dev	N
Examiner	Cat A	Male	1-5yrs	28.52	2.096	31
			6-10yrs	29.13	2.801	23
			11-15yrs	32.08	.641	13
		Female	1-5yrs	31.13	3.399	8
			6-10yrs	29.00	2.828	5
	Cat B	Male	1-5yrs	28.88	.833	25
			6-10yrs	31.36	1.433	11
			11-15yrs	31.76	.664	17
			16-20yrs	34.00	.000	2
			above 20yrs	26.67	4.926	6
		Female	1-5yrs	30.60	.843	10
			6-10yrs	36.00	NA	1
			11-15yrs	36.00	.000	2
			above 20yrs	31.33	2.066	6
			1-5yrs	28.50	2.844	12
Cat C	Male	6-10yrs	30.92	3.673	25	
		11-15yrs	31.14	2.316	14	
		16-20yrs	29.33	2.646	9	
	female	above 20yrs	28.55	4.251	11	
		1-5yrs	24.00	NA	1	
		6-10yrs	33.20	4.087	5	
Teacher	Cat C	female	16-20yrs	28.00	.000	2
			above 20yrs	34.00	NA	1
			1-5yrs	28.95	.844	22
		Male	6-10yrs	31.00	.000	7
			11-15yrs	32.00	.000	14
	Cat B	Female	1-5yrs	31.00	.000	7
			1-5yrs	28.00	2.582	19
		Male	6-10yrs	28.61	2.973	18
			11-15yrs	32.25	1.258	4
			1-5yrs	31.25	5.188	4
	Cat A	Female	6-10yrs	29.00	2.828	5
			1-5yrs	28.43	3.552	7
			1-5yrs	28.80	2.678	15
		Male	11-15yrs	30.44	1.333	9
			1-5yrs	24.00	NA	1
6-10yrs			34.30	3.093	10	
Cat C	Female	16-20yrs	28.00	.000	8	

Source: Field data (2020)

Table 21 shows the descriptive statistics of impact of the developed PBA on students learning among examiners and mathematics teachers due to School Category (Category A schools, Category B schools and Category C schools), gender and experience. The results show that for the male examiners in the Category A schools, examiners with 1-5 years of experience ($M = 28.52$, $SD = 2.096$, $N = 31$ and 11–15 years ($M = 32.08$, $SD = .641$, $N = 13$) had the lowest and highest means respectively on the educational effect of the developed PBA. For the female examiners in the Category A schools, examiners with 1-5 years of experience ($M = 31.13$, $SD = 3.399$, $N = 8$) had the highest means respectively on the educational effect of the developed PBA.

The Table also shows that for male examiners in the Category B schools, examiners with 1-5 years of experience ($M = 28.88$, $SD = .833$, $N = 25$) and 16–20 years ($M = 34.00$, $SD = .000$, $N = 2$) had the lowest and highest means respectively on the educational effect of the developed PBA. For female examiners in the Category B schools, examiners with 1-5 years of experience ($M = 30.60$, $SD = .843$, $N = 10$) and 6–10 years experience ($M = 36.00$, $N = 1$) and 11–15 years ($M = 36.00$, $SD = .000$, $N = 2$) had the lowest and highest means respectively on the educational effect of the developed PBA.

The Table also shows that for male examiners in the Category C schools, examiners with 1-5 years of experience ($M = 28.50$, $SD = 2.844$, $N = 12$) and 11–15 years ($M = 31.14$, $SD = 2.316$, $N = 14$) had the lowest and highest means respectively on the educational effect of the developed PBA. For female examiners in the Category C schools, examiners with 1-5 years of experience (M

= 24.00, N = 1) and above 20 years (M = 34.00, N = 1) had the lowest and highest means respectively on the educational effect of the developed PBA.

The results of Table 21 again show that for the male mathematics teachers in the Category A schools, teachers with 1-5 years of experience (M = 28.95, SD = .844, N = 22) and 11–15 years (M = 32.00, SD = .000, N = 14) had the lowest and highest means respectively on the educational effect of the developed PBA. For the female teachers in the Category A, all were with 1-5 years of experience (M = 31.00, SD = .000, N = 7).

The Table also shows that for male teachers in the Category B schools, examiners with 1-5 years of experience (M = 28.00, SD = 2.582, N = 19), and 11–15 years (M = 32.25, SD = 1.258, N = 4) had the lowest and highest means respectively on the educational effect of the developed PBA. For female teachers in the Category B schools, teachers with 6–10 years of experience (M = 29.00, SD = 2.828, N = 4), experience expressed educational effect of the newly developed PBA.

The Table also shows that for male teachers in the Category C school, teachers with 1-5 years of experience (M = 28.43, SD = 3.552, N = 7) and 11–15 years (M = 30.44, SD = 1.333, N = 9) had the lowest and highest mean respectively on the educational effect of the developed PBA. For female teachers in the Category C schools, teachers with 1-5 years of experience (M = 24.00, N = 1) and 6–10 years of experience (M = 34.30, SD = 3.093, N = 10) had the lowest and highest means respectively on the educational effect of the developed PBA. Table 22 shows whether the difference(s) in the means is/are significant.

Table 22- *Four-Way ANOVA results of the Educational Effect of the PBA*

Source	Sum of Squares	Df	Mean Square	F	Sig.
Status	12.349	1	12.349	2.089	.149
Sch Cat	24.941	2	12.471	2.110	.123
Gender	91.223	1	91.223	15.432	.000
Experience	294.298	4	73.574	12.447	.000
status * Sch Cat	27.485	2	13.742	2.325	.099
status * Gender	.135	1	.135	.023	.880
status * Experience	15.409	3	5.136	.869	.457
Sch Cat * Gender	10.907	2	5.453	.923	.398
Sch Cat * Experience	192.928	6	32.155	5.440	.000
Gender * Experience	69.684	4	17.421	2.947	.020
status * Sch Cat * Gender	8.193	2	4.097	.693	.501
status * Sch Cat * Experience	61.285	4	15.321	2.592	.036
status * Gender * Experience	1.563	1	1.563	.264	.607
Sch Cat*Gender * Experience	117.925	3	39.308	6.650	.000
status*Sch Cat* Gender * Exp	18.018	1	18.018	3.048	.082
Error	2080.710	352	5.911		
Total	352618.000	390			

Source: Field data (2020)

The Table shows the ANOVA results of impact of the developed PBA on students learning among status, School Category, gender and experience. The results show that the status main effect was not significant on the educational effect, $F_{1, 352} = 2.089$, $p = .149$. The teachers and examiners expressed the same thing on impact of the developed PBA on students learning. School Category main effect was also not significant, $F_{2, 352} = 2.110$, $p = .123$ indicating that generally, respondents irrespective of category of school expressed the same impact of the developed PBA on students' learning. However, gender and experience main effects were significant on the impact of the developed PBA on students learning, $F_{1, 352} = 15.432$, $p = .000$ and $F_{4, 352} = 12.447$, $p = .000$ respectively. This means that, there were differences in impact of the developed

PBA on students' learning among respondents on the bases of their gender and experience.

Again, School Category * Experience, $F_{6, 352} = 5.440, p = .000$, Gender * Experience, $F_{4, 352} = 2.947, p = .020$, School Category * Gender * Experience, $F_{3,352} = 6.650, p = .000$ and status * School Category * Experience, $F_{4, 352} = 2.592, p = .036$ interactions showed significant effects on impact of the developed PBA on students' learning. This means that respondents at different locations with different experience, respondents of a particular gender with a particular experience at particular category of school showed difference in impact of the developed PBA on students' learning. However, status * School Category, $F_{2,352} = 2.325, p = .099$, status * Gender, $F_{1,352} = .023, p = .880$, status * Experience, $F_{3,352} = .869, p = .457$, School Category * Gender, $F_{2,352} = .923, p = .398$, status * School Category * Gender, $F_{2,352} = .693, p = .501$, status * Gender * Experience, $F_{1,352} = .264, p = .607$ and status * School Category * Gender * Experience, $F_{1,352} = 3.048, p = .082$ interactions showed no significant effects on impact of the developed PBA on students' learning. This means that, examiners and teachers of a particular gender, examiners and teachers with a particular experience, male and female at a particular location, male and female examiners and teachers with a particular experience at a particular category of school showed the same impact of the developed PBA on students' learning. Only the post hoc for experience main effect performed because gender has less than three levels hence post hoc cannot be performed. Because variances were

assumed equal, the Tukey HSD test was used for the post hoc analysis. The result of post hoc of education effect due to experience is presented in Table 23.

Table 23- *Post Hoc Test of Scores Experience Main Effect of educational Effect of the PBA*

	(I) Experience	(J) Experience	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	1-5yrs	6-10yrs	-1.31*	.333	.001
		11-15yrs	-2.74*	.392	.000
		16-20yrs	-.16	.638	.999
		above 20yrs	-.01	.602	1.000
	6-10yrs	1-5yrs	1.31*	.333	.001
		11-15yrs	-1.43*	.403	.004
		16-20yrs	1.15	.645	.382
		above 20yrs	1.30	.609	.211
	11-15yrs	1-5yrs	2.74*	.392	.000
		6-10yrs	1.43*	.403	.004
		16-20yrs	2.58*	.677	.001
		above 20yrs	2.73*	.643	.000
	16-20yrs	1-5yrs	.16	.638	.999
		6-10yrs	-1.15	.645	.382
		11-15yrs	-2.58*	.677	.001
		above 20yrs	.14	.817	1.000
	above 20yrs	1-5yrs	.01	.602	1.000
		6-10yrs	-1.30	.609	.211
		11-15yrs	-2.73*	.643	.000
			16-20yrs	-.14	.817

Source: Field data (2020)

Table 23 shows the post hoc test on the experience main effect of the scores on the educational effect of the PBA. It was shown that, there were significant differences in the multiple comparison of scores on the educational effect of the instrument due to experience level of both teachers and examiners (sig values less than 0.05). The significant differences were between 1-5 years and 6-10 years with a mean of 1.31 against 1-5 years (6-10 years expressed high educational effect of the instrument than 1-5 years), between 1-5 years and 11-

15 years with a mean of 2.74 against 1-5 years (11-15 years expressed high educational effect of the instrument than 1-5 years) and between 6-10 years and 11-15 years with a mean of 1.43 against 6-10 years (11-15 years expressed high educational effect of the instrument than 6-10yrs). It was also observed between 11-15 years and 16-20 years with a mean of 2.58 in favour of 11-15 years (11-15 years expressed high educational effect of the instrument than 16-20 years) and between 11-15 years and above 20 years with a mean 2.73 in favour of 11-15 years (11-15 years expressed high educational effect of the instrument than above 20 years).

Hypothesis 4

The hypothesis sought to find out if significant differences exist in the educational and catalytic effects of the PBA items among teachers and examiners due to School Category, gender and experience. The normality assumption was checked for each of the levels of the independent variables using the Shapiro-Wilk Test and the overall (total) using the Q-Q plot. The result of the Shapiro test is presented in Appendix G4. The results show that with the exception of scores for 16-20yrs of experience, all the scores for all the levels of the independent variables for catalytic effect of the newly developed PBA were normally distributed. This is because the Shapiro-Wilk sig values are greater than 0.05. The overall normality of the scores of catalytic effects is presented in Appendix G4. The plot shows that the scores on the ability of the PBA to provide immediate feedback of students' learning are normally distributed with few skewed ones.

The result of homogeneity of variance with the Levene test is presented in Appendix G4. The result shows the test of homogeneity of variance of the scores of catalytic effect among gender, School Category, status and experience, $F(37, 352) = 4.541, p = 0.213$. The results of the Levene's test show that the variances of scores on the Catalytic effect of the newly developed PBA items are assumed equal. This is because the Levene sig value of 0.213 is greater than 0.05.

The descriptive statistics of the results of the four-way ANOVA for the catalytic effect of the newly developed PBA for SHSs are presented in Table 24.

