International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 3, May-Jun 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Questionnaire in the Measurement of Mathematics Achievement? An Application of Bifactor Modeling in Assessing Criterion Validity

Kesiki, Samuel[1]; Nekang, Fabian Nfon[2]

[1]Post Graduate Student, Department of Curriculum Studies and Teaching,
[2]Department of Curriculum Studies and Teaching,
[1,2]Faculty of Education, University of Buea, Buea, Cameroon

ABSTRACT

The concepts of mathematics performance and mathematics achievement have often been taken to mean the same thing. Though mathematics performance has been known to predict achievement to a certain degree, the extent to which student-reported outcome measures predict or substitute standardized tests in the measurement of mathematics learning outcomes has been given very limited attention. The present study utilized a cross-sectional survey research design on a sample consisting of 206 first cycle students of secondary school, and applied a bifactor analysis model, to investigate if there was a common trait of mathematical ability measured across all three-mathematics learning outcome sub-scales utilized in the present study. Additionally, the study investigated if self-assessment scores, together with mathematics performance scores, and achievement tests scores provide the ultimate criterion for the assessment of mathematics learning outcomes. Findings revealed that; mathematics performance and perceived mathematics performance differentially predicts mathematics achievement while controlling for mathematics self-efficacy, and mathematics self-concept. Moreover, of all three bifactor models hypothesized in the study, the augmented oblique bifactor model provided the best fit for the data. Calculated bifactor statistical indices placed the relative bias of standardized bifactor loadings above the 10% benchmark revealing that some multidimensionality was severe enough to disqualify the interpretation of the instrument as primarily unidimensional. Though no significant evidence was found in support of the fact that a questionnaire on student-reported mathematics outcome measures possesses predictive criterion validity, findings however revealed that the aggregated scores from all three subscales explored in the study provided a much realistic criterion for measuring mathematics learning outcomes. Furthermore, students perceived or self-rating of their mathematical abilities on specific math tasks were grossly exaggerated compared to their actual or demonstrable mathematical abilities. In spite of the fact that no direct relationship was found, mathematics performance had an indirect effect on mathematics achievement through mathematics self-efficacy, and self-concept. Findings have implications for the modern measurement theory. It was recommended that standardized mathematics tests should not be substituted with neither questionnaire measures nor with class tests or school-based tests in the measurement of mathematics achievement, given that each subscale was shown to be a reliable-enough measure of their individual constructs, but that in order to improve students' performances in the subject the aggregation of subscale scores could be considered.

***How to cite this paper**:* Kesiki, Samuel | Nekang, Fabian Nfon "A Questionnaire in the Measurement of Mathematics Achievement? An Application of Bifactor Modeling in Assessing Criterion Validity" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-3, June 2025, pp.721-744, URL: www.ijtsrd.com/papers/ijtsrd79998.pdf

IJTSRD79998

***KEYWORDS:** Perceived Mathematics Performance, Mathematics Performance, Mathematics Achievement, Questionnaire, Bifactor Model*

INTRODUCTION

Since professional mathematicians first expressed concerns about the process of teaching and learning in mathematics around the early twentieth century (Furr, 1996; Kilpatrick, 2014), teachers and researchers working in cooperation with scholars from other disciplines, have struggled to improve students' performances in the discipline (Manizade et al., 2023). Hitherto, research efforts have mostly been focused on designing suitable instruments that effectively assess students' learning of mathematics (Suurtamm et al., 2016; Yantini et al., 2021; Yudha et al., 2019). Students' performances in mathematics especially regarding the ability to demonstrate the mastery of core competencies have continued to be a subject of intense debate (Anqui-Laja & Laja, 2022; Reusser, 2000). In particular, issues surrounding the effectiveness of traditional assessment methods and the relevance of classroom learning of mathematics in the solution of real-life problems have constituted the focus of this debate (Arthur et al., 2018; Ernest, 2015; Vos, 2005). In response, many researchers have rolled out guidelines on how to tackle the many unintended learning outcomes of mathematics learning including personal, curricular, and environmental difficulties in learning the discipline (Arthur et al., 2017; Febriyanti et al., 2021; Kesiki & Nekang, 2023). In addition, beside the traditional achievement test which measures students' content knowledge and provide scores used in important decision making, researchers are experimenting with subjective assessment instruments like self-reported (self-assessment) tests which are aimed first at fostering the development of students' self-directed learning abilities (Gruppen, n.d; Max et al., 2022), and secondly by addressing attitudes known to be the main antecedents of students' unintended learning in mathematics (Chamberlain, 2010; Code et al., 2016; Gyamfi, 2022). According to Orrill et al. (2023), this paradigm shift in the assessment of mathematics learning is consistent with the emergence of new theories of learning (socio-cultural and critical theories) and new definitions of what constitutes learning. The present study investigates the reliability of students' subjective mathematics performance (self-assessed scores) as the main evidence for basing important decision-making compared to objective teacher/criterion-oriented test scores.

In psychometrics, the effective measurement of latent constructs warrants that, not only should the appropriate test be used, but that the test be reliable, bias free, and associable with the standard (Borsboom & Molenaar, 2015; Boyle et al., 2015; Kyriazos & Stalikas, 2018). Different tests measure learners' abilities at different domains of learning (Cliff & Yeld, 2006). While direct test such as achievement test measure knowledge at the cognitive domain, indirect test such as self-report inventories (questionnaires) measure attitudes and perceptions at the affective domain (Oakland, 1997; Saftari & Fajriah, 2019; Sayegh, n.d). For researchers concerned with the measurement of mathematics performance and mathematics achievement (Antara et al., 2020; Grootenboer et al., 2015; Learning Outcome Framework., 2015), the contention has always been whether to utilized scores from either teacher-made test (which are narrower in scope), and standardized test (which are better adapted for general usage, has a variety of forms, and test on a wide range of competences), or scores from student-reported tests in the measurement of the constructs (Qassimi, 2021; Reys & Rea, 1970; Sarifah et al., 2024). While the former is recommended in the measurement of both mathematics performance and achievement, the later has however found broad usage in research studies lately. Though self-report tests are considered to be suitable in the measurement of learning outcomes in general since they mostly measure psychological processes driving human learning (Lyonga, 2022; Peckrun, 2020), their use in the assessment of mathematics learning outcomes in particular (which are better assessed with direct tests) presents challenges due to a missing 'one-size-fit-all' design for self-report test (Niss et al., 1998; Radišić, 2023). However, the extent to which scores from meticulously designed and validated self-report tests predict scores from standardized mathematics test has not been given attention in the context of Cameroon. Moreover, the extent to which scores from mathematics self-report tests and teacher-made tests jointly predict standardized mathematics achievement test scores has received limited attention as well. The present study also addresses two levels of assessment (ordinal and interval) and tries to draw the link between them.

Historically, self-report tests (questionnaire measures) were first developed and utilized in the comparison of experiences and attitudes of different groups of people in the late 19th century by Francis Galton (Horvat, 2014; Midena & Yeo, 2022). Specifically, in 1817, Marc Antoine Jullien de Paris designed a 34-page international survey of national education systems (Creswel, 2012). During the period between the two world wars, modern surveys began to emerge. Today, in the 21st century remarkable progress has been made in self-report tests, with regards to their design, development, administration, analysis, and interpretation (Willis, 2019). According to the National Education Association (NEA, 2020), standardized testing on the other hand, started being utilized in America in 1838, and has since been adopted in different parts of the world as the most reliable tool for assessing learning and learning outcomes. Standardized testing was first introduced in English Cameroon in 1977 by the University of London Examination and Assessment Council, and was popularly known as the London GCE (The Cameroon GCE Board, 2023). When the University of London Examination and Assessment Council withdrew from Cameroon in 1990, the Cameroon GCE board was created in 1991 as a replacement body in charge of the organization of final year examinations for the Ordinary and Advanced levels, and later on started organizing certificate examinations in technical, and vocational education as well.

Conceptually, self-report tests measure non-academic competencies and are standardized such that, they carry the same sets of questions, same response format, and a uniform style in which responses are recorded (Boynton & Greenhalgh, 2004; He & Van de Vijver, 2019). Self-report tests are designed differently depending on the objectives they seek to achieve (Demetriou et al., 2015; Korb, 2011). From closed-ended questionnaires which range from just two response options (dichotomous), and Likert-scales with response categories which lie in a continuum between polar opposites; to open-ended questionnaires which request for more insight on a particular measured variable (Hyman & Sierra, 2016; Saris & Gallhofer, 2014). There are basically two types of self-report tests; self-administered and researcher-administered questionnaires (Robins et al., 2007; UNESCO, 2018). While the former can be completed by pencil and paper or online, the later require face-to-face discussions which can take place physically or virtually. Self-report test can lead to the collection of nominal, discrete (requiring students to count e.g number of cars passing on a road in a day), or ordinal data (Learning Outcome Framework, 2015). It is worthy to note that in the present context self-report test simply refer to ordinal data (5-point Likert scales).

Theoretically, academic and non-academic tests (cognitive or affective) are generally required to follow strict guiding principles in their construction, validation, administration, analysis, and interpretation (Vos, 2005). In the design of self-administered test however, the problem often primarily revolves around the language used (choice of item phrases and framing) and issues surrounding convergence of test measures or the spatial arrangement of information within the questionnaire (Jenkins & Dillman, 1995). In addition, the design has to provide evidence that the test not just only measures what it intends to measure but does so consistently (Kubai, 2019; Taherdoost, 2016). In other words, the construct dimension and the scale dimension (internal structure) of the test has to be ascertained, so as to correctly inform decision making (Yusof et al., 2021). Moreover, random errors which ensue from conditions of testing should be minimized during administration (Platek, 1985). Furthermore, the models for linking the associating constructs and the tools for computing the datasets has to be homogenous (there should be the possibility for mathematical analysis of data, and it should be guided by same rules and constraints) and to conform to the standards. Finally, the inferences arrived at from the evidence should be guided by theory (Hawkin et al., 2020).

Psychometric and personality tests are generally concerned with the measurement of the abilities and personality traits of individuals respectively (Boyle & Saklofske, 2004; Kyllonen & Kel, 2018). Three main theories guide the construction of psychometric and personality tests (Schuwirth & Van der Vleuten, 2012). The classical test theory (CTT), the item response theory (IRT), and the modern measurement theory (MMT). In CTT, the aggregate of an individual's responses to test items fully demonstrates the individual's ability, and every test-taker's test score (observed) has a true score and a certain degree of error due to biases and other factors (Cappelleri et al., 2015; Schuwirth & Van der Vleuten, 2012). The IRT on the other hand assumes that an individual's response to a particular test item is associable with some personality trait or ability and answers to items are chosen on a continuum (Mahmud, 2017; Yang & Kao, 2014). Consequently, a single type of test will suffice in the effective measurement of the construct, and when more people answer the test items, a much clearer picture of the behaviour is gotten. The MMT utilizes the IRT model in combination with Rasch's model (Cavanagh & Sparrow, 2010; Sarifah et al., 2024) and purport's that the probability that a test-taker gets the correct answer on an item is dependent on the test-taker's ability and on the item's difficulty. In addition, MMT assumes that a given construct could have multiple dominant traits other than the one-dimensional trait assumed in the IRT. The MMT therefore require different types of not just objective but also subjective (academic and non-academic) tests to be administered for a construct to be sufficiently measured. In the context of the present study, while the CTT provided a better framework for interpreting mathematics achievement scores, the IRT and the MMT on the other hand provided a much reliable framework for interpreting results from the questionnaire on mathematics achievement.

Contextually, surveys and other subjective tests are very popular methods of data collection in the Cameroon research milieu and have been widely utilized in studies as a primary data collection tool. This is due to the fact that they are relatively cheap to produce and distribute, and secondly because secondary data collection in the entire Country is still riddled with administrative bottlenecks. Consequently, in the face of such challenges, constructs such as mathematics performance and achievement which require standardized direct measurement tools and resources in their effective measurement pose a headache. Standardization is systematic and demands resources which makes it difficult for many researchers to develop these instruments using their limited personal resources. As such, the use of indirect measurement tools presents an easy escape, and for that reason are increasingly gaining consideration in the place of direct measurement tools. As evidence, a number of studies have utilized self-report tests in the measurement of mathematics achievement, without clear proof of unrefuted measurement quality of the instruments used in terms of instrument validation, and reliability/stability in the face of such considerations.

Statement of the Problem

In correlational studies that model the association of psychological variables, usually between one or more predictor variables and a response variable, the use of standard measurement instruments or substitutes with demonstrable convergent validity power have often been recommended in their effective measurement. Upholding higher measurement quality standards in the measurement of unobserved constructs, ensures that third variables (covariates, extraneous, and confounding variables) which have the ability to influence the outcome of measured variable can be effectively controlled (Suresh, 2017). To guarantee quality in the measurement of mathematics achievement in studies, often requires that a standardized written test be completed by the participants for the effective measurement of respondents' mathematical content knowledge, skills, and abilities (Vos, 2005). Elsewhere, in cases where teacher-made tests are utilized, there is always need to show that scores from such instruments sufficiently provide evidence of criterion validity, judged from how well such scores compare with those from the criterion variable concurrently or in the future (Bellamy, 2015).

Recently, however a trend is emerging in social science studies in which researchers (especially beginning researchers and graduate students) are increasingly utilizing scores from self-reported, sample-based instruments (ordinal scale), for example, questionnaire measures on students' abilities in solving specific mathematical tasks in the measurement of mathematics achievement in different research endeavours and inquiries. The points of contention are that; studies in which self-reported items have been utilized in the measurement of mathematics achievement, have failed to provide convincing evidence of effective construct validation on the one hand, and scale validation on the other hand. In addition, some studies have failed to clarify the construct actually being measured, perceived mathematics performance (performance behaviour or attitudes) or mathematics performance (knowledge, skills on a limited number of mathematics measures), given that such a distinction will obviously translate into the type of datasets to be obtained. Moreover, even in situations where performance behaviour is measured, items have often captured general mathematics abilities, rather than testees abilities in specific mathematics topics, and often no efforts are made in the control of covariates. The consequences are that if the situation is not addressed, novice researchers in particular will continue to utilize instruments of doubtful quality or unsuitable substitutes in the measurement of mathematics achievement. The research problem addressed in this study is twofold: assessing the conditions under which self-assessment instruments (questionnaires measuring mathematics performance behaviour), teacher-made tests measuring mathematics performance, and standardized mathematics tests measuring mathematics achievement, would actually measure a common mathematical ability. Furthermore, the study explores the conditions under which questionnaires, teacher-made tests, and mathematics achievement tests together would provide a much more holistic measurement of mathematics learning outcomes.

The specific research objectives of the study were to determine if;

1. Student-reported mathematics outcomes, and mathematics performance differentially predict mathematics achievement.
2. Student-reported mathematics outcomes, mathematics performance, and mathematics achievement all measure a common ability.
3. Student-reported mathematics outcomes, mathematics performance, and mathematics achievement together provide a better measure of mathematics learning outcomes.

The research objectives for the study were transformed into statistical hypotheses as follows;

H_{01}: Student-reported mathematics outcomes, and mathematics performance do not significantly differ in their prediction of mathematics achievement measures.

H_{02}: There is significantly no common ability measured across all three mathematics learning outcome subscales.

H_{03}: Student-reported mathematics outcomes, mathematics performance, and mathematics achievement together do not significantly provide a holistic measurement of mathematics learning outcomes

The following observations were made as a justification to the present study. According to Manizade et al. (2023, p. 198), "both affective and self-belief constructs may be specified as learning outcomes, given that competent participants within a field also hold certain beliefs about the field itself". It is on this basis that a self-report (self-assessment) test was approached as a 'could be' credible substitute instrument for mathematics achievement in the present study. Consequently, in some studies, we have witnessed indirect measurement instruments utilized in the measurement of mathematical content knowledge, skills and abilities (Buchwald & Schwarzer, 2010). Additionally, in situations where indirect measures have been utilized in the measurement of knowledge, skills and abilities, general measures of perceived mathematics performance rather than measures of perceived mathematics abilities on specific topics have been utilized, further putting in doubt the validity and reliability of such measurements. Moreover, several studies have utilized scores from teacher-made test in measuring mathematics achievement (Maree et al., 2006; Montanya, 2018; Ndlovu, 2017), and in a few other cases, teachers and some stakeholders (school administrators, parents) have been required to assessed students' learning by rating the quality and quantity of their learning (Schunk, 2012). Furthermore, a few studies have requested students to report their grades in previous performances (Bieleke et al., 2022; Mazana et al., 2019). Though the use of the above-mentioned instruments in the measurement of mathematics achievement are fair to an extent, their use especially in situations requiring subjective assessments of content knowledge must however be supported with convincingly best measurement quality practices which demonstrate the degree to which such instruments compare to the standards. In contrast to these however, the situation is different in some studies, in which a clear distinction is made between the measurement of mathematics performance and 'perceived mathematics performance' or attitudes (Ngeche, 2017; Saini & Arora, 2017).

It was projected that findings from the present study could be significant to a number of stakeholders including social science researchers in general and mathematics education researchers in particular, and in improving the measurement quality of latent constructs in surveys. In addition, the study could inform accurate data collection procedures leading to better interpretability and generalizability of findings (Jenn, 2006). Moreover, by comparing scores from three major mathematics learning outcome instruments, and by applying bifactor modelling in doing so, the present study sought to establish the degree of integration between three different theoretical perspectives regarding the nature of test items (IRT), their difficulty levels and the ability of test-takers (MMT), and standard measurement procedures (CTT).

Operational Definition of Terms

According to Roopa and Rani (2017), a questionnaire is a primary data collection instrument that is defined as a self-assessment tool for collecting and recording information about a particular issue of interest and consist of a list of questions (items or measures) that respondents answer either by supplying required information (open-ended) or by choosing responses from a continuum of options (closed-ended). A questionnaire basically gathers the attitudes and opinions of respondents on some general or specific topic or issue of interest. In the context of the present study, a self-report questionnaire refers to ordinal data measures (5-point Likert scales). While mathematics performance or teacher-made achievement tests are constructed by the teacher to assess teaching effectiveness and the learning progress of learners and tend to cover just a limited amount of content (Milawati, 2019), a standardized test on the other hand, is a test that follows certain norms (possesses a mean score for measuring achievement) and is intended for general use, covering a wider content (Qassimi, 2021). In the context of the present study, mathematics performance was measured through the pre-mock (teacher-made test), while mathematics achievement was measured through the regional mock examination. According to Cronbach and Meehl (as cited in Shou et al., 2022) criterion validity 'indicates how well the scores or responses of a test converge with the criterion variables with which the test is supposed to converge and in the context of the present study, criterion validity refers to the degree to which scores from the questionnaire on students' perceived performance in mathematics on specific mathematics topics, and scores from teacher-made test individually and collectively compare with scores from the mock GCE examination (a standardized examination).

Methodology

Research Design

The cross-sectional survey research design was utilized as the study's inquiry strategy. This research design is characterized by a single wave of data collection from a representative sample of the population of the study and there is no manipulation of the variables of the study. The area of the study was the Fako Division of the South West Region (SWR) of Cameroon. The Fako Division is a vibrant educational hub, teeming with different institutions of learning consisting of; primary, secondary, and tertiary institutions with experts in different research fields who manage and work in these institutions. In addition, the Fako Division is one of six administrative Divisions in the SWR. The other Divisions are; Manyu, Meme, Lebialem, Ndian and Kupe-Muanenguba. The Fako Division has six sub-divisions consisting of Buea, Muyuka, Tiko, Limbe I, Limbe II, and Limbe III sub-divisions.

Population of the Study

The population of the study consisted of all secondary school students from the Fako Division. Due to the heterogeneous nature of the population, first cycle students from public, confessional, and lay private schools were targeted from the Buea sub-division. The sampling frame for the study therefore consisted of a list of all 154 secondary schools in the Buea sub-division made up of 46 public, 25 confessional, and 83 lay private schools. With the help of the study's sampling frame, an accessible population made up of form 4 & form 5 students from three secondary schools in the Tiko and Buea sub-divisions was selected. Finally, a sample of 206 form 5 students was selected from the three secondary schools. To resolve the heterogeneous population, such that uniformity could be achieved in all possible sample units of the study, a suitable probability sampling technique was employed in the selection. Specifically, the proportionate sampling technique was used to select the sample of the study. This technique ensured that homogeneity and uniformity were achieved in individual sample units. This ensured that all elements (students in this case) chosen for the sample possessed an identical trait in terms of their classes, ages, and level of study. The proportionate sampling technique, entailed that students were selected from each stratum (public, confessional, and lay private) in proportion to the population representing the particular stratum, such that the most populated schools received the greater percentage of respondents and vice versa. Numerically speaking, 90 students were selected from public schools, 60 from confessional, and 56 from lay private schools. The population of girls was 116, while there were 90 boys. These gender and school type quotas presented above ensured that the desired uniformity was achieved in sampling units.

Instruments of Data Collection

Data for the study were collected using three different instruments. The instruments consisted of a student-reported test, a teacher-made test, and a standardized achievement test. The student-reported test was a 5-point Likert scale questionnaire measuring perceived mathematics (behavioural) performance which requested students to rate their abilities on a set of 10 tasks. The pre-mock examination which is designed to mimic the regional Mock for the General Certificate of Education Examination (GCE), was utilized as the instrument for measuring mathematics performance. The pre-mock does not undergo any form of harmonization or standardization, and is therefore more or less a school-based or teacher-made test, given that its design, content, and weighting of items are determined solely by the classroom teacher and therefore vary from one school to another. Finally, the regional mock examination for mathematics for the 2023/2024 academic year was utilized as the main instrument for measuring students' mathematics achievement. The regional mock paper for mathematics provides evidence of predictive criterion validity in relation to the GCE Ordinary Level (a certificate examination) mathematics paper (Nekang & Anyi, 2022). The regional mock mathematics paper was therefore taken to be the criterion variable in the present study. The GCE 'O' Level mathematics paper's assessment objective is to test learning at 5 levels of learning from knowledge (30%), understanding (40%), application (20%), to analysis and synthesis (10%). The examination is made up of two papers (I & II). Paper I, contains 50 MCQ type questions, written for $1\frac{1}{2}$ hours and carries a total weight of 30% of the examinations. Paper II contains two sections (A & B) and has two questions types; structural, and essay. There are 15 structural questions, and 5 essay questions which together make up 70% of the total examination. Test questions are pretested and item analysis is carried out to vet and remove any misleading test items from the examination. In addition, examination conditions, and rules guiding test corrections, moderation, and interpretation of results are applied the same for every candidate in the region. The "O-level" regional mock mathematics paper was therefore taken to constitute the criterion variable for the study, and therefore acted like the benchmark to which the questionnaire and pre-mock datasets could be compared to.

Measures

Mathematics achievement measures comprised of eleven factors or competencies, consisting of topics prescribed by the GCE syllabus for Ordinary Level Mathematics. These factors consisted of numbers, sets and logic, functions, euclidean geometry, mensuration, rectangular coordinate geometry and graphs, algebra and networks, trigonometry, vectors, matrices and transformations, and statistics and probability. Following this, the teacher-made test and the mathematics achievement test were each made up of eleven items, selected such that an item represented each competence of the measured variable.