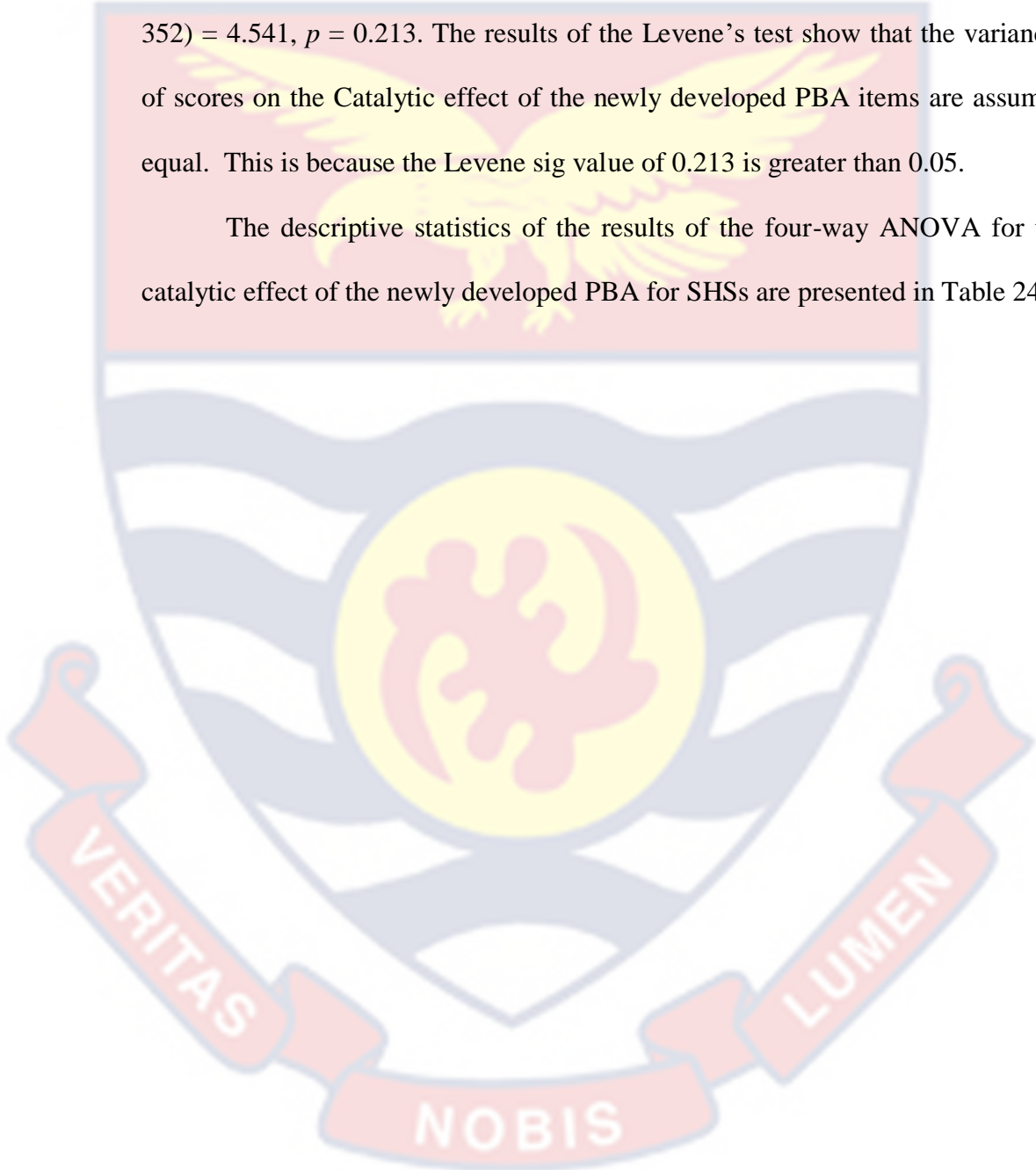


Table 24-Descriptive Statistics of the Results of Catalytic Effect

Status	Sch Cat	Gender	Experience	Mean	Std. Dev	N
Examiner	Cat A	Male	1-5yrs	26.45	3.139	31
			6-10yrs	28.87	2.437	23
			11-15yrs	27.00	3.000	13
		Female	1-5yrs	29.00	1.414	8
			6-10yrs	26.20	2.280	5
			11-15yrs	25.92	2.900	25
	Cat B	Male	6-10yrs	28.55	1.508	11
			11-15yrs	26.59	2.425	17
			16-20yrs	28.00	.000	2
		Female	>20yrs	28.33	3.141	6
			1-5yrs	29.60	.843	10
			6-10yrs	32.00	NA	1
	Cat C	Male	11-15yrs	32.00	.000	2
			>20yrs	26.67	1.862	6
			1-5yrs	28.25	2.864	12
Female		6-10yrs	28.04	3.458	25	
		11-15yrs	27.14	2.316	14	
		16-20yrs	24.89	1.764	9	
Cat C		Male	>20yrs	27.36	2.693	11
			1-5yrs	28.00	NA	1
			6-10yrs	29.00	4.123	5
	Female	16-20yrs	24.00	.000	2	
		>20yrs	29.00	NA	1	
		1-5yrs	26.18	2.954	22	
Teacher	Cat A	Male	6-10yrs	29.00	.000	7
			11-15yrs	26.50	2.594	14
		Female	1-5yrs	30.00	.000	7
			1-5yrs	26.21	3.242	19
	Cat B	Male	6-10yrs	28.83	2.771	18
			11-15yrs	28.75	3.403	4
		Female	1-5yrs	28.00	1.414	4
			6-10yrs	26.20	2.280	5
	Cat C	Male	1-5yrs	29.43	2.299	7
			6-10yrs	26.67	3.352	15
		Female	11-15yrs	26.44	1.333	9
			1-5yrs	28.00	NA	1
	Female	6-10yrs	30.30	3.129	10	
		16-20yrs	24.00	.000	8	

Source: Field data (2020)

Table 24 shows the descriptive statistics of feedback provided by the developed PBA to stimulate students' learning among examiners and mathematics

teachers due to School Category (Category A, Category B and Category C), gender and experience. The results show that for the male examiners in the Category A schools, examiners with 1-5 years of experience ($M = 26.45$, $SD = 3.139$, $N = 31$) and 6–10 years of experience ($M = 28.87$, $SD = 2.437$, $N = 23$) had the lowest and highest means respectively on the catalytic effect of the developed PBA. For the female examiners in the Category A schools, examiners with 1-5 year of experience ($M = 29.00$, $SD = 1.414$, $N = 8$) had the highest mean on the catalytic effect of the developed PBA.

The Table also shows that for male examiners in the Category B schools, examiners with 1-5 years of experience ($M = 25.92$, $SD = 2.900$, $N = 25$ and above 20 years ($M = 28.33$, $SD = 3.141$, $N = 6$), had lowest and highest means respectively on the catalytic effect of the developed PBA. For female examiners in the Category B schools, examiners with 6–10 years of experience ($M = 33.00$, $N = 1$) and above 20 years ($M = 26.67$, $SD = 1.862$, $N = 6$) had highest and lowest means respectively on the catalytic effect of the developed PBA.

The Table also shows that for male examiners in the Category C schools, examiners with 6–10 years of experience ($M = 28.04$, $SD = 3.458$, $N = 25$) and 16–20 years ($M = 24.89$, $SD = 1.764$, $N = 9$) had the highest and lowest mean respectively on the catalytic effect of the developed PBA. For female examiners in the Category C schools, examiners with 6–10 years of experience ($M = 30.00$, $SD = 4.123$, $N = 5$) and 16 – 20 years ($M = 24.00$, $SD = .000$, $N = 2$) had the highest and lowest means respectively on the catalytic effect of the developed PBA.

The results of Table 24 again show that for the male mathematics teachers in the Category A schools, teachers with 1-5 years of experience ($M = 26.18$, $SD = 2.954$, $N = 22$) and 11–15 years ($M = 26.50$, $SD = 2.594$, $N = 14$), had the lowest and highest means on the catalytic effect of the developed PBA. For the female teachers in the Category A, all were with 1-5 year of experience ($M = 30.00$, $SD = .000$, $N = 7$).

The Table also shows that for male teachers in the Category B schools, examiners with 1-5 year of experience ($M = 26.21$, $SD = 3.242$, $N = 19$) and 11–15 years ($M = 28.75$, $SD = 3.403$, $N = 4$), had the lowest and highest means respectively on the catalytic effect of the developed PBA. For female teachers in the Category B schools, teachers with 1-5 years of experience ($M = 28.00$, $SD = 1.414$, $N = 4$) the highest means on the catalytic effect of the developed PBA.

The Table also shows that for male teachers in the Category C schools, teachers with 1-5 years of experience ($M = 29.43$, $SD = 2.299$, $N = 7$) and 11–15 years ($M = 26.44$, $SD = 1.333$, $N = 9$) had the highest and lowest means respectively on the catalytic effect of the developed PBA. For female teachers in the Category C schools, teachers with 6–10 years of experience ($M = 30.30$, $SD = 3.129$, $N = 10$) and 16–20 years ($M = 24.00$, $SD = .000$, $N = 8$) had the highest and lowest means respectively on catalytic effect of the developed PBA. Table 25 shows whether the difference(s) in the means is/are significant.

Table 25- *Four-Way ANOVA Results of Catalytic Effect of the PBA*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Status	3.444	1	3.444	.482	.488
Sch Cat	5.540	2	2.770	.387	.679
Gender	21.647	1	21.647	3.028	.083
Experience	104.712	4	26.178	3.662	.060
status * Sch Cat	9.022	2	4.511	.631	.533
status * Gender	4.715	1	4.715	.659	.417
status * Experience	4.541	3	1.514	.212	.888
Sch Cat * Gender	1.959	2	.980	.137	.872
Sch Cat * Experience	34.490	6	5.748	.804	.567
Gender * Experience	34.830	4	8.708	1.218	.303
status * Sch Cat * Gender	46.983	2	23.491	3.286	.039
status * Sch Cat * Experience	18.894	4	4.724	.661	.620
status * Gender * Experience	.027	1	.027	.004	.951
Sch Cat * Gender * Experience	66.687	3	22.229	3.109	.027
status * Sch Cat * Gender * Exp.	14.652	1	14.652	2.049	.153
Error	2516.547	352	7.149		
Total	296203.000	390			

Source: Field Data (2020)

The Table shows the ANOVA results of the feedback provided by the developed PBA to stimulate students' learning among status, School Category, gender and experience. The results show that all (status, School Category, gender and experience) main effects were not significant on the feedback provided by the PBA to stimulate students' learning, $F_{1, 352} = .482$, $p = .488$, $F_{2, 352} = .679$, $F = .679$, $F_{1, 352} = 3.028$, $p = .083$ and $F_{4, 352} = 3.662$, $p = .060$ respectively. This means that generally, there were no significant differences in

the feedback provided by the PBA to stimulate students' learning as expressed by male and female respondents, the teachers and the examiners as well respondents in the Western Region.

Again, with the exception of status * School Category * Gender, $F_{2, 352} = 3.286, p = .039$ and School Category * Gender * Experience, $F_{3, 352} = 3.109, p = .027$ interactions that showed significant effects on feedback provided by PBA to stimulate students' learning, all the others, status * Gender, status * School Category, $F_{2, 352} = .631, p = .533$, status * Gender, $F_{1, 352} = .659, p = .417$, status * Experience, $F_{3, 352} = .212, p = .888$, School Category * Gender, $F_{2, 352} = .137, p = .872$, School Category * Experience, $F_{6, 352} = .804, p = .567$, Gender * Experience, $F_{4, 352} = 1.218, p = .303$, status * School Category * Experience, $F_{4, 352} = .661, p = .620$, status * Gender * Experience, $F_{1, 352} = .004, p = .951$ and status * School Category * Gender * Experience, $F_{1, 352} = 2.049, p = .153$ showed no significant effects in feedback provided by the PBA to stimulate students' learning.

This means that, respondents of a particular gender at a particular category of school and respondents with a particular experience at particular category of school and a particular gender showed different feedback provided by the PBA to stimulate students' learning. However, examiners and teachers of a particular gender, examiners and teachers with a particular experience, male and female at a particular location, male and female examiners and teachers with a particular experience at a particular category of school showed no difference in the feedback provided by the PBA to stimulate students' learning.

Discussions of key Findings

Feasibility of the developed PBA

The study found that the PBA is feasible to be used for SHSs and scripts as in the traditional system could be used in the marking, construction of alternate forms is feasible, item constructions will not require much extra time and skills and that representative content could be covered and learned for a single test. It was further found that the use of PBA would not produce extra cost to the assessment system and that it is practicable for a large number of examinees. It was also found that the status main effect had no significant difference in feasibility of the developed PBA. That is the teachers and examiners expressed the same thing on feasibility of the developed PBA. School Category main effect also showed no significant difference, indicating that generally, respondents in the Western Region expressed the same thing on feasibility of the developed PBA. However, gender and experience main effects showed significant effects in concepts that attract examination malpractice. This means that there were differences in feasibility of the developed PBA among respondents on the bases of their gender and experience.

Contrary to the finding of this study is that of Uzun, Aktaş, Aşiret & Yorulmaz (2018) which found out that, for PBAs there is a high rate of error in generalizing over tasks irrespective of how well the tasks are designed. The controversy of this study and that of Teker (2019) could be attributed to the difference in the form of PBA used. While this study was limited to on-demand PBA, the study of Uzun, Aktaş, Aşiret & Yorulmaz (2018) was limited to extended performance which Brennan (2006) has alluded that it is difficult to

generalize performance over tasks. This is because the tasks differ in every instance. Janssen, Meier and Trace (2014) thus stated that PBA is a careful specification of the task.

Suuramm et al., (2016) admitted that the challenges of PBA but could not rule out its usability in the classroom. They stated that PBAs can be challenging to the changes in general teaching paradigms. However, specific behaviours and procedures in the classroom could be changed with PBA under some circumstances. This is in affirmation to the findings of this study, though the developed on-demand PBA does come with some challenges, it is feasible for use in schools and for WAEC examinations.

Credibility of the developed PBA

The study found that the PBA results reflect students' true performance, malpractice associated with examination is reduced and that differences in students' performance become real. The study also showed that the status main effect had no significant difference in credibility of the developed PBA. That is the teachers and examiners expressed the same thing on credibility of the developed PBA. School Category main effect also showed no significant effect, indicating that generally, male and female expressed the same thing on credibility of the developed PBA. However, gender and experience main effects showed significant effect in credibility of the developed PBA. This means that there was a difference in credibility of the developed PBA among respondents on the basis of gender and experience.

In line with the findings of this study, Wiggins and McTighe (2015) did not mince words as they asserted that, in fact, authentic assessments go beyond

just responding to a test item. Authentic assessment for that matter, PBA teaches both students and teachers what is meant by “doing of a subject” and essential performances of a profession. Performance assessment in mathematics is about what is supposed and expected to be demonstrated as having learnt mathematics. As a principle, every classroom assessment is to serve the needs of the learners (Asamoah-Gyimah & Anane, 2018). This means that assessment should help the students learn how to apply every learned concept to solve real life problems. This study has found that developed PBA could help students learn better as held by Asamoah-Gyimah & Anane (2018). Falk, Ort and Moirs (2007) and Shepard (2009) also opined that with a well-designed measurement tool in the likes of a scoring rubric, PBA can explain how and the why a student might be struggling in learning mathematics. In effect, PBA can actually help teachers to find out how best their students can learn. This means that PBA brings out individual differences which are reflected in their performance. In this, lies the spirit of validity, in that, validity of assessment results seeks to establish a sound bases of comparison of students’ performance (Nitko, 2004).

Also, Darling-Hammond (2009) stated that PBA enables differentiatonal assessment to take place. All students, including exceptional students, have all opportunity to demonstrate their understanding of what is learned. This is, an indication that PBA seeks to focus on individualised learning as the real-life situation differs from student to student. This reveals individual standings on every assessment. Stone and Lane (2006) affirmed that opinion of Darling-Hammond, (2009) by stating that PBA has multiple correct procedures to a task

and therefore has multiple correct answers. This characteristic tends to reduce copying from colleagues or teachers copying answers to students since they cannot have the procedures written for each student. Also, performance assessment requires students to perform the tasks which cannot be done by a third party. Some of the performance assessment tasks are limited to an individual student, therefore leakages and copying and their sources could easily be detected. Students are required to report on the procedures that were used in completing the task. This means that the findings of this study support literature that PBA reveals individual true performance.