Perceived mathematics performance items were selected on an ordinal scale, while mathematics performance, and mathematics achievement tests yielded continuous data. The greatest challenge to overcome in the construction of the study's data analysis model therefore, was how to transform the continuous scales in the teacher-made test and achievement test into an ordinal scale such that a common scale that could fit in a confirmatory factor analysis measurement model, could be established in all three instruments. In other words, the challenge was to make each item in the classical test which relies heavily on an aggregation of scores from several items to be as important and as informative as items in an item response theory-based test. Continuous scores had to be immediately transformed into ordinal categories. The performance and achievement tests were made up of MCQ, structural, and essay type questions. While the achievement test followed the standard prescribed by the GCE syllabus of 50 MCQ, 15 structural, and 5 essay questions from 11 topics. The final scope of the mathematics performance test questions consisted of 25 MCQ's, 10 structural, and 4 essay questions from 10 topics (probability and statistics were not tested bringing the exams coverage to *55.7% with respect to number of questions and 90.9% with respect to number of topics*)**.** MCQ's were weighted equally across both tests, a point per item. The weighting (mark distribution) of structural and essay type questions varied across items due to different demand intensities and different levels of difficulty for each item. The largest percentage of test items and consequently the largest weighting came from algebraic networks.

The total scores for test questions on different measured variables of the study (topics) ranged from a minimum of 1 point per question in paper I to 12 points per question in Paper II. To resolve these discrepancies within the framework of the CTT, the marks were all converted to a scale of 4. To do this, a student's mark for a given question or group of questions from the same topic, was divided by the total mark for that question or group of questions, and the dividend was then multiplied by 4.

*Converted mark=((Marks scored on a given item)/(Total marks per item)*4)*

The procedure brought all scores to lie within the range from 0 to 4, with 0 representing the lower bound and lowest test score possible, and 4 representing the upper bound and highest score possible for any given test item. To transform continuous scores created from the conversion process into ordinal categories, such that they could be utilized in structural equation modelling (SEM) measurement models, the converted scores were recoded into different variables as demonstrated next. The range of marks from 0 to 0.4 was coded as 0 and evaluated as *'Failed'*, the range of marks from 0.5 to 1.4 was coded as 1 and evaluated as *'Weak'*, the range of marks from 1.5 to 2.4 was coded as 2 and evaluated as *'Average'*, the range of marks from 2.5 to 3.4 was coded as 3 and evaluated as *'Good'*, and finally, the range of marks from 3.5 to 4.4 was coded as 4 and evaluated as *'Very Good'*.

During the transformation of mathematics performance and mathematics achievement scores, separate *'data-forms'* (researcher's constructed tables that show the distribution of marks from different topics assessed in the tests) were developed in line with the tasks selected by the different schools. One such *'data-form',* for example, from one of the sample schools, consisted of 10 questions (9 MCQ's 1 mark each, and 1 essay question for 15 marks) tested on numbers. This made the total score on numbers alone to be 24 marks (these scores were later standardized on a 4-point scale). In addition, 3 questions (1 MCQ, 1 structural, and 1 essay) tested on sets and logic. This made the total score on sets and logic alone to be 20 marks (and just like for numbers, the scores were later standardized on a 4-point scale) and so on. Though the tasks from the different schools were not standardized (different schools selected dissimilar tasks both in topics and weighting), some were alike, but most were however equivalent. According to Medley (as cited in Manizade et al., 2023, p. 198), the "successful assessment of students' outcomes involves the above three essential steps".

The student-reported test was made up of 10 questions (with a *90.9%* coverage rate with respect to topics) whose responses were selected on a 5-point Likert scale. Data from a 5-point scale is essentially continuous, even for a small sample in contrast to Likert scales with 4 or less response options which require a very large sample size to be considered continuous (Johnson & Creech, 1983; Sullivan & Artino, 2013). By that fact, a sample size of 206 was selected for the study. On the 5-point Likert scale, students were requested to rate their abilities on each of the items on a continuum from 0 to 4, with 0 being the lowest score and 4 being the highest score. Under the framework of the MMT (Rasch's model) given variations in the difficulty of test items, and the discrepancies that exist in individual test-taker's abilities, students were not timed while completing the questionnaire, and were therefore given maximum time to reflect on individual items. Following that, students were then requested to rate their abilities in correctly solving the tasks. The design and selection of the scale was guided by Ortiz (2016), 4-stage model of problem-solving of task in a mathematics classroom. According to this model, a student should effectively solve a given mathematical task in four stages; *understanding the problem, devising a plan to solve the problem, carrying out the plan, and looking back.* A student selecting a score of zero by implication did not have any knowledge on how to solve the problem. A score of 1 implied that the student could interpret the questions, a score of 2 implied a student could provide a valid algorithm leading to the solution, a score of 3 implied a student could solve the problem but could not guarantee accuracy, and a score of 4 meant that a student could essentially solve the problem and get an accurate answer. Students' performances on the scale were evaluated as follow; *0=No Idea on task, 1=Can interpret task, 2=Can provide valid steps, 3=Can solve correctly/not sure of answer, 4=Can solve and obtain correct answer*. Students were drilled for 15 minutes prior to completing the test, in order to familiarize and ensure accurate self-rating of tasks. Items were specific mathematical tasks for example, *"I can get the correct answer in evaluating 15-4*2+3"*.

Validity

The instruments of data collection in the study were validated in four ways. To ensure face validity, the instruments were handed to two independent researchers who made an appraisal of the physical appearance of the instruments, in order to ascertain the extent to which the outlined items were in line with the stated objectives of the study. Secondly, content validation of the instruments entailed assessing the degree of homogeneity of different groups of items for the different measured variables in the study. Within the framework of the IRT, items for any given measured variable were formulated such that they were individually different but collectively measuring the same construct. In addition, instructions were clearly delineated on the instruments for clarity so as to effectively manage bias in the measurement process introduced by extraneous and confounding variables. Moreover, the demands of the tests were kept to a minimum such that unsolicited anxiety and tension did not characterize test-taking, given that test items were of varying difficulty, and that test-takers had varying abilities in answering them. In line with the MMT, test consisting of student-reported measures were not timed. Situational variables including classroom conditions and other environmental factors which interfere with test administration and test-taking were also checked. For example, classroom lighting, seating, and serenity/noise free classrooms were ensured during administration. In this light, the researchers hoped to dismiss any spurious correlations from the findings, leading to an accurate interpretation of findings.

Construct validity was measured for each of the study's variables through convergent and discriminant validity, and scale dimensionality was ascertained through a bifactor model analysis discussed under data analyses. Specifically, discriminant and convergent validities were ascertained through the calculation of average variances extracted (AVE) and composite reliability (CR) statistics respectively. AVEs and CRs were calculated for each of the five univariate (single-factor) measurement models through confirmatory factor analysis (CFA) with the aid of the measurement quality calculator developed by Kesiki (2023). Besides the 5 single-factor CFA models (see table 1 for individual univariate model indices), a bifactor model was used to test the main research hypothesis for the study. The models were assessed in an attempt to establish measurement quality in the holistic measurement of mathematics learning outcomes.

Covariates

Though mathematics self-efficacy and mathematics self-concepts were not the focus of the present study, their potential to introduce spurious correlations in the link between perceived mathematics performance and mathematics achievement was taken into consideration. Mathematics self-efficacy refers to the extent to which students believe in their own ability to understand specific mathematics topics (behaviour), while mathematics self-concept refers to students’ beliefs in their own mathematics abilities (Betz & Hacket, 1983; OECD, 2013). Beside requesting students to rate their abilities in specific mathematical tasks in measuring perceived mathematics performance, students’ self-efficacy and self-concepts of mathematics ability were also measured. Self-efficacy and self-concepts of mathematics ability were then taken to be covariates of perceived mathematics performance in addition to gender and school type; because of their potential to share with the later a considerable amount of the variance in mathematics achievement (Arens et al., 2017). Students were requested to complete a 5-point Likert type questionnaire by stating their level of belief and confidence regarding their ability to demonstrate mastery on certain mathematics behaviours (topics). On that scale; 0=Never, 1=Seldom (Not often), 2=sometimes, 3=Often, 4=Always. The 6 items scale (5-point Likert) from PALMA, a 5-year large-scale long study on mathematics self-concept’s link with mathematics achievement by Arens et al. (2017) was adopted in the measurement of mathematics self-concept. The responses ranged from 0=Not at all true to 4=completely true. In the present study, items involved very specific mathematical tasks measuring students’ perceived mathematics performance, and ability in certain mathematics topics measuring self-efficacy, to attitudes on students’ general abilities in the subject of mathematics measuring self-concept. For example, *“I can get the correct answer in evaluating 15-4*2+3”*, *“I believe I can distinguish with no errors between whole, natural, integers, rational, and real numbers”*, and *“In math, I am a talented student”*, were typical items measuring perceived mathematics performance, self-efficacy, and self-concept, respectively.

Measurement Models

The measurement models for the five variables in the study (three main variables consisting of perceived mathematics performance, mathematics performance, and mathematics achievement; and two covariates, mathematics self-efficacy, mathematics self-concept) together with the measurement quality for each variable were presented. The fit statistics for the proposed 6-item mathematics performance model, 6-item perceived mathematics performance model, 9-item mathematics achievement model, 8-item mathematics self-efficacy model, and the 10-item mathematics self-concept model, each revealed good internal consistency for the group of measures in each observed variable of the study. Over four error terms were correlated in the mathematics performance model, three in the mathematics self-efficacy model, and over six in the mathematics self-concept model in order for best model fits to be achieved in the respective models. Fit indices, average variance extracted, composite reliability, and the Cronbach alpha coefficients for all three observed variables and covariates were presented in the table below. The values revealed that each measurement model provided a good fit for the data.

Table 1: Measurement Models Fit Indices for Mathematics Performance, Perceived Mathematics Performance, Mathematics Achievement, Mathematics Self-Efficacy, and Mathematics Self-Concept

Fit index Category	Model fit index	Obtained values					Cut off values Best Fit
		MP	PMP	MA	MSE	MSC	
Absolute fit indices	CMIN	31.811	11.811	6.483	31.939	76.712	
	DF	16	9	9	17	28	> 1
	SRMR	.0461	.0361	.0306	.0418	.0470	< .08
	GFI	1.000	.981	.990	.963	.922	< .95
	RMSEA	.069	.039	.000	.065	.092	< .08
Comparative fit indices	CFI	.979	.986	1.000	.964	.960	> .95
	IFI	.979	.986	1.037	.965	.961	> .95
Parsimonious Correction fit index	PNFI	.548	.567	.550	.563	.585	> .5
	CMIN/DF	1.988	1.312	.720	1.879	2.740	< 3
Internal Consistency	AVE	.45	.32	.18	.36	.54	≥ .50
	CR	.82	.73	.57	.81	.90	≥ .70
	Sig.	.011	.224	.691	.015	.000	> .05
DECISION	**Best Fit**	**Best Fit**	**Best Fit**	**Best Fit**	**Best Fit**		

MP: Mathematics Performance; PMP: Perceived Mathematics Performance; MA: Mathematics Achievement; MSE: Mathematics Self-Efficacy; MSC: Mathematics Self-Concept; DF: Degree of Freedom; CMIIN: Chi Square; CMIN/DF: Chi Square/Degree of Freedom; CFI: Comparative Fit Index; GFI: Goodness-Of-Fit Index; RMSEA: Root Mean Square Error of Approximation; PNFI: Parsimonious-Adjusted Measures Index; IFI:

Incremental Fit Index; SRMR: Standardized Root Mean Square Residual; CR: Composite Reliability; AVE: Average Variance Extracted; Sig.: Significance Level; α: Cronbach Alpha Coefficient.

Reliability

Construct reliability for the study's measured variables were assessed in two ways; through the calculation of the Cronbach alpha coefficient, and composite reliability following a test-retest procedure. The questionnaire measuring perceived mathematics performance, self-efficacy and self-concepts of mathematical ability was first examined for reliability through a pilot test. The reliability of the instrument was first tested with a sample of 40 students during the first phase of pilot testing, and later to a different sample of 20 more students after a period of 3 months to determine the stability of the instrument (re-test). The Cronbach alpha coefficients for perceived mathematics performance, self-efficacy and self-concepts, for the study were respectively .797, .828, and .922. The sample for the pilot test involved students drawn from a school that did not constitute part of the final sample for the study. The final questionnaire that was answered by the students contained three sub-scales; perceived mathematics performance, self-efficacy, and self-concepts. Given that mathematics performance, and mathematics achievement possessed continuous data, the Cronbach alpha reliability coefficients for both measured variables was ascertained in two ways. Firstly, regarding the continuous data, an intraclass correlation coefficient was computed and was found to be .84 for the instrument. Secondly, after the continuous data were converted to a 4-point scale and recoded into a 5-point ordinal Likert scale data, the Cronbach reliability for the scores from both tests were respectively .810 and .726.

Methods of Data Analyses

Data for the study were analyzed descriptively and inferentially. The specific descriptive statistics tools included frequencies, percentages, means, and standard deviations. The specific inferential statistics tools/models that were utilized in the study consisted of a confirmatory factor analysis (CFA) model, the bifactor model (BM), ANOVA, and multivariate regression analysis. In analyzing the main objective's variable datasets, 5 univariate CFA models were first constructed to test the measurement quality for each of the five measured variables (see table 1). Following this, a bifactor model was used to test the dimension of the scale (or internal structure). The bifactor model was applied to determine whether the three subscales; the questionnaire, premock (teacher-made test), and the mock (mathematics achievement test) were all measuring a common mathematical ability or three separate/unique mathematical abilities. The bifactor model for the study, measuring, perceived mathematics performance, mathematics performance, and mathematics achievement was abbreviated the MLOM, standing for the *'Mathematics Learning Outcome Model'*. The MLOM initially consisted of three subscales, with a total of 30 items. The finalized MLOM still retained three subscales, but had a total of 19 items. The 11 items deleted from the initial model showed extremely low correlations with the respective domain specific scales and were therefore removed from the model. Moreover, to determine the likelihood that students' perceived mathematics performance and mathematics performance differentially predicted mathematics achievement, descriptive statistics, multivariate regression analysis, analysis of variance, and a line chart were used to analyzed and visualize the datasets respectively.

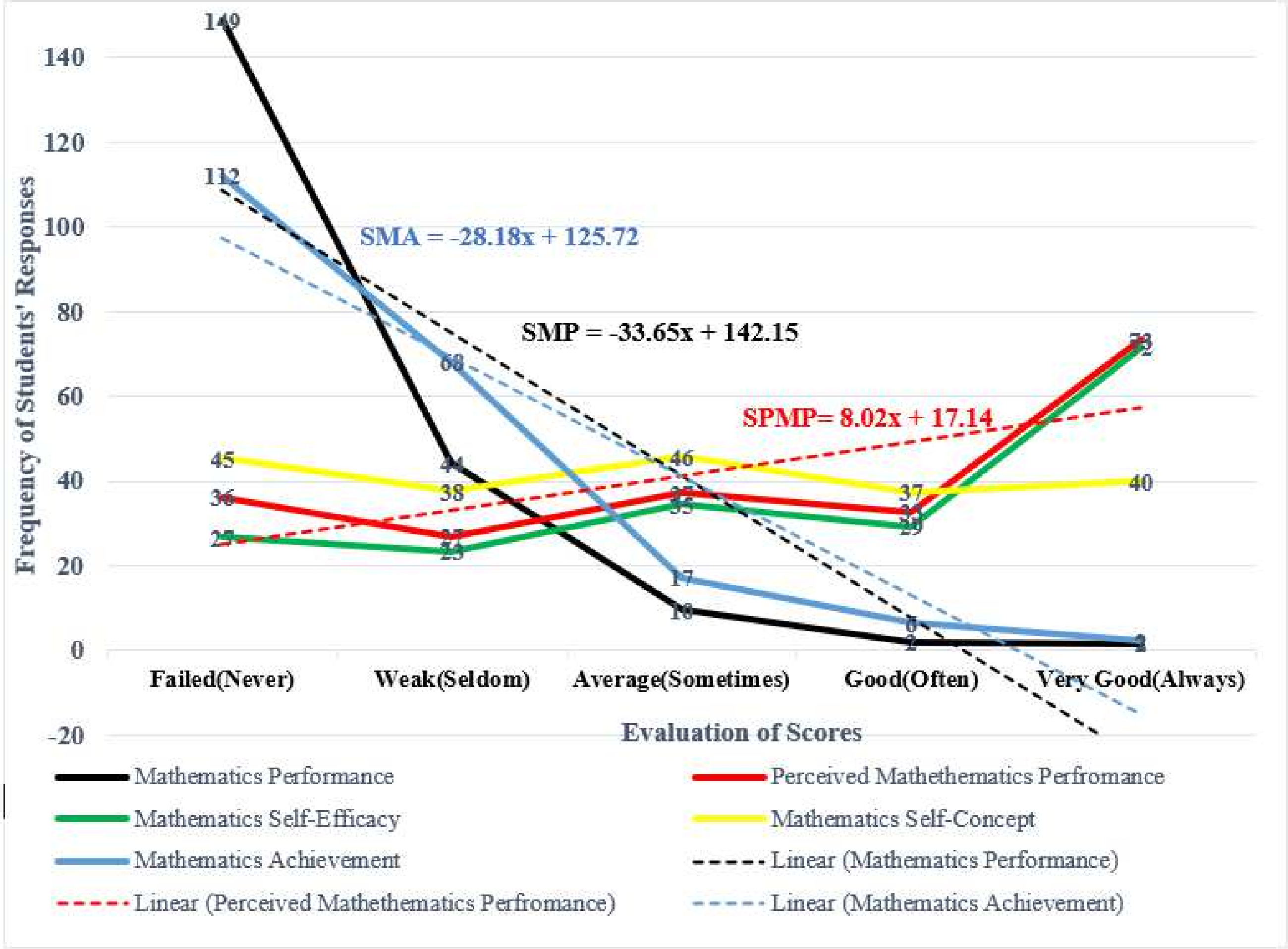
According to Reise et al. (2007), a BM has two kinds of factors; a general factor and domain specific factors. The BM best addresses the question about the dimensionality of items of a construct (internal structure of the scale or a test). The combined sets of items across all domain specific factors load onto the general factor, and capture a domain (trait) common in all subscales. The general factor represents the variance common in all domain specific factors in the model, while domain specific factors represent variance unique in their respective domains (Fang et al., 2020; Reise, 2012). In the present study, the general factor (labelled as mathematics learning outcome) represented a common or general scale to which all the items of all three sub-scales loaded, and captured the variance common across domain-specific items and domain specific scales. The BM in the present study therefore tested the null hypothesis that the general scale did not significantly explain most of the variance in individual items across all three sub-scales (questionnaires, teacher-made test, and achievement test). In other words, the scales were hypothesized not to measure a common mathematical ability. Finally, to determine if the three subscales together provided a much more holistic measurement of mathematics learning outcomes, descriptive statistics tools were used to study trends and patterns that existed in the datasets.

Findings

The findings were presented according to the specific research questions of the study. Firstly, a visual representation of the results with regards to how respondents performed on all constructs under measurement including covariates were presented. Secondly different bifactor models were hypothesized in order to find a fitting model for the data, and finally, an ANOVA test was carried out to determine differences in the means of the students' scores in the different subscales.

Research Question One: To what extent do student-reported mathematics outcomes, and mathematics performance predict mathematics achievement?

The line chart below visualizes number of students against different categories of students' scores from five different tests. The tests included student-reported scores or perceived mathematics performance (red line), teacher-made test or mathematics performance (black line), achievement test or standardized test (blue line), self-efficacy (green line), and self-concept (yellow line). Students' scores were evaluated in categories as follows; 0% < Weak ≤ 25%, 25% < Average ≤ 50%, 50% < Good ≤ 75%, and 75% < Very Good ≤ 100%.



SPM: Frequency of Student Responses on Mathematics Performance Items; **SMA:** Frequency of Student Responses on Mathematics Achievement Items; **SPMP:** Frequency of Student Responses on Perceived Mathematics Performance Items.

Figure 1: Comparing Trends in Students' Performances Across Different Instruments

The visualization of students' tests scores on the line chart revealed two different sets of patterns. Mathematics performance recorded the highest number of fails compared to mathematics achievement and perceived mathematics performance. While mathematics performance and mathematics achievement scores showed patterns that were similar (Pattern 1), perceived mathematics performance scores followed a pattern consistent with those of self-efficacy and self-concept scores (pattern 2). In particular, the first pattern of curves revealed that there were fewer students with higher scores, and that more 'failed' evaluations of the mathematics performance and mathematics achievement items were associated with a great number of students' responses (as seen from the negative slopes in the equations for mathematics performance and achievement; -33.65 and -28.18 respectively) compared to the second category of curves in which there were more students with higher scores, and more 'very good' evaluations were associated with a greater number of students responses of the perceived mathematics performance, self-concept, and self-efficacy items (as seen from the positive slopes of the dotted lines for perceived mathematics performance; 8.02). The equations of the straight lines for mathematics performance and mathematics achievement in particular revealed that for every unit rise in the quality of students' evaluation of learning, the number of students' responses on mathematics performance and mathematics achievement items actually drooped by an average of 34 and 28 responses respectively. In general, more students performed better in the mathematics achievement test than in the mathematics performance tests. There were no differences in the number of 'very good' performances for students in both the mathematics achievement and the mathematics performance tests, given that an equal number of such performances were recorded in both tests. Moreover, in contrast perceived mathematics performance scores followed a pattern completely different from that of mathematics achievement and mathematics performance. Its pattern was similar to that followed by self-efficacy and self-concept.

A multiple regression analysis of two predictor variables was first conducted without the inclusion of covariates, which were later added into the model. When covariates were added, this resulted in model misspecification, and neither mathematics performance, nor perceived mathematics performance showed any significant effects on mathematics achievement. However, when some covariates (gender and school type) were removed, both the model specification and the effect of perceived mathematics performance on mathematics achievement were significant. A further analysis (mediation and moderation informed by initial challenges in fitting the model) was conducted on the variables (mathematics self-concept and self-efficacy) to test for possible colliding and intermediary effects on the individual links between mathematics performance, perceived mathematics performance and mathematics achievement. When that was done both supposed covariates (mathematics self-concept and self-efficacy) turned out to be partial mediators of each of the association between mathematics performance and mathematics achievement, and perceived mathematics performance and mathematics achievement. Both the direct and indirect effects of MSC and MSE in the models were significant.

Table 2: Regression Model Summary on the Effect of mathematics performance and perceived mathematics performance on mathematics achievement

Independent Variables	Unstandardized Coefficients		t-ratio	Sig.
	Slope (B)	Std. error		
Mathematics Performance	.080	.073	1.082	.281
Perceived Mathematics Performance	.065	.031	2.124	.035
Mathematics Self-Efficacy	-.050	.038	-1.300	.195
Mathematics Self-Concept	.045	.034	1.342	.181

Constant = 10.740
R^2 = .024
F-Ratio = 2.449, P = .038 < .05
SEE = 10.45108
n = 206				

The data was significantly fitted into a linear model (p < .05) and together mathematics performance and perceived mathematics performance accounted for 2.4% of the movements or variations in students' mathematics achievement. In addition, perceived mathematics performance was found to be a better predictor of students' mathematics achievement (p < .05) than mathematics performance (p > .05). The researchers therefore rejected the null hypothesis and concluded that students reported outcomes and mathematics performance significantly differ in their predictions of mathematics achievement.