Educational effect of the PBA

This study found that the developed PBA could stimulate students' learning; students are motivated to learn, students are compelled to learn, PBA encourages students to think differently on issues and that PBA encourages students to learn extensively. It was further found that the status main effect had no significant effect in impact of PBA on students' learning. That is the teachers and examiners expressed the same thing on impact of the developed PBA on students learning. School Category main effect also showed no significant effect, indicating that generally, respondents irrespective of category of school expressed the same thing on the impact of the developed PBA on students' learning. However, gender and experience main effects showed significant effects in impact of the developed PBA on students learning. This means that there were differences in impact of the developed PBA on students' learning among respondents on the basis of their gender and experience.

The finding is consistent with literature as Stone and Lane (2006) performance assessment has multiple correct procedures to a task and therefore has multiple correct responses. This characteristic tends to reduce copying from colleagues or teachers copying answers to students since they cannot have the procedures written for each student. Also, performance assessment requires students to perform the tasks which cannot be done by a third party. Some of the performance assessment tasks are limited to an individual student. Students are required to report on the procedures that were used in completing the task. These characteristics of PBA compel students to learn since it would be difficult for colleagues or any third party to assist them during examinations. The idea that collusion is impossible is a motivating factor for students to do their best by making sure they perform any task given them on their own whiles in classroom before the time of examination. This was well said by Nitko (2004) that PBA as a form of assessment presents a “hand on task” which requires students to perform an activity which requires them to apply the knowledge and skills acquired from different learning experiences. It allows students to show how well they have learnt. Simply put, a PBA is an assessment for requiring students to show in practices, the specific skills and competencies they have mastered.

A study by Topping (2015) and Arhin (2015) found that teacher’s feedback in PBA can improve learning and that classroom instructional strategies is positively influenced by the use of PBA strategies. This is clear indication that as found by this study, PBA has the potential to influence the instructional

strategy to improve students' learning through prompt and effective feedback from teachers.

In a convergent view, Sun-Geun and Eun-Hui's (2015) study found that performance assessment has positive effect on the educational value in the teaching and learning of science. The results of Sun-Geun and Eun-Hui (2015) is affirmed by this study. Even though, the research designs were different, the studies came up with the same findings that PBA has an educational value. Whiles Sun-Geun and Eun-Hui (2015) used an experimental design, this study used a descriptive design.

Kone (2015) also found that the performance assessment has a positive effect on the motivation of the students. That is, PBA could be used to motivate student to learn. In this study, the participants reported that PBA could enhance effective classroom teaching and learning. Like Sun-Geun and Eun-Hui (2015), Kone (2015) used an experimental design and still arrived at the same findings as this study.

Catalytic effect of the PBA

The study found that PBA reveals students' true performance, immediate feedback can be given to students, students will be able to reflect on their performance and that PBA could be used in the classroom to give prompt feedback to students. The result further showed that the status, School Category and gender main effects were not significant in the feedback provided by PBA to stimulate students' learning. This means that generally, there was no significant difference in the feedback provided by PBA to stimulate students' learning as expressed by male and female respondents, the teachers and the examiners as well

respondents in the Western Region. However, experience main effects was significant in the catalytic effect of the PBA. This means that there was difference in feedback provided by PBA to stimulate students' learning among respondents on the base of their experience.

The finding supports the finding of Palm (2008) that PBA enables students to synchronize their knowledge and apply the knowledge to a new situation outside classroom setting. This means that PBA provides prompt feedback to students that stimulate their learning. In similar manner, Wiggins and McTighe (2015) found that irrespective of the type of performance, all PBAs have one thing in common and that is, the performance of an authentic task that depicts a real-life experience and also mimics challenges in real world. It could be seen that the findings of this study do not deviate from works of Wiggins and McTighe (2015). Performance-based assessment stimulates students learning in a real-life experience from the feedback of teachers. Topping (2015) supported this assertion by stating that teachers' feedback in PBA can improve learning. It is therefore not surprising that the developed PBA of this study was found to possess the characteristics of providing feedback that stimulates students learning.

Commenting on PBA and students learning, Darling-Hammond and Pecheone, (2019) stated that timely feedbacks are provided to students work when PBA is used as a formative assessment than large-scale standardized tests. This is because sometimes, it takes more than a month to produce results of standardised tests. For PBA, teachers could meaningfully modify the assessment at the time of teaching their current students. This means that the findings of this study support

literature that PBA can provide prompt feedback to students in the classroom which stimulate students learning. This is, PBA could be used as a formative assessment in the classroom with prompt feedback.

The studies of Sun-Geun and Eun-Hui (2015), Kone (2015) and Sung-Eun (2015) all found that performance assessment has positive effect on students learning. This is confirmed by this finding that the PBA has catalytic effects. Again, the study of Kone (2015) revealed that students' motivation varied across experience. This is an indication that like this study, there is statistically significant difference in the catalytic effect of PBA due to experience. The convergent findings could be as a result of the similarity of methodology (instrument and design). The only line of difference in this study and that of Sun-Geun and Eun-Hui (2015), Kone (2015) and Sung-Eun (2015) is that this study was conducted in mathematics other than the other subjects.

Arhin (2015) in his experimental study with PBA found that the experimental group performed better and showed a positive attitude than the control group. The finding of Arhin (2015), revealed an important aspect of the catalytic effect of PBA-motivation. The use of PBA encourages students to learn for understanding so that they can apply the knowledge to solve their immediate problems. Hence, this study which found that PBA has catalytic effect is in line with the study of Arhin (2015) even though, this study employed a descriptive design which perceptions of stakeholders were measured.

Reliability of the PBA

The result of the study revealed a high inter-rater reliability of the instrument. Chan and Malim (2017) used Cronbach's alpha to estimate the

reliability of their instrument which was a questionnaire. The reliability coefficient was 0.939 indicating high reliability for internal consistency. The result of this also reported a high inter-rater reliability for the instrument which was of the graded response type both within the items and the entire instrument. It could be concluded that the results of this is in line with Chan and Malim (2017) that an assessment instrument should have a high reliability.

Also, Reid (2014) estimated the reliability of a questionnaire on two different occasions with Cronbach alpha. Reliability coefficients were 0.851 and 0.822 respectively indicating a high reliable instrument. In this vain, this study though different in format from that of Chan and Malim (2017) and Reid (2014) produced a high inter-rater reliability for the items and the scale. This also indicates that this instrument is a good one as far as reliability of an instrument is concerned. Hasnida and Ghazali (2016) like Chan and Malim (2017) estimated the reliability of the instrument using internal consistence reliability, which is measured by alpha coefficient reliability or Cronbach Alpha. The finding of the study showed that the instrument was reliable.

Validity of PBA

The results of the study revealed that both the items and the scale designed have good CVR statistics. That is, the key stakeholders considered the items and the entire instrument to be relevant to the objective of the mathematics curriculum. The result also revealed that all the five items measure a single construct; mathematical ability and that each item significantly contribute the construct being measured. The construct validity of the instrument is strong. Zamanzadeh, Ghahramanian, Rassouli, Abbaszadeh, Alavi-Majd and Nikanfar

(2015) found in their content validity study that the instrument enjoys an appropriate level of content validity S-CVI with the average approach, which was equal to 0.93. The study of Zamanzadeh, et al. (2015) used the Lawshe (1927) method to estimate content validity ratio while this current study used the modified Kappa statistic. However, both reported a good level of content validity ratio. Unlike this study, the construct validity was not estimated.

Further, the result of Chan and Malim (2017) which estimated the divergent and convergent construct with principal component analysis extraction and Varimax rotation reported that the items had good factor. Sixty-two (62) items retaining with the factor loadings that was above 0.4. Majority of the items were considered to be good. The difference between this study and that of Chan and Malim (2017) was in the format of the items. While Chan and Malim (2017) used a question, this study studied graded responses type in mathematics. Therefore, this study did not estimate factor loading but unidimensionality of items. That is whether the items measure a common construct. The goodness of items in this study was in the unidimensionality (convergent validity), that of Chan and Malim was in different construct (divergent validity and convergent validity) with good factor loadings. Another line of difference was that Chan and Malim (2017) did not estimate the content validity of their instrument.

Similar to Chan and Malim (2017), Reid (2014) used the confirmatory factor analysis factor ranged from 0.453 to 0.859 high loading on attitude factor. Reid (2014) therefore looked at convergent validity of the instrument because, the divergent validity had already been estimated. The instrument was as good as this

instrument but for different purposes. Unfortunately, content validity was not evaluated.

Hasnida and Ghazali (2016) also evaluated the content validity by the experts. A qualitative analysis was done and reported that the instrument which was a questionnaire has content relevance. This study on the other hand used a statistical procedure to estimate the content validity ratio which was found to be good. Construct validity which was measured by Exploratory Factor Analysis (EFA) found that 69 out of the 72 items were retained based on the loadings. Both the divergent and convergent validity were estimated. The instrument in this study also produced acceptable indices of validity, both content and construct validity.

Chapter Summary

The study sought to develop and validate PBA for use in SHSs. It was found the self-developed PBA meet the criteria for good assessment. Specifically, the newly developed PBA is feasible and credible to be used in the SHSs in the classroom as expressed by the teachers and in external examination as expressed by the examiners. Also, the newly developed PBA has both educational and catalytic effects. The instrument could stimulate students' learning and immediate feedback could be provided to students to enhance their learning.

The instrument (odPIM) was found to have good inter-rater reliability coefficient for items and scale. The reliability coefficient ranged from 0.895 to 0.988. The instrument was found have a good level of content validity with

coefficient ranging from 0.834 to 1.000. The construct validity also revealed that the instrument has unidimensionality characteristics.



CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

This chapter presents the summary and the key findings of the study. It also presents the conclusion, recommendation and suggestions for further studies.

Summary of the Study

The purpose of this study was to develop and validate a PBA for SHSs. This study implemented a four-phase instrument development process: (a) planning, (b) construction, (c) quantitative evaluations, and (d) validation. The study used employed quantitative instrumentation research design with a four-phase instrument development and validation process: planning, construction, qualitative evaluations, and quantitative validation for the development and validation of the developed instrument. The study made use of stratified, simple random and purposive sampling techniques.

In all, sample of 240 mathematics examiners, 150 mathematics teachers and 750 SHS three students in the Western Region of Ghana were used for the validation phase of the instrument development. The instruments for the data collection of the study were the PBA items in mathematics and a questionnaire. The test was used for the quantitative evaluation of the instrument designed in terms of reliability and validity while the questionnaire was used for the quantitative validation of the instrument in terms of feasibility, credibility, educational and catalytic effects.

In the first phase, questionnaire was administered to examiners at the centres at the time of conference marking and coordination with a sample of the

PBA test attached to the questionnaire. In the second phase, the questionnaire with a sample of the PBA test was administered to the selected teachers in their respective schools. The purpose of the questionnaire was to illicit information on examination malpractice on the face of the PBA test and to validate the PBA in terms of feasibility, credibility, educational and catalytic effects of the newly developed PBA. In the final phase, selected class of students sat for the PBA test. Means and standard deviation, Pearson Moment correlation coefficient, modified Kappa statistics, Principal Component Analysis and four-way ANOVA were used for the analyses. The following were the findings of the study:

1. Performance-based assessment is feasible to be used for SHSs.
2. Performance-based assessment is credible to be used for SHSs.
3. Performance-based assessment could stimulate students' learning; students are motivated to learn, students are compelled to learn, PBA encourages students to think differently on an issue and that PBA encourages students to learn extensively.
4. Performance-based assessment was reported to provide immediate feedback to students, students will be able to reflect on their performance and that PBA could be used in the classroom to give prompt feedback to students.
5. It was found that the instrument has good inter-rater reliability coefficient ranging from 0.879 to 0.988.
6. It was found that the items have good CVR ranging from 0.834 to 1.00. This means the instrument has a good content validity. Also, the result

indicated the items on the instrument measures one construct. The results revealed one component with an eigenvalue greater than 1 with 89.06%. All the items have acceptable level of factor loading. This means that the instrument has a good convergent construct validity.

7. The results show that all the main effects (status, gender, status and experience) had no statistically significant effect in feasibility of the developed PBA. The teachers and examiners in the Western Region of all experience levels expressed the same thing on feasibility of the developed PBA.
8. The results show that all the main effects (status, gender, status and experience) had no statistically significant effect in credibility of the developed PBA, except for experience, which showed a significant effect in the credibility of the newly developed PBA. However, teachers' and examiners' expression of the credibility of the instrument was different based on their years of experience. The results further revealed that mathematics teachers and mathematics examiners who have 6-10 years of experience expressed a higher credibility of the instrument than colleagues of other levels of experience.
9. The results show that while all the main effects (School Category, status, gender and experiences) had no statistically significant effect in catalytic effect of the PBA, gender and experience main effects had significant effect in catalytic effect of the PBA. That is the teachers and examiners in the Western Region of all years of experience expressed the same thing on

the possibility of the developed PBA to provide immediate feedback on students learning.

10. In the case of educational effect, there was statistically significant difference in the expression of the possibilities of the developed PBA to motivate students' learning among respondents on the bases of their gender and experience. The result further revealed that males expressed a higher educational effect of the instrument than females. Also, teachers and examiners who have 11-15yrs of experience expressed higher possibility of the educational effect of the instrument than their counterparts of other years of experience.

Conclusion

The PBA for SHSs has been found to have the validity, reliability, feasibility, educational effect, catalytic effect and credibility after the validation. However, some limitations were identified in how well the instrument meets the criteria of a good assessment during the validation of the instrument. For instance, mathematics teachers and examiners showed significant difference in the evaluation of the instrument due to experience. The teachers and examiners with 1-5 years and above 20 years of experience expressed little faith in the instrument. This means that further discussion with the mathematics teachers and WAEC mathematics examiners would be required to strengthen the developed instrument for use. For instance, a nationwide coordination would strengthen the reliability of raters.

The traditional type of items in mathematics for SHSs could be modified a bit to make them a PBA, where students would be required to apply knowledge and skills acquired in mathematics to real life situation. It has also revealed that PBA of this nature could be used in the SHSs to have the educational and catalytic effects required. This assessment is also feasible for use in the SHSs. This study would make a significant contribution to knowledge in the area of PBA for SHSs.

Contribution to Knowledge, Practice and Policy

Knowledge

The study would provide a guide on how to validate polytomous items using modified Kappa statistics, PCA and Pearson Product Moment correlation. There is no known validation of polytomous items at the SHS in Ghana. Perhaps, assessors in Ghana such as WAEC and teachers do not know the procedures for validation of the polytomous items except the dichotomous items.