Research Question Two: Is there any common ability measured by student-reported mathematics outcomes, mathematics performance, and mathematics achievement subscales?

Three bifactor models were tested at this stage of the study. The aim was to assessed the internal structure or the dimensionality of the MLOM scale. Firstly, a classic orthogonal bifactor model with three domain specific factors (mathematics performance, perceived mathematics, and mathematics achievement) and a general factor to which all subscales loaded was developed. Secondly, a bifactor S-1 model, with two subscales (mathematics performance, and mathematics achievement) loaded onto a general factor. According to Eid et al. (2017), and later supported by Pekmezci (2022), the bifactor S-1 model allows correlation between specific factors and enables items that do not form a common specific factor to be loaded only on the general factor. It further resolved the problem of negative factor loadings which were recorded in the classic orthogonal bifactor model and persisted even after the items were reversed coded. During the analyses, in the classic orthogonal bifactor model, 4 items with negative factor loadings on the domain specific factors were recoded. Though the model fit was improved, data could not still be parsimoniously fitted into the model. This was exacerbated by negative loadings which could not be completely eliminated despite steps to remove them. To resolve the issue of negative loadings in the classic orthogonal bifactor model, a 1-factor hierarchical model which according to Markon (2019) is equivalent to the classic orthogonal bifactor model was constructed. The 1-factor hierarchical model still did not provide a better fit for the data. In the bifactor S-1 model, perceived mathematics performance items were allowed to load onto the general factor but no domain specific factor was established for that measured variable. The model therefore had two subscales that loaded onto the general factor. Finally, an augmented oblique bifactor model which according to Zhang et al. (2023), is superior to the bifactor S-1 model (claim contended by Koch and Eid (2024)) was the third and final model to be tested in the study. In the augmented oblique bifactor model, items with negative loadings in the PMP subscale were removed. The augmented oblique bifactor model allows correlations between subscale factors that form a common factor, and in the present study, two subscale factors were correlated (MP and MA subscales). Out of all the three models, the augmented oblique bifactor model parsimoniously fitted the data. The results output for the augmented oblique bifactor model utilized in the present study is presented in figure 2 below.

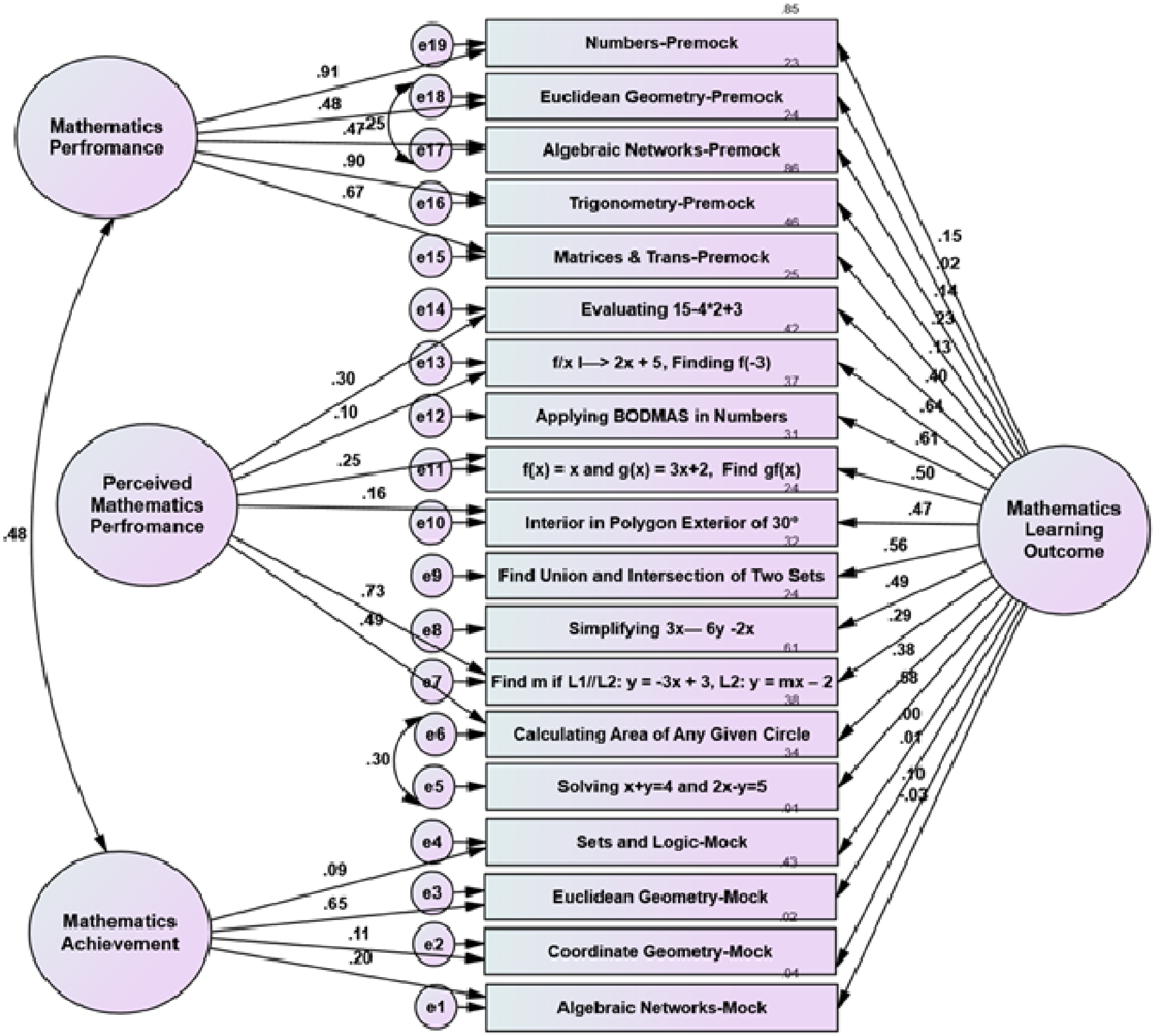


Figure 2: Results Output for the Augmented Oblique Bifactor Model

The augmented oblique bifactor models produced better fit for the data compared to the bifactor S-1 and the 1-factor hierarchical models. On that basis and for purposes of ancillary bifactor indices calculation, the augmented oblique model was selected over the other models despite having the second largest Akaike's information criterion (AIC) value of the three hypothesized models. In the S-1 model specifically, the complete absence of loadings on the items of the PMP subscale made it impossible for the calculation of certain vital ancillary bifactor indices to be completed. The problem was resolved with the use of loadings from the augmented oblique bifactor model, which retained 6 out of the 10 loadings for that subscale.

Table 3: Measurement Models Fit Indices for the Bifactor S-1, 1-Factor Hierarchical, and the Augmented Oblique Bifactor Models

Fit Index Category	Model Fit Index	Obtained Values			Cut Off Values Acceptable Fit
		Bifactor S-1 Model	1-Factor Hierarchical Model	Augmented Oblique Bifactor Model	
Absolute fit indices	CMIN	198.370	214.320	200.911	
	DF	122	120	134	> 1
	SRMR	.0728	.0740	.0641	< .08
	RMSEA	.055	.059	.049	< .08
Comparative fit indices	CFI	.918	.898	.931	> .9
	IFI	.920	.902	.934	> .9
Parsimonious Correction fit index	PNFI	.651	.629	.646	> .5
	CMIN/DF	1.626	1.786	1.499	< 3
	AIC	296.370	316.320	312.911	Smallest taken
DECISION		**Acceptable Fit (Rejected)**	**Bad Fit (Rejected)**	**Acceptable Fit (Accepted)**	

To determine the internal structure or dimensionality of the mathematics learning outcome scale for the augmented oblique bifactor model, a number of ancillary bifactor indices were calculated. The calculation of these indices was facilitated by the use of the bifactor indices calculator developed by Dueber (2017). These indices included, the percent of uncontaminated correlations (PUC) which refers to the percentage of covariance which uniquely reflects variance from the general factor (MLO). According to Zhang et al. (2019), PUC values that exceeds .7 would reflect less bias in structural coefficients, and thus indicates that the MLO scale (general factor) can be treated as unidimensional. Another index that was calculated was the explained common variance (EVC) which according to Stuckey and Edelen (2015) represents the amount of common variance attributed to the general factor. An ECV value that exceeds .7 would be indicative of a strong general factor and would thus suggest unidimesionality of the MLO scale. In addition, the individual explained common variance (IECV), referring to the common variance in individual items attributed to the general factor was also computed. An item with an IECV that exceeds .8 would reflect the MLO scale more than the subscale factors. Moreover, omega coefficients were used in judging the reliability of the MLOM with higher Omega's representing more reliable general and subscale factors. Finally, the degree to which latent constructs were well-defined in the study were assessed through the Hancock H indices, with values greater than .8 signifying greater internal consistency of the measures.

Table 4: Calculated Augmented Oblique Bifactor Model's Statistical Indices for Judging Scale Dimensionality

	ECV (S&E)	ECV (NEW)	Omega/ OmegaS	OmegaH/ OmegaHS	Relative Omega	H	FD	Crucial Ancillary Bifactor Indices
GF	.402	**.402**	.822	**.539**	.656	.791	.876	**ECV = .402**
PMP	.144	.396	.777	.259	.333	.622	.863	**PUC = .819**
MP	.381	.956	.947	.911	.962	.912	.991	**Cutoffs for some Unidimesionality PUC < .8, $ECV_g > .7$, $\omega H_g > .7$**
MA	.073	.980	.241	.240	.995	.446	.668	