Practice

The study has shed light on the use of PBA in mathematics in SHS. That is, the study would be a guide on how to develop PBA items in mathematics for SHS.

Policy

The study has provided information on the feasibility, credibility, educational and catalytic effects of PBA in mathematics. This is the first-time feasibility, credibility, educational and catalytic effects of PBA in mathematics for SHS has been investigated. This would help Ghana Education Service and

Ministry of Education to formulate a policy on PBA in mathematics as core in the assessment of students in mathematics at the SHS level.

The study would inform curriculum developers of the SHS to develop curriculum that would centre on PBA for SHS. This help realise the educational and catalytic effects of PBA on SHS students.

Recommendations

Based on the findings of the study, the following recommendations and suggestions have been listed for consideration by authorities and stakeholders in Ghanaian education to help the teaching and learning and assessment of mathematics at the SHS level:

1. The teacher educators should make PBA an integral part of assessment lessons and course at both the colleges of education and universities where mathematics teachers are trained by the curriculum developers in mathematics education. This would help provide the knowledge and skills on PBA needed by the mathematics teachers to have an effective and efficient assessment in mathematics.
2. Mathematics teachers should make use of the developed PBA as an assessment strategy in teaching and learning of mathematics. In the teaching practice, student-teachers should be made to employ PBA so that its usage would be internalised in the mathematics teachers.
3. The PBA should be used in SHS by mathematics teachers for high-stake examinations such as end of semester and mock examinations. This would

help the teachers and students have a feel of the PBA as an external examination format.

4. The West African Examination Council should give a try-out of the PBA in the SHS for some selected schools to further ascertain the strength and weakness of the developed PBA. This would help address any limitation to strengthen it for use in WAEC examinations at the SHS level.

Suggestions for Further Research

Based on the findings of this research, the following areas would be suggested for further research:

1. The same study should be carried out in other subjects to find out if the PBA would have favourable characteristics for use on other subjects other than mathematics.
2. The same study should be conducted with a larger sample of teachers, examiners, schools and students. This will help have sample characteristics closer to the population characteristics for better generalization of the findings.

REFERENCES

- Adib, D. A., Rusilowati, A., & Hidayah, T. (2018). Development of authentic appraisal instruments basic skills for playing football of Junior High School Students. *Active Learning in Higher Education*, 7(1), 9–18.
- Adjei, E., & Tagoe, M. (2009). *Research methods in information studies*. Accra: IAE (UG).
- Agu, N. N., Onyekuba, C., & Anyichie, A. C. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: Need for a test construction skill inventory. *Educational Research and Reviews*, 8(8), 431-439.
- Ainsworth, L., & Viegut, D. (2006). *Common formative assessments: How to connect standards-based instruction and assessment*. Thousand Oaks, CA: Corwin.
- Airasian, P. W. (2001). *Classroom assessment: Concepts and applications*. New York: McGraw-Hill.
- Allen, M. J. (2004). *Assessing academic programs in higher education*. San Francisco: Jossey-Bass.
- Amedahe, F. A. (2000). *Assessment in schools*. University of Cape Coast: Unpublished Memeograph.
- Amedahe, F. K. (2012). *Fundamentals of educational research methods*. University of Cape Coast: Unpublished Mimeograph.
- Amedahe, F. K., & Asamoah-Gyimah, K. (2015). *Measurement and evaluation*. Cape Coast: CODE.

American Association for Public Opinion Research. (2016). *Best practices for research*. Retrieved from <http://www.aapor.org/Standards-Ethics/Best-Practices.aspx>.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA, APA & NCME.

American Educational Research Association, American Psychological Association & National Council on Measurement (1999). *Standards for educational and psychological testing*. USA: American Educational Research, American Psychological Association & National Council on Measurement.

American Educational Research Association. (2014). Evaluating evidence regarding relationships with criteria. *Measurement and Evaluation in Counselling and Development*, 50, 264–269.

Anamuah-Mensah, J., Mereku, D. K., & Asabere-Ameyaw, A. (2004). *Ghana junior secondary school students' achievement in mathematics and science, Results from Ghana's participation in the 2003 trends in international mathematics and science study*. Ministry of Education, Youth, and Sports: Accra.

Ankomah, F. (2020). *Predictors of adherence to test construction principles: The case of senior high school teachers in Sekondi-Takoradi Metropolis.*

Unpublished master's thesis, Department of Education and Psychology
University of Cape Coast.

Annan-Brew, R. (2020). *Differential item functioning of West African senior secondary certificate examination in core subjects in southern Ghana.*

Unpublished doctoral thesis, Department of Education and Psychology
University of Cape Coast.

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(2), 112–114.

Arias-Estero, J., & Castejón, F. (2014). Using instruments for tactical assessment in physical education and extra-curricular sports. *European Physical Education Review*, 15, 38–51.

Armah, C. (2018). *Test construction and administration practices among lecturers and staff of examinations unit of the University of Cape Coast.*

Unpublished master's thesis, Department of Education and Psychology
University of Cape Coast.

Asamoah-Gyimah, K., & Anane, E. (2018). *Assessment in schools.* University of Cape Coast: Unpublished Mimeograph.

Australian Association of Mathematics Teachers (2002). *Standards for excellence in teaching Mathematics in Australian Schools*. Adelaide, South Australia: Author.

Azwar, S. (2012). *Reliability and validity*. Yogyakarta: Pustaka Pelajar.

Bahr, D. L., Monroe, E. E., & Mantilla, J. (2018). Developing a framework of outcomes for Mathematics teacher learning: Three Mathematics educators engage in collaborative self-study. *Teacher Education Quarterly*, Spring 2018.

Bardes, B., & Denton, J. (2001). *Using the grading process for departmental and program assessment*. Paper presented at the American Association for Higher Education Conference. New York: Denver.

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. *American Journal of Occupational Therapy*, 36(12), 789–800. <https://doi.org/10.5014/ajot.36.12.789>.

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. *The American Journal of Occupational Therapy*, 36, 789-800.

Bichi, A. A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: A review of empirical studies. *Austrian Journal of Basic & Applied Science* 9(7): 549-556.

Bollen, K. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, 35(2), 359-372.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education*, 31(4), 399 - 413.

Boursicot, K., Kemp, S., Wilkinson, T.J., Findyartini, A., Canning, C.A., Cilliers, F., & Fuller, R. (2020). Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. *Medical teacher*, 1-

10.

Brennan, R. L. (2002). *Elements of generalizability theory*. Iowa City, IA: ACT.

Brennan, R. L. (2002). Generalizability theory. In A. J. Nitko (Ed). *Educational measurement* (3rd). USA: American Council on Education.

Brennan, R. L. (Ed). (2006). *Educational measurement* (4th ed). USA: American Council on Education, Praeger Series on Education.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9), 2812-2831.

Burdis, J. R. (2014). *Designing and evaluating a Russian elicited imitation test to be used at the Missionary Training Center*. Unpublished doctoral dissertation, State University, Russia.

Burkhardt, H., & Swan, M. (2008). Designing assessment of performance in mathematics. *Educational Measurement: Issues and Practice*, 9(4), 1–24.

Burt, R. S. (2017). Structural holes versus network closure as social capital. *Social capital*. 31-56.

Butakor, P. K. (2016). Hierarchical linear modeling of the relationship between attitudinal and instructional variables and mathematics achievement. *International Journal of Research in Education Methodology*, 7(5), 1328-1336.

California Department of Education (2013). *Replace the proposed new California math curriculum framework*. Open Letter to Governor Gavin Newsom, State Superintendent Tony Thurmond, the State Board of Education, and the Instructional Quality Commission.

Camilli, G. (2006). Test fairness. In R. L. (Ed.), *Educational measurement* (4th ed., pp. 220-256). Westport, CT: American Council on Education.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, 33(4) 453–472.

Chan, L. L., & Malim, T. (2017). Validity and reliability of the instrument using exploratory factor analysis and Cronbach's alpha. *International Journal of Academic Research in Business and Social Sciences*, 7(10), 1-13.

Choudrie, J., & Dwivedi, Y. K. (2005). *Investigating broadband diffusion in the household: towards content validity and pre-test of the survey instrument research-in-progress*. *Proceedings from European Conference on Information Systems (ECIS)*. Germany: Association for Information Systems.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring and reporting, In R. L. Brennan (Ed.) *Educational Measurement* (4th ed.), (pp. 355 - 386). Washington, DC: National Council on Measurement in Education and American Council on Education.

Cohen, L., Manion, L., & Marrison, K. (2000). *Research methods in education*. London: Routledge Falmer.

Collins, A., Hawkins, J., & Frederiksen, J. (1990). *Technology-based performance assessments*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April 2-5, 1990.

Cooper, D. R., & Schindler, S. S. (2009). *Business research methods*. (10th ed.). New York: McGraw-Hill Higher Education.

Creswell, J. W. (2002). *Educational Research: planning, conducting and evaluating quantitative and qualitative research*. New Jersey, Pearson Education, Inc.,

Creswell, J. W. (2009). *Research design: Qualitative, quantitative and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed). Thousand Oaks, California: Sage Publications.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. USA: Lengage Learning.

Cullinane, A. (2011). Formative assessment classroom techniques. *Resource & Research Guides*, 2(13), 1-4.

Darling-Hammond, L., & Pecheone, P. (2019). Equity issues in performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89-114). Boston: Kluwer.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5(4), 194–197.

DeMars, C. E. (2018). Item information function (ed.) In *SAGE Encyclopaedia of Educational Research, Measurement, and Evaluation*. Thousand Oak SAGE Publications.

Dhindsa, H. S., Omar, K., & Waldrip, B. (2007) Upper Secondary Bruneian science students' perceptions of assessment. *International Journal of Science Education* 29(10):1261-1280.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Draugalis, J., Coons, S., & Plaza, C. (2008). Best practices for survey research reports: A synopsis for authors and reviewers. *American Journal of Pharmaceutical Education*, 72(1), 1-6.

Drost, E. A. (n.d). Validity and reliability in social science research. *Education Research and Perspectives*, 38 (1), 105-124.

Dunbar, N. E. et al. (2017). Mitigation of cognitive bias with a serious game: Two experiments testing feedback timing and source. *International Journal of Game-Based Learning*, 7(4), 86-100.

Edmunds, J. (2006). *How to assess student performance in history: Going beyond multiple-choice tests*. Produced by the SERVE center at the University of North Carolina at Greensboro.

El-sehrawy, M. G. (2020). Developing and validating an instrument for evaluating research protocols. *Journal of Nursing and Health Science (IOSR-JNHS)*, 9 (4), 10-17.

Erzoah, K. K., Gyamfi, A., Yeboah, A., & Langee, P. (2022). Teachers' knowledge and practices of classroom assessment in the Ellembelle District of Ghana. *Advances in Research*, 23(4), 1-10.

Estacio, R. D. (2015). *Development and validation of learning assessment tool and instructional material in physics 1 (mechanics)*. Unpublished Master Thesis, Eulogio "Amang" Rodriguez Institute of Science and Technology.

Etsey, Y. K. A. (2012). *Assessment in education*. University of Cape Coast: Unpublished Mimeograph.

Etsey, Y. K. A., & Gyamfi, A. (2017). Improving assessment of learning in mathematics through assessment as learning. *Journal of Educational Assessment in Africa*, 12(1), 11-20.

Ewetan, T. O., & Ewetan, O. O. (2015). Teachers' teaching experience and academic performance in mathematics and English language in public secondary schools in Ogun State, Nigeria. *International Journal of Humanities Social Sciences and Education (IJHSSE)*, 2 (2), 123-134.

Falk, B., Ort, S. W., & Moirs, K. (2007). Keeping the focus on the child: Supporting and reporting on teaching and learning with a classroom-based performance assessment system. *Educational Assessment*, 12 (1), 47-75.

Feldt, L. S., & Brennan, R. L. (2001). Reliability. In A. J. Nitko (Ed). *Educational measurement* (3rd). USA: American Council on Education.

Fowler, F. J. (2008). *Survey research methods* (4th ed.). Los Angeles, CA: Sage.

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education*. seventh ed. New York: McGraw-Hill.

- Gable, R. K., & Wolf, M. B. (2012). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings* (4th.). Boston, MA: Kluwer Academic Publishers.
- Galbraith, P. (2016). Students, mathematics, and technology: assessing the present – challenging the future. *International Journal of Mathematical Education in Science and Technology*, 37 (3), 277-290.
- Gao, M. (2012) Classroom assessments in mathematics: High school students' perceptions *International Journal of Business and Social Science*, 3(2), 63-74.
- García-López, L. M., González-Víllora, S., Gutiérrez, D., & Serra, J. (2013). Development and validation of the Game Performance Evaluation Tool (GPET) in soccer. *Revista Euroamericana de Ciencias Del Deporte*, 2(1), 89–99.
- Garrison, C., Chandler, D., & Ehringhaus, M. (2020). *Effective classroom assessment: Linking assessment with instruction*. Westerville, Ohio: Measured Progress.
- Ghana Education Service (2019). *Computerised school selection and placement manual*. Accra: Author.
- Goodrum, D., Hackling, M., & Rennie, L. (2001). *The status and quality of teaching and learning of science in Australian schools*. Canberra: Department of Education, Training and Youth Affairs.
- Gordon, M. (2008) *Assess notes nursing assessment and diagnostic reasoning*. Philadelphia: F.A. Davis Company Press.

- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11, 255–274. <https://doi.org/10.3102/01623737011003255>
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427–438
- Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability. (4th Edition)* Gaithersburg: Advanced Analytics, LLC.
- Gyamfi, A. (2017a). *Impact of assessment as learning on academic achievement and attitudes towards mathematics of senior high school students in Ahanta west and Mpohor districts*. Unpublished master's thesis, Department of Education and Psychology, University of Cape Coast.
- Gyamfi, A. (2017b). *Measurement and evaluation*. Unpublished Mimeograph.
- Gyamfi, A. (2022a). Controlling examination malpractice in Senior High Schools in Ghana through performance-based assessment. *Journal of Advances in Education and Philosophy*, 6(3), 203-211.
- Gyamfi, A. (2022b). Application of Classical Test Theory (CTT) in the validation of teacher made Mathematics Multiple Choice Test (MMCT) items in Ghana. *International Journal of Research and Innovation in Social Science*, 6(2), 78-93.
- Gyimah, E. K., Ntim, K., & Deku, P. (2012). *Assessment in special education*. Cape Coast: Hamilton Press.

- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis (7th Edition)*. New York: Pearson Education
- Hamavandy, M., & Kiany, G. R. (2014). A historical overview on the concept of validity in language testing. *Advances in Language and Literary Studies*, 5(4), 86-91.
- Hasnida, N., & Ghazali, M. (2016). A Reliability and Validity of an Instrument to Evaluate the School-Based Assessment System: A Pilot Study. *International Journal of Evaluation and Research in Education (IJERE)*, 5(2), 148-157.
- Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidenced Based Nursing*, 18 (3), 66-67.
- Hibbard, K. M (2017). *Performance-Based Learning and Assessment in middle school science.*, United Kingdom: Taylor & Francis Ltd.
- Hild, P., Gut, C., & Brückmann, M. (2018). Validating performance assessments: measures that may help to evaluate students' expertise in 'doing science', *Research in Science & Technological Education*, DOI: 10.1080/02635143.2018.1552851.
- Hodges. L. (2014). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher* 35(7), 56-68.
- Iji, C. O., & Omenka, J. E. (2014). Mathematics Teachers' Perception of Difficult Concepts in Secondary School Mathematics Curriculum in Benue State, Nigeria. *Asia Pacific Journal of Education, Arts and Sciences*, 2 (1), 23-33.

Janssen, G., Meier, V., & Trace, J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. *Colombian Applied Linguistics Journal*, 16(2), 37-54.

Jiraro, S., Sujiva, S., & Wongwanich, S. (2014). An application of action research for teacher empowerment to develop teachers' test construction competency development models. *Procedia-Social and Behavioral Sciences*, 116, 1263-1267. Available at: <https://doi.org/10.1016/j.sbspro.2014.01.380>.

Johnson, R. C. (2011). *Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context*. Unpublished doctoral thesis, Macquarie University, UK.

Kamaldeen, A., Buhari, A. M., & Parakoyi, D. B. (2012). Perception, attitude and practices of parents in Okene, Nigeria towards girls-child education. *International Journal of Scientific and Research Publication*, 2(8), 1-7.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29, 3-17.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kelley, K., Clark, B., Brown, V., & Sitzia, J. (2003). Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, 15, 261-266.

Koh, C. E., & Nam, K. T. (2005). Business use of the internet: A longitudinal study from a value chain perspective. *Industrial Management & Data Systems*, 105 (1), 82-96.

Koné, K. (2015). *The impact of performance-based assessment on University ESL learners' motivation*. Unpublished Master's thesis, Minnesota State University, Mankato.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, 43, 489-493.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>

Leonelli, E. D., & Schmitt, M. J. (2012). Bringing reform to adult numeracy instruction. *Field Notes*, 11(2), 226–246.

Liaquat, H, Asif, J. M., Siraji, J., & Maroof, K. (2012). Development and standardization of intelligence test for children. *International Journal of Learning & Development*, 2(5), 190-202.

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437- 448.

Manning, S. G. (2015). *The development and validation of a measure of disengagement*. Unpublished masters' thesis, Colorado State University, USA.

- Marzano, R. J.; Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: Mid-Continent Research for Education and Learning, Aurora, & Association for Supervision and Curriculum Development.
- Messick, S. (2001). Validity. In A. J. Nitko (Ed). *Educational measurement* (3rd). USA: American council on education.
- Metin, M. (2013). Teachers' difficulties in preparation and implementation of performance task. *Educational Sciences: Theory and Practice*, 13 (3), 1644-1673.
- Metzger, S., Gut, C., Hild, P., & Tardent, J. (2014). *Modelling and assessing experimental competence: An interdisciplinary progression model for hands-on Assessments*. In E-proceedings of The ESERA 2013 conference in Nicosia.
- Ministry of Education (2012). *Core mathematics syllabus for senior high schools*. Accra: CRDD.
- Ministry of Education (MOE) (2018). *Core mathematics syllabus for senior high schools*. Accra: CRDD.
- Morrell, P. D., & Carroll, J. B. (2010). *Conducting educational research: A primer for teachers and administrators*. Rotterdam, The Netherlands: Sense Publishers.
- Mpuangnan, K. N., & Adusei, O. (2021). Implementation of standard-based curriculum in Ghana: The concerns of basic school teacher. *International Journal of Education and Research*, 9 (3), 53-66.

Mussawy, S. A. J. (2009). *Assessment Practices: Student's and teachers' perceptions of classroom assessment*. Unpublished master's Thesis, University of Massachusetts.

National Association of Testing Authorities, (2012). *Guidelines for the validation and verification of quantitative and qualitative test methods: Technical note 17*. Australia: Author.

National Council of Teachers of Mathematics (1987). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics (1995). *Assessment Standards for School Mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics (2002). *Assessment standards for school mathematics*. Reston, VA: Author

National Council of Teachers of Mathematics (2010). *Teaching mathematics in the middle school*. Reston, VA: Author.

National Council of Teachers of Mathematics (NCTM). (1995). *Curriculum and evaluation standards for schools mathematics*. Reston, VA: Author

Neuman, W. L. (2003). *Social research methods: Qualitative and quantitative approaches*. Boston: Allyn and Bacon.

Newman, I., Lim, J., & Pineda, F. (2013). Content validity using a mixed methods approach: Its application and development through the use of a table of specifications methodology. *Journal of Mixed Methods Research*, 7(3), 243–260.

Newton, P. E. (2014). *Validity in educational and psychological assessment*. Thousand Oaks, CA: Sage.

Nitko, A. (1996). *Educational Assessment of Students* (3rd Ed). Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Nitko, A. (1996). *Educational Assessment of Students* (8th Ed). Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Nitko, A. (2012). *Educational Assessment of Students* (6th Ed). Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Nitko, A. J. (2001). *Educational measurements* (3rded.) (Ed) USA: American council on education.

Nitko, A. J. (2002). *Educational Tests and Measurements* (2nded.). USA: Prentice-Hall, Inc.

Nitko, A. J. (2004). *Educational measurement* (4rd Ed.). USA: American Council on Education & Praeger.

Nitko, A. J. (2004). *Educational Tests and Measurements* (3rded.). USA: Prentice-Hall, Inc.

Nugroho, M. D., & Tomoliyus, T. (2019). Validation of performance assessment instrument on futsal game in extracurricular activities. *Jurnal SPORTIF: Jurnal Penelitian Pembelajaran*, 5(2), 175-183.

Office of Educational Research and Improvement (2009). *Grant announcement: National Educational Research and Development Center Program*. Washington, DC: U.S. Department of Education.

Onwuegbuzie A. J., & Combs J. P. (2010). Emergent data analysis techniques in mixed methods research: A synthesis. In Tashakkori A., Teddlie C. (Eds.), *Sage handbook of mixed methods in social and behavioral research* (2nd ed., pp. 397-430). Thousand Oaks, CA: Sage.

Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. T. (2008). Interviewing the Interpretive Researcher: A Method for Addressing the Crises of Representation, Legitimation, and Praxis. *International Journal of Qualitative Methods*, 7(4), 1–17.
<https://doi.org/10.1177/160940690800700401>.

Onwuegbuzie, A., Daniel, L., & Collins, K. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity: International Journal of Methodology, Springer*, 43(2), 197-209.

Onwuegbuzie, J., Bustamante, R., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, 4(1), 56-78.

Oslin, J. L., Mitchell, S. A., & Griffin, L. L. (2016). The Game Performance Assessment Instrument (GPAI): Development and preliminary validation. *Journal of Teaching in Physical Education*, 17(2), 231–243.

Osterlind, S., & Merz, W. R. (1994). Building a taxonomy for constructed-response test items. *Educational Assessment* 2(2):133-147.

Osuola, E. (2001). *Introduction to research methodology* (2nd. ed). Onitisha, Nigeria: Africana EEP Publishers Ltd.

Ottawa Conference (2010). *Assessment of Competence in Medicine and the Healthcare Professions*. Canada: Author.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, 13(4), 1-11.

Park, D., Bahrudin, F. I., & Han, J. (2020). The evolution of research through a strategic construction of research methodologies. *International Journal of Quantitative and Qualitative Research Methods*, 8(3), 1-23.

Pegg, J., (2013). Assessment in mathematics: A developmental approach. In J. Royer (Ed.), *Mathematical Cognition* (pp. 227–259). Greenwich, CT: Information Age Publishing.

Pelegriano J. W., Chubowsky, N., & Glaser, R. (Eds) (2013). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington DC: National Academy Press.

Pellegrino, J. W., & Goldman, S. R. (2008). Beyond Rhetoric: Realities and Complexities of Integrating Assessment into Classroom Teaching and Learning. In C. A. Dwyer (Ed). *The future of assessment (1st ed)*. New York: Routledge.

Pineda, M. F. (2012). *Developing a process to create and validate an instrument assessing student attainment of competencies at an Intercultural University in Mexico*. Unpublished doctoral dissertation, Federal Intercultural University, Mexico.

Pishghadam, R. P., Baghdei, P., & Shayesteh, S. (2012). Construction and validation of an English Language Teacher Creativity Scale (ELT-CS). *Journal of American Science*; 8(3), 497-508.

Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS *Educational Measurement: Issues and Practice*, 23(2), 12-16.

Polit, D. F., & Yang, F. M. (2016). *Measurement and the measurement of change*. Philadelphia, PA: Wolters Kluwer.

Polit, D. F., Beck, C. T., & Owen, S. V. (2017). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health*, 30(4):459–67.

Quansah, F. (2020). Students' evaluation of the quality of teaching using generalisability theory: A case of a selected university in Ghana. *South African Journal of Higher Education*, 34 (5), 136–150.

Ramsenthaler, C., Gao, W., Siegert, R. J., Schey, S. A., Edmonds, P. M., & Higginson, I. J. (n.d.). Longitudinal validity and reliability of the Myeloma Patient Outcome Scale (MyPOS) was established using traditional , generalizability and Rasch psychometric methods. *Quality of Life Research*, 26(11), 2931-2947.

Reid, T. (2014). *Development and validation of an instrument assessing preschool children attitude towards science*. Unpublished master's thesis, University of Hawaii, Hawaii.

Resnick, L. B., & Resnick, D. L. (2001). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *future assessments: Changing views of aptitude, achievement, and instruction*. Boston: Kluwer.

Retnawati, H. (2017). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *Research and Evaluation in Education*, 12(2), 12-25.

Rosaroso, R. C., & Rosaroro, N. A. R. (2015). Performance-based assessment in selected higher education institutions in Cebu City, Philippines. *Asia Pacific Journal of Multidisciplinary Research*, 3(4), 72-77.

Royal, K. D. (2017). Four tenets of modern validity theory for medical education assessment and evaluation. *Advances in medical education and practice*, 8, 567–570. <https://doi.org/10.2147/AMEP.S139492>.

Royal, K. D., & Gonzalez, L. M. (2016). An evaluation of the psychometric properties of an advising survey for medical and professional program students. *Journal of Educational and Developmental Psychology*, 6(1), 195 – 203.

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment in special and remedial education*. Boston: Houghton Mifflin.

Sam, N. I, Gyamfi, A., & Yeboah, A. (2019). Examination malpractice in West African Senior Secondary School Examination (WASSCE) in the central region of Ghana. *International Journal for Innovation in Educational Assessment*, 9(8), 23-36.

Sam, N. I. (2012). *Examination malpractice in West African Senior Secondary School Examination (WASSCE) in the central region of Ghana*. Unpublished master's thesis, Department of Education and Psychology University of Cape Coast.

Sarantakos, S. (2000). *Social research*. South yarra: Macmillan education Australia.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: Praeger.

Schoenfeld, A. H. (2000). On mathematics as sense-making: an informal attack on the unfortunate divorce of formal and informal mathematics. In D. N. Perkins, J. Segal, and J. Voss (Eds.), *Informal reasoning in education*. Hillsdale, NJ: Erlbaum.

Schreiber, N., Theyßen, H., & Schecker, H. (2016). "Process-Oriented and Product-Oriented assessment of experimental skills in physics: A Comparison." In *insights from research in science teaching and learning, contributions from Science Education Research*, edited by N. Papadouris, A. Hadjigeorgiou, Angela, C. Constantinou, 29–43. Switzerland: Springer-Verlag.

Shavelson, R. J., Baxter, G. P., & Pine J. (2009). *What alternative assessments look like in science*. Paper presented at Office of Educational Research and Improvement Conference, The Promise and Peril of Alternative Assessment, Washington, DC: October.

Shavelson, R. J., Mayberry, P., & Li, W. (2012). Generalizability of military performance measurements: Marine Corps Infantryman. *Military Psychology*, 6(9), 23-35

Shavelson, R. J., Mayberry, P., & Rowley, A. (2012). Alternative assessment in schools. *Journal of Educational and Developmental Psychology*, 2(8), 98-110.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.

Shepard, L. A. (2009). *The role of classroom assessment in teaching and learning*. CSE Tech Rep. No. 517. Los Angeles: University of California, Center for Study of Evaluation and Santa Cruz, CA: Center for Research on Education, Diversity, and Excellence.

Sireci, S. G. (2013). *A Theory of Action for test validation*. Paper presented at the 13th Annual Maryland Assessment Conference. College Park, MD: University of Maryland.

Sireci, S. G. (2015). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, 1-10. Retrieve on 2nd April, 2019 on <http://dx.doi.org/10.1080/0969594X.2015.1072084>.

- Smisko, A, Twing, J. S. & Denny, P. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education* 1, 3(4), 333-342.
- Smith, T. C., Polloway, E. A., Patton, J. R., & Dowdy, C. (1995). *Teaching students with special needs in inclusive setting*. Boston: Allyn & Beacon
- States, J., Detrich, R., & Keyworth, R. (2018). *Overview of summative assessment*. Oakland, CA: The Wing Institute.
- Stone, C. A., & Lane, S. (2006). Performance assessment. In R. L. Brennan (Ed). *Educational measurement* (4rd). USA: American council on education
- Straub, D., Boudreau, M. C., & Gefen, D. (2004). Validation guidelines for IS positive research. *Communications of the Association for Information Systems*, 13, 380-427.
- Stromquist, N. P. (2007). *The gender socialization process in schools: A cross-national comparison*. Background Paper Prepared for the Education for All Global Monitoring Report 2008. 2008/ED/EFA/MRT/PI/71.
- Strube, M. J. (2002). Reliability and generalizability theory. In L.G. grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.
- Struyven, K. Dochy, F. & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education* 30(4), 331-347.