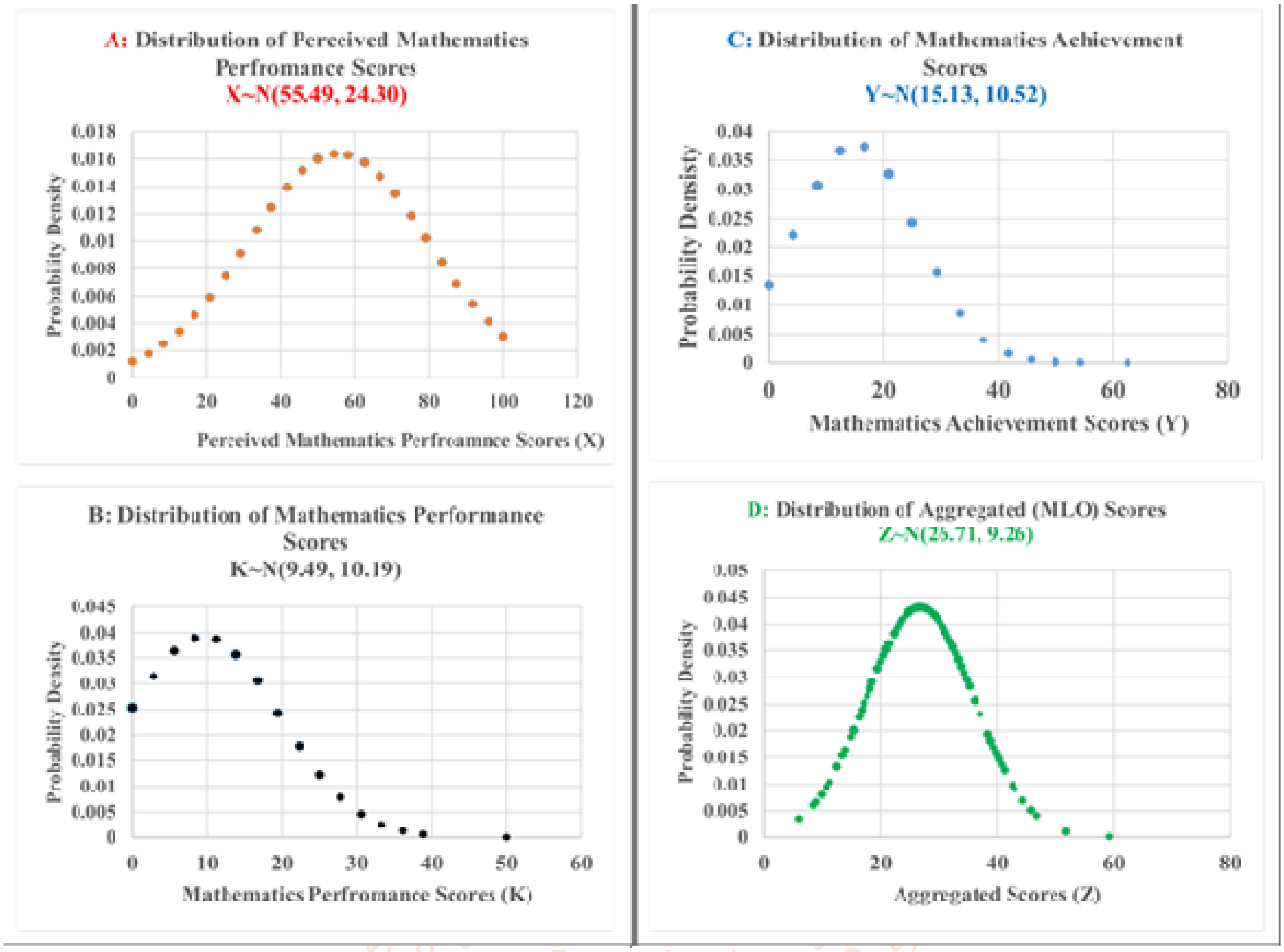
GF: General Factor; PMP: Perceived Mathematics Performance; MP: Mathematics Performance; MA: Mathematics Achievement; PUC: Percent of Uncontaminated Correlations; ECV: Explained Common Variance; ωH: Omega Hierarchical, H: Hancock & Mueller Correlation; FD: Factor Determinacy; g: The subscript g refers to the General Factor.

In the present study, the ECV for the MLO was .402, meaning that only 40% of the variance was attributed to the MLO (general factor), while the remaining 60% was attributed to the subscale factors. ECV's for the subscale factor were between .396 and .980, meaning that between 39.6% and 98% of the variance was attributed to the subscale factors. Items 9, 10, 11, 12, 13 and 14 had ICVE's that exceeded .8, and the average of all IECV's was .46, meaning that only 46% of the common variance was attributed to the MLO. The PUC index was .819, meaning that the MLO reflected more bias in structural coefficients, and thus indicates that the MLO could not be treated as unidimensional. The omegaH value for the MLO was .539, and those for the subscales were between .240 and .911, signifying that the subscale factors were more reliable than the MLO factors. Finally, the Hancok (H) values reflected well defined constructs and good internal consistency of items of the MLO and its subscales, except for the mathematics achievement subscale with H value below .7. According to Reise et al. (2013, p. 22), when PUC values are lower than .80, general ECV values greater than .60 and OmegaH greater than .70 (of the general factor) "suggest that the presence of some multidimensionality is not severe enough to disqualify the interpretation of the instrument as primarily unidimensional". In the present study as detailed in table 4 above, the **PUC = .819 > .80**, general **ECV = .402 < .6**, **ωH = .539 < .70** for the general factor, and 6 out of all 19 items had IEVC values that exceeded .8. According to Bonifay et al. (2015, p. 6), "it appears that when ECV is above .70, relative bias is below the 10% benchmark and when ECV is above .80, relative bias is less than 5%". The researchers therefore deduced from the calculated bifactor indices and noted that; relative bias was above the 10% benchmark leading to the conclusion that, the presence of some multidimensionality was severe enough to disqualify the interpretation of the instrument as primarily unidimensional. Guided by these results and by the recommendations of Reise et al. (2013), regarding relative bias exceeding the 10% benchmark, it was concluded that, the MLO was primarily multidimensional. Therefore, the null hypothesis was rejected and it was concluded that, all three mathematics outcome subscales measure separate mathematical abilities. Evidence was contrary to a common mathematical ability measured across all subscales in the study.

Research Question Three: Do student-reported mathematics outcomes, mathematics performance, and mathematics achievement significantly provide a holistic measurement of mathematics learning outcomes?

To answer this research question, a one-way ANOVA test was carried out on the datasets of the study to investigate the effect of the three subscales on the mathematics learning outcome of secondary school students. The researchers hypothesized that there were significant differences in the means of the three mathematics learning outcomes subscales. If the null hypothesis were to be upheld, it would mean that evaluating the average of the three subscales would significantly influence students' final scores. In addition, a justification for taking the mean of the three subscales to make the final score came from the fact that a multidimensional internal structure was found for the MLO scale (see table 4), meaning that such a method of scoring the instrument was correct and consistent with theory. Further proof to demonstrate the reliability of aggregating subscale scores was provided through a normality test which assessed and compared the likelihood of finding a students' score within the mean of the various subscales, and a plot of trendlines to visualize patterns in the MLO dataset so as to assess and compare the distances between trendlines estimated values and actual subscale scores.

The mean test scores and standard deviations of the subscales and the MLO scale were as follows; mathematics performance (M = 9.49, SD = 10.19), perceived mathematics performance (M = 55.49, SD = 24.30), mathematics achievement (M = 15.13, SD = 10.52); mathematics learning outcome (M = 24.15, SD = 9.26). The one-way ANOVA revealed a significant effect of the subscales on the mathematics learning outcome scale; indicating that there was a statistically significant mean difference in subscale scores between at least two groups, F (2, 608) = 478.715, p < .001. The effect size, eta squared (η^2), was .61, indicating a large effect. Tukey's HSD post hoc test for multiple comparisons showed that the mean subscale score was significantly different between mathematics performance and perceived mathematics performance (p = .000, 95% C.I. = [-49.87, -42.25]). In addition, the mean subscale score was also significantly different between mathematics performance and mathematics achievement (p = .001, 95% C.I. = [-9.56, -1.89]). In summary, students scored significantly higher on the perceived mathematics performance subscale (p < .001), than on both the mathematics achievement (p < .001) and the mathematics performance (p < .001) subscales. The effect size of .61 confirmed that these differences were practically significant. More evidence that the aggregated subscale scores provided a better measurement for the MLO was provided by the test of normality. The Kolmogorov-Smirnov tests of normality revealed that only the dataset for perceived mathematics performance was normally distributed (Statistic = .049, p = .200) as compared to mathematics performance and mathematics achievement datasets (Statistic = .246, p = .000; Statistic = .186, p = .000 respectively). Irrespective of that, the Kolmogorov-Smirnov tests of normality for the aggregated dataset also revealed normally distributed scores (Statistic = .060, p = .071 respectively). The normally distributed scores for the different scales are visualized in the curves in figure 3 below.



A: Distribution of Perceived Mathematics Perfromance Scores X~N(55.49, 24.30)

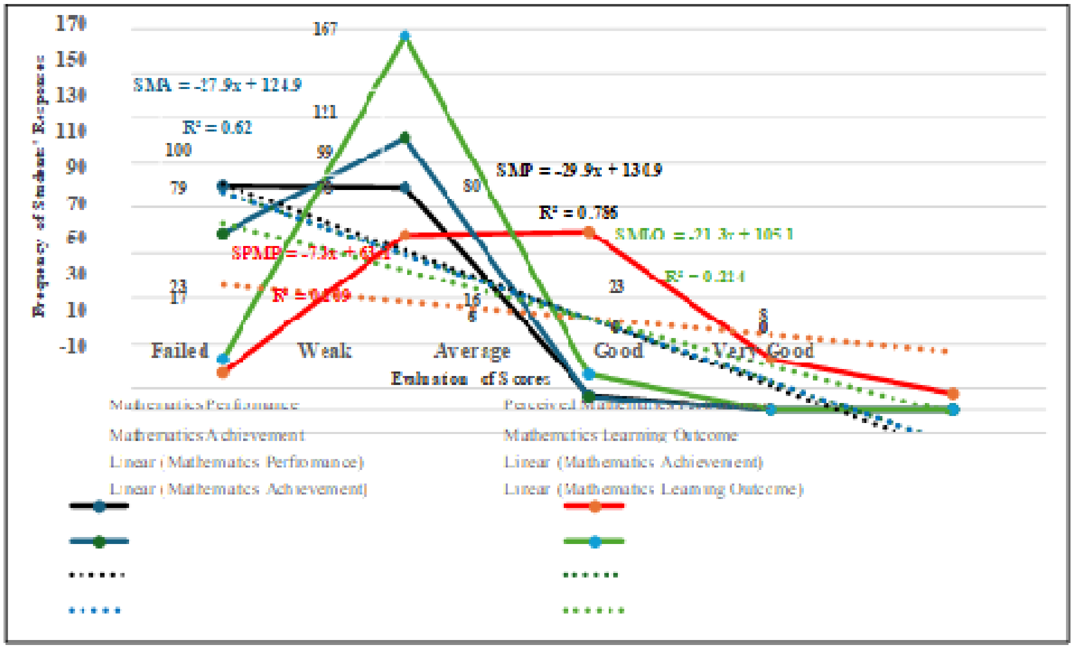
C: Distribution of Mathematics Achievement Scores Y~N(15.13, 10.52)

B: Distribution of Mathematics Performance Scores K~N(9.49, 10.19)

D: Distribution of Aggregated (MLO) Scores Z~N(26.71, 9.26)

Figure 3: Comparing Distributions of Subscale and Aggregated Scale Scores X~N($\overline{X}$, SD)

According to a normal distribution curve, 95% of students' scores in all three subscales in the present study fall between the mean plus or minus two standard deviations of the mean (Mean ± 2SD). According to curve A, 95% of students' perceived mathematics performance scores fall between (55.49 ± 48.6) % around the mean, and there is a high probability of finding a students' score within ±48.6% of the mean on this scale. For the mathematics performance test, there is a high probability of finding a students' score within ± 28.1% of the mean. According to curve C, 95% of students' mathematics achievement scores fall between (15.13 ± 21.04) % around the mean. For the aggregated scale, there is a high probability of finding a students' score within ± 18.52% of the mean. The perceived mathematics performance and aggregated scales (Curve A and D) are perfectly bell shaped compared to the mathematics performance and mathematics achievement tests. While curve D is taller and narrower, Curve A is shorter and wider in shape depicting that there is a higher likelihood for curve D, than A for measuring a particular students' score within a narrower range of scores. In other words, it is more common to find a student with an average score in the aggregated scale than it is to find a student with a very low or a very high score in the same tests. For the mathematics performance and mathematics achievement tests, the scores appear more centered towards the left, making it more common to finding a student with a very low score than it is to finding one with say an average or a very high score.