Sun-Geun, B., & Eun-Hui, H. (2015). A quasi-experimental research on the educational value of performance assessment. *Asia Pacific Education Review*, 6(2), 179-190.

Sung-Eun, K. (2015). *Effect of implementing performance assessment on students' learning: Meta-analysis using HLM*. Unpublished doctoral thesis, The Pennsylvania State University.

Suuramm, C. *et al.* (2016). Assessment in Mathematics Education. In: *Assessment in Mathematics Education*. ICME-13 Topical Surveys. Springer, Cham.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education.

Taherdoost, H. (2016) Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in research. *International Journal of Academic Research in Management (IJARM)*, 5 (3), 28-36.

Taras, M. (2002). Using Assessment for learning and learning from assessment. *Assessment & Evaluation in Higher Education*, 27(6), 501-510.

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, 46 (5), 45–60. doi:10.1016/j.learninstruc.2016.08.003.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Los Angeles, CA: Sage.

Teker, G. T. (2019). Coping with unbalanced designs of generalizability theory: G String V. *International Journal of Assessment Tools in Education*, 6(5), 57–69.

The Duke Endowment (2002). *The student resilience and well-being. A five-year initiative to assess and promote the conditions that help college students flourish*. Duke university, US.

Tomoliyus, T., Sumaryanti, S., & Jadmika, H. M. (2016). Development of validity and reliability of net game performance-based assessment on elementary students' achievement in physical education. *International Journal of Assessment and Evaluation in Education*, 6(4), 41–49.

Topping, K. J. (2015). Trends in peer learning, *Education Psychology*, 25 (6), 631-645.

Torrance, H. (2005). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education*, 14(3), 281-294.

Truxillo, C. (2003). *Multivariate statistical methods: practical research applications course notes*. Cary, N.C.: SAS Institute.

Uzun, N. B., Aktaş, M., Aşiret, S., & Yorulmaz, S. (2018). Determinants of job satisfaction of Colleges of Education lecturers: A Study of Nasarawa State College of Education, Akwanga. *Asian Journal of Education and Training*, 4(2), 85-90.

VanTassel-Baska, J. (2013). Performance-based assessment: The road to authentic learning for the gifted. *Gifted Child Today*.
<https://doi.org/10.1177/10762175135096.1>

WAEC (2016). *Report on the conduct of the May/June 2015 WASSCE in Ghana*. Accra, Ghana: Author.

WAEC (2016, February 18). Candidates involved in examination irregularities during the conduct of the May/June 2015 WASSCE. *The Ghanaian Times* (No. 156645).

WAEC (2017). *Instructions for the conduct of the May/June 2017: West African Senior School Certificate Examination*. Accra, Ghana: Author.

WAEC (2018). *Report on test development and related issues*. Accra, Ghana: Author.

WAEC (2019). *Report on the irregularity and clemency for the May/June West African Senior School Certificate Examination (WASSCE) conducted in Ghana*. Accra, Ghana: Author.

Wang, X., French, B., & Clay, F. P. (2015). Convergent and discriminant validity with formative measurement: A mediator perspective. *Journal of Modern Applied Statistical Methods*, 14 (1) 83-106.

Wanner, V. J. (2004). *Development and validation of a performance-based assessment in work and family life personal development*. PhD Dissertation: Ohio State University.

Wee, Y.S., & Quazi, H.A. (2005). Development and validation of critical factors of environmental management. *Industrial Management & Data Systems*, 105 (2), 96-114.

Werner, L., Denner, J., Campe, S., & Kawamoto, D.C., (2012). *The Fairy Performance Assessment: Measuring Computational Thinking in Middle School*. Proceedings of the 43rd ACM technical symposium on Computer Science Education, 215-220. New York: ACM.

White, M. C. (2017). *Generalizability of scores from classroom observation instrument*. Unpublished doctoral thesis, University of Michigan.

Wiggins, G., & McTighe, J. (2015). *Understanding by Design (4th Ed)*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wilkins, J. L. M., Norton, A., & Boyce, S. J. (2013). Validating a written instrument for assessing students' fractions schemes and operations. *The Mathematics Educator*, 22 (2), 31–54.

Williams, B., Brown, T., & Onsmann, A. (2012). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, 8(3), 1-13. Retrieved from <http://ro.ecu.edu.au/jephc/vol8/iss3/1/> on 4th February, 2020.

Wren, D. G. (2009). Performance assessment: A key component of a balanced assessment system. *Research Brief No. 2*. Report from the Department of Research, Evaluation and Assessment. Virginia Beach City Public Schools.

Wyatt, C. (2016). *The development and validation of an instrument to measure teachers' perceptions of the effect of mobile technology initiatives on classroom climate*. University of Poland: Graduate Theses and Dissertations.

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2013). Two quantitative approaches for estimating content validity. *Western journal of nursing research*, 25(5), 508–518. <https://doi.org/10.1177/0193945903252998>

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2013). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, 25(5), 508–18.

Yeboah, A. (2017). *Relevance of assessment course: A follow-up study of graduate teachers in Ghana*. Unpublished master's thesis, Department of Education and Psychology University of Cape Coast.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: Development of an instrument for measuring Patient-Centered Communication. *Journal of caring sciences*, 4(2), 165–178.

APPENDICES

APPENDIX A- DISTRIBUTION OF THE TARGET AND ACCESSIBLE
POPULATION

Category	School	Target Population		Accessible Population	
		Student	Teacher	Student	Teacher
A	Sekondi college	513	13	513	13
	St John's	403	12	403	12
	GSTS	511	13	511	13
	Archbishop Porters	517	12	517	12
	Faijai	501	13	501	13
B	Shama SHS	309	11	X	X
	Daboase SHS	222	10	222	10
	Adiembra SHS	368	8	X	X
	Bompe SHTS	328	6	X	X
	St. Mary's	352	11	352	11
	Nsein SHS	342	8	X	X
	Esiama SHTS	321	12	321	12
	Half Asini SH	423	9	X	X
	Tarkwa SHS	411	10	411	10
	Amenfiman	301	8	X	X
	Benso SHTS	174	10	174	10
	Ahantaman	324	7	X	X
C	Method. High	316	8	X	X

Takoradi SHS	411	10	X	X
Baidoo Bonsoe	323	11	323	11
Sankor Day	27	5	X	X
Axim Girls	212	7	X	X
Bonzu Kaku	232	8	X	X
Nkroful SHS	304	9	X	X
Annor Adjei	221	9	221	9
Fiaseman SHS	211	9	X	X
Hunivalley	243	11	243	11
St. Augustine	276	8	X	X
Prestea SHTS	223	7	X	X
Asankragua SH	201	8	X	X
Asankragua SHT	142	9	142	9
Gwiraman SHS	98	6	X	X
Mpohor SHS	312	9	X	X
Diabene SHS	132	9	132	9
Uthman Bin Afam	93	5	X	X
Total	7498	321	4986	165

APPENDIX B1-TEST SPECIFICATION FOR THE PERFORMANCE BASED ASSESSMENT ITEMS

Every test construction begins with defining the target construct to be assessed and translating that into test specification. The test specification allows alternate forms for the tasks to be constructed (Nitko, 2004).

Question One

Content: Transformation

Objectives: Ability to reflect, rotate, enlarge and translate objects to get an image.

Description: The items should focus any three of the transformation methods. The item should allow students to form any four pair of numbers within a given range. (Students Plot their ordered parts using appropriate scale). Items should ask students to name their plane shape. Students rotate their object either 90° or 270° clockwise or anticlockwise on the graph sheet to get image 1. Item should ask students to select a scale factor from a given range and enlarge their object to get image 2. Provide a range of numbers for the translation vector for translation. For reflection, the item should be specific on the line of reflection other than the x and y axes. For translation, range of setting translation vector should be given.

Sample

- Form any four pair of numbers each within the range of -5 to 5.
- Using an appropriate scale, plot the ordered parts.
- What is the specific name of the plane shape drawn
- Rotate your object through 90° anticlockwise about the origin to form image 1. Label your image appropriately.

- e. Using a scale factor with the range of -2 to 2 , enlarge your object to form image 2. Label your image appropriately.
- f. Reflect your object in the line $y=2$

The task is a performance assessment task because it seeks students to draw knowledge and skills from different mathematical concept and discipline. For example the aspect on reflection will expect students to apply the principles of reflection on a plane mirror. Also the task does not have a single correct answer. The correctness is depends on using the appropriate procedure. The numbers involved differ.

Question Two

Content: Statistics

Objectives: Ability to keep records of data, represent data graphically, idea of sample space, range and accurate recordings, processing of data using frequency and accurate measurement, accurate graphical representation of data, good knowledge of types of data and estimation of measures of central tendencies..

Description: The items should give range of numbers to be recorded. The numbers should continuous in nature. The range should be such it will lend itself to group data. The number of numbers should be within the range of 40 to 60. The task should students to draw a frequency table either for a histogram or cumulative curve. Measures to be estimated should the measures of central tendencies.

Sample

A teacher conducted an end of term examination and scored over 100%.

- a. Record 50 of the possible outcomes with a range of at least 90
- b. Construct a frequency table for the recorded outcomes
- c. Draw a histogram for your frequency table
- d. Estimate the mode, median and mean of your scores

The item above is a performance assessment task which requires students to apply varieties of knowledge and skills such as recording, computations of statistics, graphical representation of data and many more. The task is a construct – centred tasks which has stated the knowledge and skills students are to exhibit. It also lends itself to multiple correct responses even though the procedures are the same.

Question Three

Content: Area

Objectives: Good understanding of idea of measurement (conversion of one linear unit to the other), calculating of area, basic arithmetic (performing of operations-addition, multiplication and division) and approximations.

Description: The items should create a scenario of authentic assessment. The item should be limited to the floor of rooms of a residential house. The size of the rooms should real as much as possible. For example, a normal bedroom size ranges from 11 to 15ft. The sizes of the tiles should be as those on the market. Item should two different size of tiles for the estimation. The bedrooms and its accessories should not exceed 8.

Sample

Mr Mensah has a two-bedroom flat. The bed rooms measure 12×12 ft., the hall measures 20×25 ft., the dining hall measures 10×15 ft., the kitchen measures 10×12 ft, the washroom with toilet measures 7×5 ft and the porch measures 15×7 ft. Mr. Mensah wants to tile the floor of all the rooms. Two sets of tiles are available, one measures 50×50 mm and the other, 40×40 mm. There are 7 pieces in the box of the 50×50 mm and 15 pieces in the box of the 40×40 mm.

How many boxes of each size will be needed to finish all tiling (explain your answer either mathematical communication or verbally).

This is a task-centred item because the knowledge and skills required to be exhibited in order to accomplish the task is not stated in the task. The task also exhibit one of the key characteristics of performance assessment which authenticity. It is the expected final product that is stated.

Question Four

Content: Construction

Objectives: Good idea of types of triangles, constructing of lines and angles, bisection of angles and lines, accurate measurement of angles and lines and correct use of instruments.

Description: The items should ask students to draw a triangle with angle given. The angle should a basic angle (30° , 45° , 60° 90° and 120°). The length of the given lines should not exceed 10cm. The item should ask students the unknown lines and angles (one each). Two of the lines of the triangle should be made to be bisected to meet at a given point. The item should ask students to construct a perpendicular bisector of one point to meet the opposite line at a given point. Students should be ask to draw a circle with the meeting point of the two bisectors

as a centre and one point of the triangle as a radius. The item should ask students to name the triangle drawn and justify their students.

Sample

Using a ruler and a pair of compasses only,

- a. Construct triangle ABC such that $AB = 6\text{cm}$, angle $ABC = 60^\circ$ and $AC = 8\text{cm}$.
- b. Measure line BC and angle BCA
- c. Bisect line AC and BC to meet at P
- d. Construct a perpendicular bisector of C to meet AB at Q
- e. Using PC as a radius with P as the center, construct a circle
- f. What is the name of the triangle ABC? (justify your answer)

The task is a construct-centered task which has stated the knowledge and skills students are to exhibit to produce the product. It requires students to apply the skills and knowledge of the use of the drawing instrument. In this task, it is the process which is assessed. The task, like all performance assessment tasks, requires students to apply various knowledge and skills in and outside the content, construction, to produce the product.

Question Five

Content: Linear equation

Objectives: Ability to translate statement in symbols, Good level of English comprehension and isolating a variable from other terms.

Description: The items should be in word problem and should be limited to ages. The relationship subjects should be authentic. The resultant ages between for the subjects should also be authentic and the results should be discrete. For example it is father daughter then the ages to should real age difference of a father and

daughter. Phrases like in four years, twice the sum, two years ago are permitted. The translation should not be a simple linear equation, at least expansion or equating two binomials.

SAMPLE

- a. Assuming that your father is 25 years older you. If in six years time the sum of your ages will be 63 years, how old is your father now?
- b. Suppose further that your father is 28 years older than your sibling. If in six years time, the difference of their ages will be 7 less than twice the sum of their ages, how old is your sibling now?
- c. Between you and your sibling, who is the eldest?



APPENDIX B2-PERFORMANCE-BASED ASSESSMENT TEST**UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES****Performance-based assessment in mathematics****SHS 3****Duration 1hr 45 mins****Instructions: Answer all questions. All questions carry equal marks of 20.****Question 1**

At the wedding ceremony of Mr and Mrs Ayebine-Gyamfi, the photographer took a picture of the couples. The photographer realised that the original picture (object) lies within the range of 1 to 5 on a Cartesian plane both axes.

- a. Record four possible coordinate of the picture
- b. Using an appropriate scale, plot the ordered pairs and join the points to form a shape.
- c. What is the specific name of the plane shape drawn?
- d. Rotate your picture through 90° anticlockwise about the origin to form image 1. Label your image appropriately.
- e. Using a scale factor within the range of -2 to 2, enlarge your picture to form image 2. Label your image appropriately.
- f. Reflect your picture in the line $y=2$

Question 2

A teacher conducted an end of term examination and scored over 100% for a class of 50 students.

- a. Record the possible outcomes with a range of at least 90
- b. Construct a frequency table for the recorded outcomes

- c. Draw a histogram for your frequency table
- d. Estimate the mode, median and mean of your scores.

Question 3

Mr Mensah decided to put up a two-bedroom flat. The house has two bed rooms, living hall, dining hall, kitchen, two washrooms with toilet and a porch. The dimension of the bedrooms and kitchen are between 12-15ft, dining hall is 10-12ft, living hall is 25 -30ft, washroom with toilet is 5-7ft and the porch 7-12 ft. Mr. Mensah wants to tile the floor of all the rooms. Two sets of tiles are available, one measures 50×50 mm and the other, 40×40 mm. There are 7 pieces in the box of the 50×50 mm and 15 pieces in the box of the 40×40 mm.

Choose an appropriate dimension of each room within the dimensions given, find how many boxes of each size will be needed to finish all tiling (explain your answer in either mathematical or everyday English).

Question 4

There is a-three- sister communities in the Ahanta West District of the Western, Himakrom, Bonsokrom and Npanyinasa. The distance from Himakrom to Bonsokrom is 2km, the distance of Npanyinasa from Himakrom is 1600m. The bearing of Npanyinasa from Bonsokrom is 300° . The Municipal Assembly intended to build a school for the three communities so that the school will be equidistant from the communities.

Using a ruler and a pair of compasses only,

- a. Make a geometric construction of the communities and where the school will be situated.
- b. What is the distance of the school to Bonsokrom?

- c. What is the distance of Bonsokrom from Npanyinasa?
- d. What is the specific name of the shape formed by the position of the communities? (Justify your answer)

Question 5

- a. Assuming that your father is 25 years older than you. If in six years time the sum of your ages will be 63 years, how old is your father now?
- b. Suppose further that your father is 28 years older than your sibling. If in six years time, the difference of their ages will be 7 less than twice the sum of their ages, how old is your sibling now?
- c. Between you and your sibling, who is the elder?



**APPENDIX B3-SCORING RUBRIC FOR THE PERFORMANCE BASED
ASSESSMENT ITEMS**

General Instructions

1. When a student misses an **M** mark, the preceding **M** or **A** marks are scored zero.

Question One

- a. B2 for 4 pairs correct (- ½ for each number outside the range)
- b. B1 for using scale that covers all points

B1 for labelling. One is implied

B1 for calibration (½ for each axis) - ½ error, once on each axis

B2 for plotting of points (- ½ each error). NB. Errors include none joining, plotting wrong of points, none labelling, not writing coordinates of point

- c. B1 for name the shape - ½ without justification or incorrect justification

- d. B3 for draw image under 90 °clockwise rotations. (- ½ each error).

NB. Errors include none joining, plotting wrong of points, none labelling , not writing coordinates of point

- e. B1 for using factor within the range

B3 for using (his) scale factor to draw enlargement of the object. (- ½ each error). NB. Errors include none joining, plotting wrong of points, none labelling, not writing coordinates of point

- f. B3 for drawing reflection in the line $y=2$ (- ½ each error) NB. Errors include none joining , plotting wrong of points, none labelling , not writing coordinates of point

Question Two

- a. B2 for all scores within the range (- ½ each score outside the range of 90)
- b. B1 for using group frequency table such that there are not more than 10 units under marks/scores

A1 for any 3 frequencies correct,

A2 for all frequencies correct (- ½ each error)

- c. B1 for using class boundaries

B1 for labelling scores/marks and frequencies. One is implied (- ½ each error such as incorrect calibration)

B3 for correct graph (- ½ each error). NB. Errors include omission of zigzag, wrong frequency from table, uniform bar size

- d. *For Mode*

B1 for finding Δ_1 and Δ_2 (½ for each)

M1 for substitution all values correctly into the formula

A1 for correct answer estimated to the nearest whole number as per the scores

For median

M1 for substituting all values correctly into the formula

A1 for correct answer estimated to the nearest whole number as per the scores

For mean

A1 for any 3 fx correct,

A1 for all frequencies correct (- ½ each error)

M1 for substituting all values correctly into the formula

A1 for correct answer estimated to the nearest whole number as per the scores

Question Three

A6 for areas of rooms (-1 for each error. ie correct area estimated by multiplying dimensions)

B1 for area of each tiles (for both ½ for each by multiplying)

B2 for either converting areas of room in ft to mm or converting dimension of tiles in mm to ft.

B4 for number of pieces of each tile size by dividing area of room by area of tiles (-2 without showing or stating division, one for each)

B4 for total number of pieces by summing up pieces of each room (-2 without showing or stating the summation)

B3 for number of boxes by dividing total number of pieces of each size by 7 and 15 (-1 without showing or stating division).

Aliter

A6 for areas of rooms (-1 for each error. ie correct area estimated)

B1 for area of each tiles (for both $\frac{1}{2}$ for each)

B4 for total area of by summing up areas of each room (-2 without showing or stating the summation

B2 for either converting areas of room in ft to mm or converting dimension of tiles in mm to ft.

B4 for number of pieces of each tile size by dividing total area of rooms by area of each tiles (-2 without showing or stating division, one for each)

B3 for number of boxes by dividing total number of pieces of each size by 7 and 15 (-1 without showing or stating division).

NB. Accept any other correct procedure.

Question Four

B1 for line $AB = 6\text{cm}$

B1 for 60° at B (- $\frac{1}{2}$ without arc constructed)

B1 for line drawn with straight edge

B1 for constructing arc of 8cm

A1 for completing triangle (- $\frac{1}{2}$ without showing C)

B1 for correctly measuring line BC (accept ± 1)

B1 for correctly measuring angle BCA (accept ± 1)

B2 for bisecting AC (- $\frac{1}{2}$ for each arc not constructed or seen) B1 for arcs B1 for line

B2 for bisecting BC (- ½ for each arc not constructed or seen) B1 for arcs B1 for line

B1 for locating P

B1 for arcs on line AB from point C - ½ for each arc not constructed or seen)

B1 for intersecting arcs from arcs on line AB (-1 for freehand sketch of arc)

A1 for constructing perpendicular line from C

B1 for using P as center (-1 for C as center)

B1 constructing circle

A2 for touching all points of the triangle (-1 each error)

A1 for type of triangle (-1 for not justifying answer)

Question Five

a. B2 for stating age of father and daughter now – ½ for wrong or missing symbol/error

B2 for stating age of father and daughter in six years time – ½ for wrong or missing symbol/error

B1 for summing (his) ages in six years and equating to 60. ½ for sum and ½ equating to 60

M1 for solving that is isolating variable

A1 for correct answer

A1 for age of the father

b. B2 for stating age of father and daughter now – $\frac{1}{2}$ for wrong or missing symbol/error

B2 for stating age of father and daughter in six years time – $\frac{1}{2}$ for wrong or missing symbol/error

B1 for difference (his) ages in six years.

B1 for summing up twice of ages (his) in six years time

B1 for equating difference of ages to sum of ages in six years time

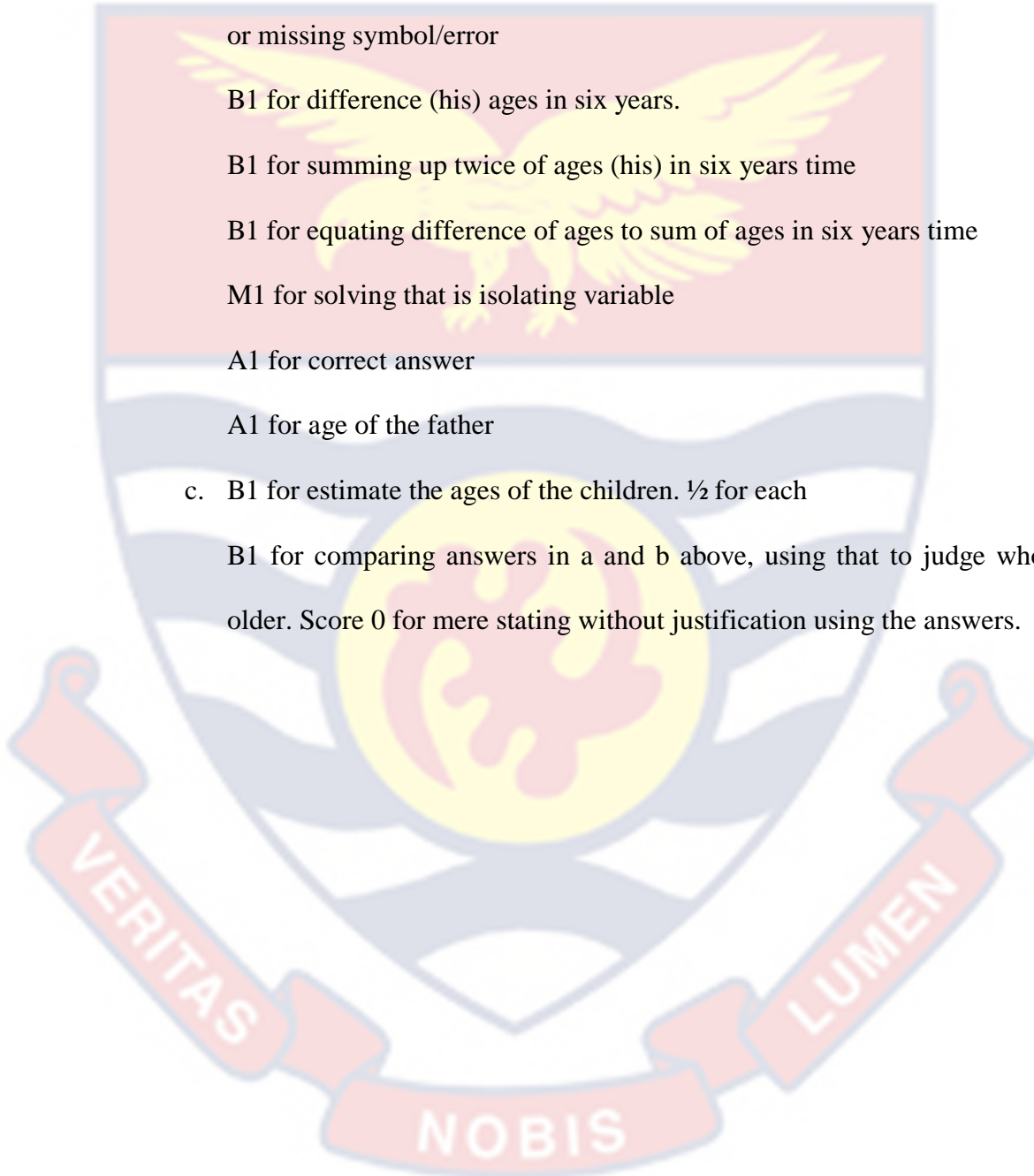
M1 for solving that is isolating variable

A1 for correct answer

A1 for age of the father

c. B1 for estimate the ages of the children. $\frac{1}{2}$ for each

B1 for comparing answers in a and b above, using that to judge who is older. Score 0 for mere stating without justification using the answers.



APPENDIX C-QUESTIONNAIRE FOR TEACHERS AND EXAMINERS**UNIVERSITY OF CAPE COAST****COLLEGE OF EDUCATION STUDIES**

Respondent's Consent:

The purpose of this questionnaire is to seek information to validate performance-based assessment for Senior High Schools. Your full participation will help make informed decisions about the assessment and learning of Mathematics in the Senior High Schools. It would therefore be appreciated if you could frankly provide responses to **all** items on the questionnaire.

You are assured of complete **confidentiality** and **anonymity** of all information provided. Your participation in this study is **completely voluntary**. However, your participation is very much appreciated and will assist in the education process of your district, and Ghana as a whole.

Please tick the appropriate response to answer this questionnaire to the best of your knowledge.

Instructions

Study the Performance-based assessment Questions attached to respondent to items on Section C to G

Questionnaire for teachers and examiners**Section A**

1. Gender
 - a) Male []
 - b) Female []

2. Marking Experience

a. 1-5 years []	b) 6-10years []
c) 11-15years []	d) 16- 20years []
e) above 20 years []	

3. School.....

Section B

**FEASIBILITY OF THE IMPLEMENTATION OF PERFORMANCE-
BASED ASSESSMENT**

Indicate by ticking [√] your level of agreement on the following activities regarding **Feasibility of the implementation of performance-based assessment for WAEC examinations**. Where: **SA = Strongly Agree**, **A = Agree**, **D = Disagree**, and **SD = Strongly Disagree**

S/N	Item	SA	A	D	SD
1	Marking of script will comparatively be of the same time as the traditional system				
2	Same number of scripts could be marked in PBA as in the traditional system could be marked by an examiner				
3	Scripts marking of PBA will be of the same difficulty as the traditional				
4	Same number examiners for the traditional system could finish marking the PBA items				
5	Constructions of the PBA items will not be difficult just like the traditional system				
6	Construction of alternate forms PBA is be feasible				
7	With a well designed test specifications, alternate forms can be created				
8	Item constructions of PBA will require much time and skills				
9	PBA Will be able to cover all content learned in a single test				
10	Student could be assessed wirh PBA within the allotted time				

11	Materials for using PBA for examinations are available				
12	Use of PBA would not produce extra cost to the assessment system				
13	PBA is practicable for a large number of examinees				

Section C

EDUCATIONAL EFFECTS OF PERFORMANCE-BASED ASSESSMENT

Indicate by ticking [] your level of agreement with the following items regarding

Effect of performance-based assessment on student's learning. Where: SA =

Strongly Agree, A = Agree, D = Disagree, and SD = Strongly Disagree

S/N	Item	SA	A	D	SD
1	Students will be compelled to learn				
2	Students are motivated to learn				
3	Encourages students to think differently on an issues				
4	Causes students to think critically on problems				
5	Encourages students to learn extensively				
6	Makes learning easier				
7	Encourages learning in every domain				
8	Can be integrated into the teaching and learning processes				
9	Encourages learning of mathematical skills				

Section D**CATALYTIC EFFECTS OF PERFORMANCE-BASED ASSESSMENT**

Indicate by ticking [] your level of agreement on the following activities regarding **examination provides feedback that stimulates learning**.