SMLO: Aggregated Frequency of Student Responses on the Mathematics Learning Outcome Scale

Figure 4: A Comparison of Trends on Students' Scores Between Individual Subscales and the Main Mathematics Learning Outcome Scale

The line chart shows trends in students' evaluations on the four different scales consisting of the three subscales and the mathematics learning outcome scale (the aggregated scale, gotten by taking the average of the three subscales). The chart revealed that students were evaluated better on the aggregated scale of subscales than on the individual subscales themselves (with the exception of the perceived mathematics performance subscale which remained better for evaluations of average, good and very good). In addition, the chart shows that when students' scores were aggregated on the three subscales, the number of students evaluated as weak and very good improved significantly as failed evaluations dipped. Moreover, the trendlines revealed a much less negative slope for the aggregated scores (-21.3) than for students' scores from mathematics performance (-29.9) and mathematics achievement (-27.9) subscales. More positive students' evaluations of the test scores were associated with far fewer number of students' responses in the mathematics performance and mathematics achievement test items than on the aggregated test. The linear regression models revealed that; 1 unit of improvement in students' evaluation was associated with drops in averagely 21 responses of students in mathematics performance items, while 1 unit of improvement in students' evaluation was associated with significant drops in averagely 28 responses of students in mathematics achievements items. The mean students score on the aggregated scale (M = 24.15) was higher than for the mathematics performance (M = 9.49) and mathematics achievement (M = 15.13) as the standard deviation for the MLO dropped slightly than for all the other subscale scores. The estimated values for the PMP subscale trendline were further away from the actual values (10.9%) than the estimated trendline values for the other subscales (78.7% and 62.6% respectively). The model estimates for the trendline values were closer to the actual values for the aggregated scale (22.5%) than for the PMP subscale (10.9%), but these estimates were lower than for the MA and MP subscales. This together with results of the ANOVA, and normality tests clearly demonstrated that the aggregation of scores from all three subscales together provided a holistic, realistic and much better measurement of mathematics learning outcomes. This led to a rejection of the null hypothesis and it was concluded that all three mathematics outcome subscales provide a much more holistic measurement of mathematics learning outcomes.

Discussions

The findings of the study revealed that perceived mathematics performance and mathematics performance differentially predict achievement. The findings supported those by Fernández-Cézar et al. (2021) who found that cognitive (mathematics performance) and behavioural variables (attitudes) differed in their prediction of mathematics achievement. In addition, the findings revealed that students' achievement improved with more positive students' perceptions of their mathematics performance abilities. Also, in the present study mathematics performance was found to have an indirect effect on mathematics achievement through mathematics self-efficacy and self-concept. The findings were similar to those by Kesiki and Nekang (2023), and Hidayatullah and Csíkos (2023) who found that prior performance and beliefs were indirectly associated with achievements through attitudes respectively. Moreover, the findings of the presents study revealed a multidimensional internal structure for the mathematics learning outcome scale consisting of both affective and cognitive measures. These findings were similar to evidence by Antara et al. (2020) pointing to the feasibility of developing an assessment instrument that effectively measures a two-dimensional measured variable. Finally, the findings from the present study supported evidence by Zhao et al. (2012), who found that a holistic measurement model for assessing mathematics learning should include both personal and contextual variables. Findings have implications for the MMT theory given that the interpretation of test scores on the MLO scale was based on the scores of individual subscales (different test types) rather than on scores from a single test.

Recommendations and Suggestions

Findings from the present study has demonstrated that, self-report scales especially closed ended questionnaires (excluding MCQs) should not substitute standard tests in the measurement of mathematics learning outcomes due to the fact that they do not provide sufficient evidence of predictive criterion validity when correlated with mathematics achievement. It's dataset in the present study appeared to show an extremely low association with the mathematics achievement dataset when controlled for mathematics performance ($r = .065$, $p < .05$). Moreover, students sampled in the present study grossly exaggerated their own perceptions of their abilities in solving specific mathematics tasks compared to their actual abilities in solving the tasks as was demonstrated through a comparison of their mathematics performance and achievement scores. The trendline model estimate for the PMP subscale was less fitting to the data and the trendline estimate was further away from the actual scores than for the other subscales (compare R^2 values for the trendlines in figure 4). As was the case in the present study students' perceptions of their own mathematics abilities were a positive predictor of mathematics achievement but their actual class test performances were not a significant predictor of mathematics achievement. So, they appear to be a total divergence between students' perceptions of their mathematical abilities and actual or concrete mathematical abilities. The mean students' perceived performance in mathematics for the study was 55.49% compared to students' mean scores in class and public test (9.49% and 15.13% respectively) which were significantly lower ($p < .05$). The difference in the mean performances of students in the three test that were carried out revealed that these differences in means were statically significant (see one-way ANOVA results).

Furthermore, the calculated statistical indices from the standardized bifactor loadings of the augmented oblique bifactor model provided enough evidence that relative bias in the model was above the 10% benchmark in which Reise et al. (2013) suggested that a unidimensional internal structure should be ruled out for such an instrument. This led to the conclusion that the MLOM was multidimensional. At best this would mean that, mathematics performance, perceived mathematics performance, and mathematics achievement subscales measure separate mathematical abilities and a single subscale therefore cannot claim to either measure all of students' mathematical abilities or act as a replacement for the other subscales. In other words, assessing a single dominant trait cannot be an effective measurement of students' mathematics learning but that an assessment of different behaviours (traits) of learners can provide an effective measurement of the construct. The raw total score was not a reliable enough measure of the MLOM (unidimesionality) but the subscale scores were a reliable-enough measure of their specific factors (multidimensionality). The questionnaire and mathematics performance test did not provide any evidence of criterion validity and should not be used as substitutes in the measurement of mathematics achievement.

Finally, the researchers reiterate that self-assessment instruments (questionnaire measures) should not be utilized as the lone instrument in the measurement of mathematics achievement. Dowrich (2008, p. 4) is emphatic on this by asserting that, '… a general rule in testing which states that no important decision should be made on the basis of one limited sample of behaviour'. However, in case a questionnaire is to be utilized in measuring students' mathematics learning outcomes, the teacher should consider using a self-assessment technique such as self-questioning or any other effective subjective performance-based intervention during the instructional and practical phases of mathematics learning to improve the accuracy of students' self-assessment (Mastnak et al., 2023). Secondly, the findings of the present study have implications for the MMT given that a multidimensional internal structure for the MLO warrants the use of different types of academic test in the effective measurement of mathematics learning outcomes. This entails that, other instruments measuring cognitive attributes should be employed as well and the aggregate of the scores from all instruments utilized should make the final performance. As seen in the present study, the aggregation of scores from different instruments improved students' mathematics performances while reducing the variations observed in the actual (individual subscales) panel test scores as they were slightly reduced (SD of aggregated scores= 9.26, SD of MP=10.19, and SD of MA=10.52). The results proved that aggregated scores from instruments testing both at the affective and cognitive domains sufficiently mimics normally distributed test scores (see figure 3).

In context, class tests (premock included) were found not to be significant direct predictors of the regional mock examination (standard test) since they did not provide any evidence of predictive criterion validity, only the indirect effect was significant through self-efficacy and self-concept of mathematical ability. To resolve this and reduce variations in teacher-made test (variations exist in test in terms of item difficulty, weighting, question format, and test duration, across classrooms and schools), the researchers recommended that class tests should be harmonized with the aim of reducing the variations that exist in the behaviours being tested. Though harmonizing class test would by no means make them standard, it would however reduce variations in test conditions.

References

[1] Anqui-Laja, A. E., & Laja. L. S. (2022). The causes of failure in basic mathematics at the college stage. *Psych Educ,* 5, 746-752, doi:10.5281/zenodo.7352857,

[2] Antara, G. W. S., Sudarma, K., & Dibia, K. (2020). The Assessment Instrument of Mathematics Learning Outcomes Based on HOTS Toward Two-Dimensional Geometry Topic. Indonesian Journal of Educational Research and Review, 3(2), 19-24.

[3] Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K., & vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology, 109*(5), 621–634. http://doi.org/10.1037/edu0000163

[4] Arthur, C., Badertscher, E., Goldenberg, P., Moeller, B., McLeod, M., Nikula, J., & Reed, K. (2017). *Strategies to improve all students' mathematics learning and achievement*. EDC

[5] Arthur, Y., E. Owusu, E., Arhin, A. K. (2018). Connecting mathematics to real life problems: A Teaching quality that improves students' mathematics interest. Mathematics, Education.

[6] Bellamy, N. (2015). Principles of clinical outcome assessment. *Rheumatology (Sixth Edition),* 1, 9-19. https://doi.org/10.1016/B978-0-323-09138-1.00002-4

[7] Betz, N. E., & Hackett, G. (1983). *Mathematics Self-Efficacy Scale (MSES)* [Database record]. APA PsycTests. https://doi.org/10.1037/t01563-000

[8] Bieleke, M., Goetz, T., Yanagida, T., Botes, E., Frenzel, A. C., & Pekrun, R. (2022). Measuring emotions in mathematics: The achievement emotions questionnaire—mathematics (AEQ-M). *ZDM – Mathematics Education*, 55, 269–284. https://doi.org/10.1007/s11858-022-01425-8

[9] Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. Structural Equation Modeling: A Multidisciplinary Journal, 22(4), 504-516.

[10] Borsboom, P, D., & Molenaar, D. (2015). Psychometrics. *International Encyclopedia of Social Science & Behavioural Sciences (2nd, ed.).* 418-422. https://doi.org/10.1016/B978-0-08-097086-8.43079-5

[11] Boyle, G. J., & Saklofske, D. H. (2004). Editors' introduction: Contemporary perspectives on the psychology of individual differences. *Humanities & Social Sciences papers,* 59. http://epublications.bond.edu.au/hss_pubs/5

[12] Boyle, G. J., Saklofske, D. H., & Matthews, G. (2015). *Criteria for selection and evaluation of scales and measures*. Elsevier/Academic Press

[13] Boynton, P. M., & Greenhalgh, T. (2004). Hands-on guide to questionnaire research: Selecting, designing, and developing your questionnaire. *The BMJ,* 328(7451). DOI:10.1136/bmj.328.7451.1312

[14] Buchwald, P., & Schwarzer, C. (2010). Impact of Assessment on students' test anxiety. *International Encyclopedia of Education (Third Edition),* 498-505.

https://doi.org/10.1016/B978-0-08-044894-7.00304-3

[15] Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2015). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clin Ther*, 36(5), 648-662. Doi:10.1016/j.clinthera.2014.04.006

[16] Chamberlin, S. A. (2010). A review of instruments created to assess affect in mathematics. Journal of Mathematics Education, 3(1), 167-182.

[17] Cliff, A., & Yeld, N. (Eds.). (2006). Test domains and constructs: Academic literacy. *In book: Access and Entry Level Benchmarks. The National Benchmark Tests Project (pp.19-27).*

[18] Code, W., Merchant, S., Maciejewski, W., Thomas, M., & Lo, J. (2016). The mathematics attitudes and perceptions survey: An instrument to assess expert-like views and dispositions among undergraduate mathematics students. *International Journal of Mathematical Education in Science and Technology*, 47(6), 917-937, DOI:10.1080/0020739X.2015.1133854To

[19] Creswel, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th Ed).* Pearson

[20] Demetriou, C., UzunOzer, B., & Essau, C. A. (2015). *Self-report questionnaires*. John Wiley & Sons, Inc. DOI:10.1002/9781118625392.wbecp507

[21] Dowrich, M. (2008). Teacher perceptions of the implementation of the national continuous assessment programme in a primary school in the St. George East Education District in Trinidad and Tobago, Unpublished Thesis. Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Education (Concentration in Curriculum) of The University of the West Indies.