Where: **SA = Strongly Agree, A = Agree, D = Disagree, and SD = Strongly Disagree**

S/N	Item	SA	A	D	SD
1	Immediate feedback can be given to students				
2	Reveals students' true performance				
3	Reveals areas of students' strength and weakness on each aspect of content learned				
4	Students will be able to reflect on their performance				
5	All domains of learning are assessed				
6	Makes learning individualistic				
7	Could be used in the classroom to give prompt feedback to students				
8	Measures diversity of behaviour				

Section E**CREDIBILITY OF PERFORMANCE-BASED ASSESSMENT RESULTS**

Indicate by ticking [] your level of agreement with the following items regarding

Different stakeholders find the examination processes and the results credible. Where: **SA = Strongly Agree, A = Agree, D = Disagree, and SD =**

Strongly Disagree

S/N	Item	SA	A	D	SD
1	Results reflect students true performance				
2	Malpractice associated with examination is reduced				
3	The results can be trusted				
4	Differences in students performance become real				
5	Knowledge level becomes the same as application level				
6	Provides accurate estimation of student performance				
7	Results could be generalized				

Thank you so much for support

APPENDIX D- EXPLORATORY AND CONFIRMATORY FACTOR

ANALYSIS

Comp	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	7.091	30.831	30.831	7.091	30.831	30.831	6.250
2	4.440	19.304	50.135	4.440	19.304	50.135	3.270
3	3.550	15.435	65.570	3.550	15.435	65.570	3.258
4	1.842	8.008	92.411	1.842	8.008	92.411	3.285
5	.877	3.811	96.222				
6	.744	3.235	99.457				
7	.125	.543	100.000				

Item	Component			
	1	2	3	4
Feasibility				
Marking of script will comparatively be of the same time as the traditional system	.747			
Same number of scripts could be marked in PBA as in the traditional system could be marked by an examiner	.747			
Scripts marking of PBA will be of the same difficulty as the traditional	.733			
Same number examiners for the traditional system could finish marking the PBA items	.724			
Constructions of the PBA items will not be difficult just like the traditional system	.707			
Construction of alternate forms PBA is be feasible	.683			
With a well-designed test specification, alternate forms can be created	.683			
Item constructions of PBA will require much time and skills	.674			
PBA Will be able to cover all content learned in a single test	.663			
Student could be assessed with PBA within the allotted time	.575			
Materials for using PBA for examinations are	.538			

available

Use of PBA would not produce extra cost to the assessment system .537

PBA is practicable for a large number of examinees .527

Credibility

Results reflect students' true performance .733

Malpractice associated with examination is reduced .724

The results can be trusted .707

Differences in students' performance become real .687

Knowledge level becomes the same as application level .678

Provides accurate estimation of student performance .666

Results could be generalized .651

Educational effect

Students will be compelled to learn .834

All domains of learning are assessed .683

Encourages students to think differently on an issue .683

Causes students to think critically on problems .683

Encourages students to learn extensively .654

Lessons reflect real life experience .612

Encourages learning in every domain .543

Can be integrated into the teaching and learning processes .515

Encourages learning of mathematical skills 515

Catalytic effect

Immediate feedback can be given to students .931

Reveals areas of students' strength and weakness on each aspect of content learned .931

Students will be able to reflect on their performance .931

Students are motivated to learn .733

Makes learning individualistic .724

Could be used in the classroom to give prompt feedback to students .707

Measures diversity of behaviour .618

APPENDIX E-INTRODUCTORY LETTER

UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
FACULTY OF EDUCATIONAL FOUNDATIONS
DEPARTMENT OF EDUCATION AND PSYCHOLOGY

Telephone: 0332091697
Email: dep@ucc.edu.gh



UNIVERSITY POST OFFICE
CAPE COAST, GHANA

Our Ref:

18th February, 2020

Your Ref:

TO WHOM IT MAY CONCERN

Dear Sir/Madam,

**THESIS WORK
LETTER OF INTRODUCTION
MR. ABRAHAM GYAMFI**

We introduce to you Mr. Gyamfi, a student from the University of Cape Coast, Department of Education and Psychology. He is pursuing Doctor of Philosophy Degree in Measurement and Evaluation and he is currently at the thesis stage.

Mr. Gyamfi is researching on the topic:

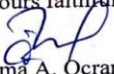
“APPLICATION OF GENERALIZABILITY AND ITEM RESPONSE THEORIES IN THE DEVELOPMENT AND VALIDATION OF PERFORMANCE-BASED ASSESSMENT IN MATHEMATICS FOR SENIOR HIGH SCHOOLS IN GHANA.”

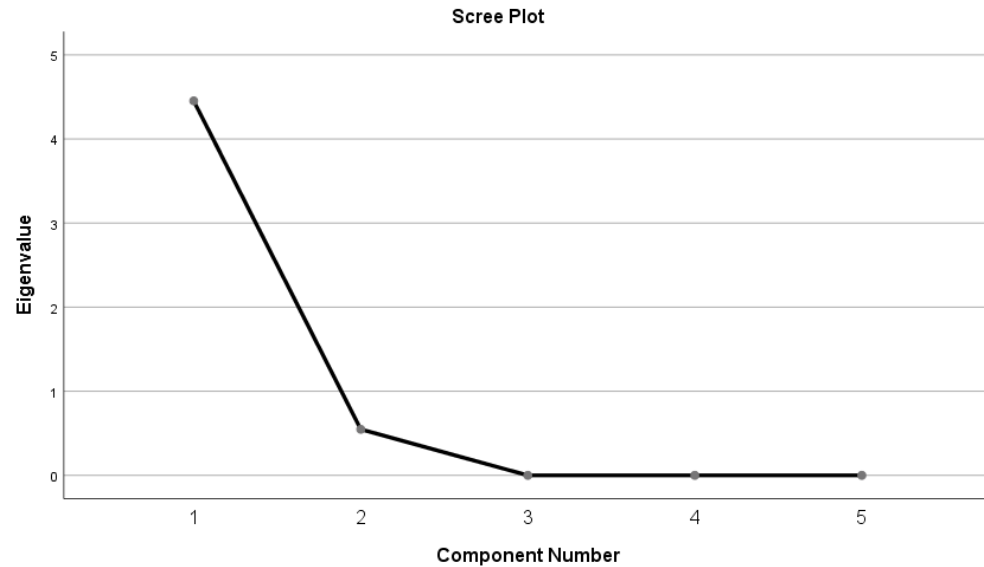
He has opted to collect data at your section for his Thesis work. We would be most grateful if you could provide him the opportunity and assistance for the study. Any information provided would be treated strictly as confidential.

We sincerely appreciate your co-operation and assistance in this direction.

Thank you.

Yours faithfully,


Ama A. Ocran (Ms.)
Principal Administrative Assistant
For: **HEAD**

APPENDIX F-UNIDIRECTIONALITY AND LOCAL INDEPENDENCE**ASSUMPTIONS OF THE PERFORMANCE BASED ASSESSMENT***Scree Plot for the Items**-Eigenvalues of Total Variance Explained*

Comp.	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cum. %	Total	% of Variance	Cum. %
1	4.453	89.058	89.06	4.453	89.058	89.058
2	.547	10.942	100.00			
3	9.890E-16	1.978E-14	100.00			
4	7.618E-16	1.524E-14	100.00			
5	-1.479E-16	-2.958E-15	100.00			

Source: Gyamfi (2020)

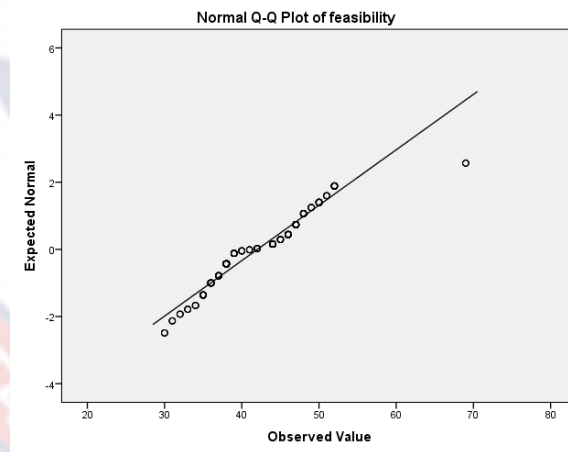
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.544
Bartlett's Test of Sphericity	Approx. Chi-Square	284.226
	df	10
	Sig.	.000

APPENDIX G1-HOMOGENITY AND LEVENE TEST OF FEASIBILITY*Test of Normality of Scores of the Feasibility of the PBA*

Variable	levels	Shapiro-Wilk		
		Statistic	df	Sig.
Gender	Male	.912	314	.070
	Female	.859	76	.061
Status	Examiner	.901	150	.080
	Teacher	.929	150	.083
School Category	Cat A	.846	130	.001
	Cat B	.902	130	.231
	Cat C	.934	130	.090
Experience	1-5yrs	.804	147	.004
	6-10yrs	.926	125	.052
	11-15yrs	.836	73	.070
	16-20yrs	.484	21	.072
	> 20yrs	.825	24	.080

Source: Field data (2020)

Q-Q plot of Normality of Scores of the Feasibility of the PBA*Test of Equality of Variances of the Feasibility of the PBA*

	F	df1	df2	Sig.
Feasibility	3.964	37	352	.070

Source: Field data (2020)

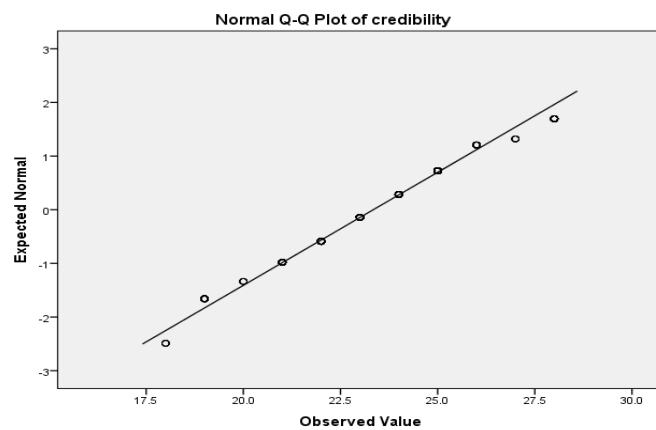
APPENDIX G2-HOMOGENITY AND LEVENE TEST OF CREDIBILITY

Test of Normality of Scores of the Credibility of the PBA

Variable	levels	Shapiro-Wilk		
		Statistic	Df	Sig.
Gender	Male	.912	314	.070
	Female	.859	76	.061
Status	Examiner	.901	150	.080
	Teacher	.929	150	.043
School	Cat A	.846	130	.101
Category	Cat B	.902	130	.231
	Cat C	.934	130	.090
Experience	1-5yrs	.804	147	.074
	6-10yrs	.926	125	.052
	11-15yrs	.836	73	.070
	16-20yrs	.484	21	.002
	> 20yrs	.825	24	.080

Source: Field data (2020)

Q-Q Plot of Normality of scores of the credibility of the PBA



Test of Equality of Variances of the Credibility of the PBA

	F	df1	df2	Sig.
credibility	5.496	37	352	.000

Source: Field data (2020)

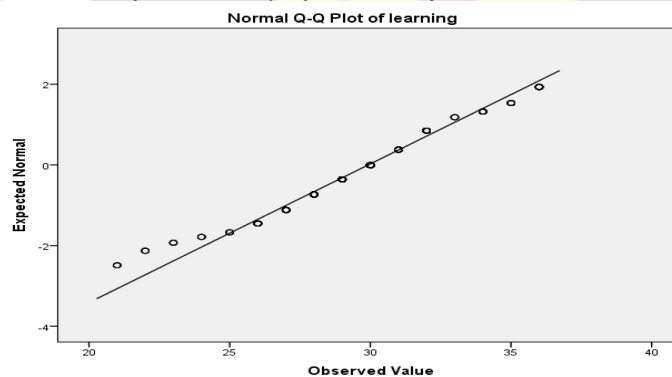
**APPENDIX G3-HOMOGENITY AND LEVENE TEST OF
EDUCATIONAL EFFECT**

Test of Normality of Scores of Educational Effect of the PBA

Variable	levels	Shapiro-Wilk		
		Statistic	df	Sig.
Gender	Male	.912	314	.070
	Female	.859	76	.061
Status	Examiner	.901	150	.080
	Teacher	.929	150	.083
School	Cat A	.846	130	.101
Category	Cat B	.902	130	.231
	Cat C	.934	130	.090
Experience	1-5yrs	.804	147	.074
	6-10yrs	.926	125	.052
	11-15yrs	.836	73	.070
	16-20yrs	.484	21	.072
	> 20yrs	.825	24	.080

Source: Field data (2020)

Q-Q Plot of Normality of Scores of the Educational Effect of the PBA



Test of Equality of Variances of Educational Effect of the PBA

	F	df1	df2	Sig.
Educational effect	5.742	37	352	.673

Source: Field data (2020)

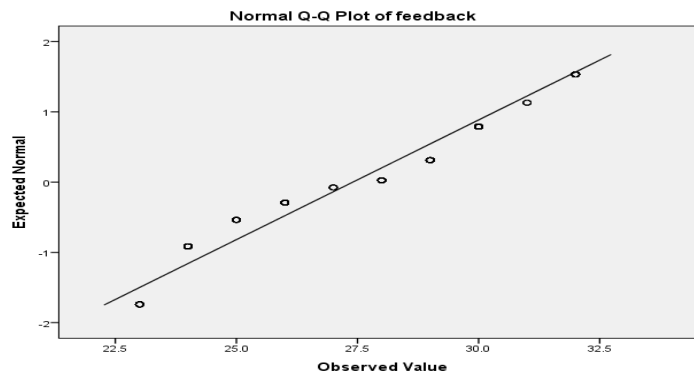
APPENDIX G4-HOMOGENITY AND LEVENE TEST OF CATALYTIC EFFECT

Test of Normality of Scores of Catalytic Effect of the PBA

Variable	Levels	Shapiro-Wilk		
		Statistic	df	Sig.
Gender	Male	.912	314	.770
	Female	.859	76	.061
Status	Examiner	.901	150	.080
	Teacher	.929	150	.343
School	Cat A	.846	130	.101
Category	Cat B	.902	130	.231
	Cat C	.934	130	.190
	Experience	1-5yrs	.804	147
	6-10yrs	.926	125	.052
	11-15yrs	.836	73	.170
	16-20yrs	.484	21	.002
	> 20yrs	.825	24	.080

Source: Field data (2020)

Q-Q Plot of Normality of Scores of the Catalytic Effect of the PBA



Test of Equality of Variances of Catalytic Effect of PBA

	F	df1	df2	Sig.
Catalytic effect	4.541	37	352	.213

Source: Field data (2020)