[22] Dueber, D. M. (2017). Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. https://dx.doi.org/10.13023/edp.tool.01 [Available at http://sites.education.uky.edu/apslab/resources/]

[23] Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. Psychological Methods, 22, 541–562. https://doi.org/10.1037/met0000083

[24] Ernest, P. (2015). The social outcomes of learning mathematics: standard, unintended or visionary? *International Journal of Education in Mathematics, Science and Technology*, 3(3),

[25] Fang, G., Guo, J., Xu, X., Ying, Z., & Zhang. S. (2020). Identifiability of Bifactor Models. https://www.researchgate.net/publication/347624335_Identifiability_of_Bifactor_Models

[26] Febriyanti, R., Mustadi, A., & Jerussalem, M. A. (2021). Students' learning difficulties in mathematics: How do teachers diagnose and how do teachers solve them? *Mathematics Education Journal.* 15(1), https://doi.org/10.22342/jpm.15.1.10564.23-36

[27] Fernández-Cézar, R., Solano-Pinto, N., & Garrido, D. (2021). Can mathematics achievement be predicted? The role of cognitive–behavioral–emotional variables. *Mathematics,* 9, 1591. https://doi.org/10.3390/math9141591 https://www.mdpi.com/journal/mathematics

[28] Furr, J. (1996). *A brief history of mathematics education in America*. University of Geogia

[29] Grootenboer, P. J., Lomas, G., Ingram, N. (2005). The affective domain and mathematics education. *OAI,* DOI:10.1007/978-6091-970-1_3

[30] Gruppen, L. D. (n.d). Self-assessment in self-directed learning: Can we trust it? Do we have a choice? University of Michigan Medical School.

[31] Gyamfi, A. (2022). *Mathematics education: Any social value?* Daily Graphic

[32] Hawkins, M., Elsworth, G. R., Hoban, E., & Osborne, R. H. (2020). Questionnaire validation practice within a theoretical framework: A systematic descriptive literature review of health literacy assessments. *BMJ,* DOI:10.1136/bmjopen-2019-035974

[33] He, J., & Van de Vijver, E. J. R. (2019). Assessment of the general response style: A cross-cultural validation. *Acta de investigación psicológica*, 9(3), 14-24. DOI:https://doi.org/10.22201/fpsi.20074719e.2019.3.320

[34] Hidayatullah, A., Csíkos, C. The role of students' beliefs, parents' educational level, and

the mediating role of attitude and motivation in students' mathematics achievement. Asia-Pacific Edu Res 33, 253–262. https://doi.org/10.1007/s40299-023-00724-2

[35] *History of standardized testing in the United States*. (June 25, 2020). The National Education Association.

[36] Horvat, J. (2011). Questionnaire. In: Lovric, M. (eds) International Encyclopedia of Statistical Science. Springer. https://doi.org/10.1007/978-3-642-04898-2_55

[37] Hyman, M. R., & Sierra, J. J. (2016). Open-versus close-ended survey questions. *Business Outlook*, 14(2).

[38] Jenkins, C. R., & Dillman, D. A. (1995). The language of self-administered questionnaires as seen through the eyes of respondents.

[39] Jenn, N. C. (2006). Designing a questionnaire. *Malays Fam Physician*, 1(1), 32-35.

[40] Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. American Sociological Review, 48, 398-407.

[41] Kesiki, S. (2023). Measurement-quality calculator for confirmatory factor analysis model factors (MQC-CFAMF): An MS Excel-based tool for calculating composite reliability and average variance extracted. https://youtu.be/bt4jVOCMQOs?si=PLS829ziz0JGF5qC

[42] Kesiki, S., & Nekang, F. N. (2023). Salient and mythical sources of secondary school students' difficulties in learning mathematics in Cameroon. *International Journal of Advanced Multidisciplinary Research and Studies,* 3(6),137-154.

[43] Kilpatrick, J. (1992). A history of research in mathematics education. In D. A. Grouws (Ed.), Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics (pp. 3–38). Macmillan Publishing Co, Inc.

[44] Koch, T., & Michael Eid, M. (2024). Augmented Bifactor Models and Bifactor-(S-1) Models are Identical. A Comment on Zhang, Luo, Zhang, Sun & Zhang (2023). Structural Equation Modeling: A Multidisciplinary Journal. https://doi.org/10.1080/10705511.2024.2339387

[45] Korb, K. A. (2011). Self-report questionnaires: Can they collect accurate information? *Journal of Educational Foundations*, 1, 5-12.

[46] Kubai, E. (2019). Reliability and validity of research instruments. https://www.researchgate.net/publication/335827941_Reliability_and_Validity_of_Research_Instruments_Correspondence_to_kubaiedwinyahoocom

[47] Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology,* 9, 2531-2560. DOI:10.4236/psych.2018.911145

[48] Learning Outcome Framework. (2015). Educators' guide for pedagogy and assessment: using a learning outcomes approach. https://www.schoolslearningoutcomes.edu.mt/files/documents/01_Mathematics.144501456314.pdf

[49] Lyonga, N. A. N. (2022). Exploring students' self-assessment to increase learning outcomes in teachers' training colleges in Cameroon. Journal of learning for development, 9(2), 317-330.

[50] Mahmud, J. (2017). Item response theory: A basic concept. *Educational Research and Reviews*, 12(5), 258-266. DOI:10.5897/ERR2017.3147

[51] Manizade, A., Buchholtz, N., & Beswick, K. (2023). *The evolution of research on teaching mathematics: International perspectives in the digital era he digital era.* Springer. https://doi.org/10.1007/978-3-031-31193-2

[52] Maree, K., Aldous, C., Hattingh, A. S., & Van der Linde, M. (2006). Predictors of learner performance in mathematics and science according to a large-scale study in Mpumalanga. South African Journal of Education, 26(2), 229-252.

[53] Markon, K. E. (2019). Bifactor and Hierarchical Models: Specification, Inference, and Interpretation. Annual Review of Clinical Psychology, 15, 51–69.

[54] Mastnak, A., Zuljan, M. V., & Magajna, Z. (2023). Self-assessment by self-questioning in the instructional and practical phases of mathematics learning. *Pedagogika,* 149(1), 163–184. https://doi.org/10.15823/p.2023.149.8

[55] Max, A. –L., Lukas, S., & Weitzel, H. (2022). The relationship between self-assessment and performance in learning TPACK: Are self-

assessments a good way to support preservice teachers' learning? *Journal of Computer Assisted Learning*, 38(4), 1160-1172. http://doi.org/10.1111.jcal.12674

[56] Mazana M. Y., Montero, C. S., & Casmir, R. O. (2019). Investigating students' attitude towards learning mathematics. *International Electronic Journal of Mathematics Education*, 14(1), 207-231. https://doi.org/10.29333/iejme/3997

[57] Midena, D., & Yeo, R. (2022). Towards a history of the questionnaire. 503-529. https://doi.org/10.1080/17496977.2022.2097576

[58] Milawati. (2019). Blueprint as a base in building better teacher-made test. *Prosodi,* 13(2), 119-128. https://doi.org/10.21107/prosodi.v13i2.622

[59] Montanya, B. N. (2018). Impact of students' attitudes on mathematics performance among public secondary school students in Masaba North Sub County, Nyamira County.

[60] Ndlovu, V. (2017). Grade 10 – 12 learners' attitude towards mathematics and how the attitudes affect performance. https://core.ac.uk/download/pdf/188776028.pdf

[61] Nekang, F. N. & Anyi, R. A. (2022). The use of mock O/L mathematics examination as a predictor of the GCE performance in mathematics in secondary schools in Fako Division, South West Region of Cameroon. *International Journal of Advanced Multidisciplinary Research and Studies,* 2(3), 407-413. https://www.multiresearchjournal.com/admin/uploads/archives/archive-1655355239.pdf

[62] Ngeche. T. N. M. (2017). Student and teacher attitudes as correlates of performance in mathematics in Cameroon secondary schools. *International Journal of Humanities Social Sciences and Education (IJHSSE),* 4(12), 1-10.

[63] Ngunjiri, M. (2022). The role of assessment in mathematics classrooms: A Review. International Journal of Advanced Research, 5(1), 156-160. DOI:10.37284/ijar.5.1.887

[64] Niss, M., Pegg, J., Gutiérrez, A., & Huerta, P. (1998). Assessment in geometry. In C. Mammana & V. Villani (Eds.), *Perspectives on the teaching of geometry for the 21st century.* New ICMI study series (Vol 5, pp. 263–295). Springer. https://doi.org/10.1007/978-94-011-5226-6_9.

[65] Oakland, T. (1997). Affective assessment. Apresentado no III CONPE.

[66] OECD. (2013). Mathematics Self-Beliefs and Participation in Mathematics-Related Activities. Students' Engagement, Drive and Self-Beliefs, III, 79-104.

[67] Orrill, C. H., Gearty, Z., & Wang, K. (2023). Continuing evolution of research on teaching and learning: Exploring emerging methods for unpacking research on teachers, teaching, and learning. *The Evolution of Research on Teaching Mathematics, Mathematics Education in the Digital Era,* 22, https://doi.org/10.1007/978-3-031-31193-2_12

[68] Ortiz, E. (2016). The problem-solving process in a mathematics classroom. MSU Libraries, 1(1),

[69] Patrick C., Kyllonen, P. C., & Kel, H. (2018). Ability tests measure personality, personality tests measure ability: Disentangling construct and method in evaluating the relationship between personality and ability. *Journal of Intelligence,* 6(32), 1-26. doi:10.3390/jintelligence6030032

[70] Pekmezci, F. B. (2022). Bifactor and bifactor s-1 model estimations with non-reverse-coded data. Journal of measurement and evaluation in education and Psychology, 13(3), DOI:10.21031/epod.1135567.

[71] Pekrun, R. (2020). Commentary: Self-report is indispensable to assess students' learning. *Frontline learning Research.* 8(3), 185 – 193. https://files.eric.ed.gov/fulltext/EJ1260612.pdf

[72] Platek, R. (1985). Some important issues in questionnaire development. *Journal of Official Statistics,* 1(2), 119-136.

[73] Qassimi, N. M. (2021). Standardized testing. https://www.researchgate.net/publication/351819226_Standardized_Testing

[74] Radišić, J. (2023). Student mathematics learning outcomes. In: Manizade, A., Buchholtz, N., Beswick, K. (eds) *The evolution of research on teaching mathematics. Mathematics Education in the Digital Era*, vol 22. Springer, Cham. https://doi.org/10.1007/978-3-031-31193-2_7

[75] Reise, S. P. (2012) The Rediscovery of Bifactor Measurement Models. Multivariate Behavioral Research, 47(5), 667-696. DOI:10.1080/00273171.2012.715555

[76] Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. Journal of Personality Assessment, 95(2), 129-140.

[77] Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Qual Life Res. 16(1), 19-31. doi:10.1007/s11136-007-9183-7

[78] Reusser, K. (2000). Success and failure in school mathematics: effects of instruction and school environment. *European Child & Adolescent Psychiatry*, 9, II/17–II/26

[79] Reys, R. E., & Rea, R. E. (1970). The comprehensive mathematics inventory: An experimental instrument for assessing the mathematical competencies of children entering school. *Journal for Research in Mathematics Education*, 1(3), 180-186. https://doi.org/10.2307/748336

[80] Robins, R. W., Fraley, R. C., & Krueger, R. F. (Eds.). (2007). *Handbook of research methods in personality psychology*. The Guilford Press

[81] Roopa, S., & Rani, M. S. (2017). Questionnaire designing for a survey. *J Ind Orthod Soc 2012*, 46(4), 273-277. https://www.researchgate.net/publication/235801675_Questionnaire_Designing_for_a_Survey

[82] Saftari, M., & Fajriah, N. (2019). Assessment of affective domain in attitude scale assessments to assess learning outcomes. *Jurnal Ilmu Pendidikan dan Kependidikan*, 7(1), 71-81. https://doi.org/10.35438/e.v7i1.164

[83] Saini, M., & Arora, M. (2017). Development of the questionnaire for academic performance in mathematics. *International Journal of Interdisciplinary Research in Science Society and Culture (IJIRSSC),* 3(1), 41-58.

[84] Sarifah, I., Kuswati, E., Suniasih, W., Fahrurrozi, F., & Susilo, G. (2024), Cognitive test instruments to measure student mathematics ability in elementary school: Rasch analysis. *Educational Administration: Theory and Practice*, 30(2), 767-799. Doi:10.53555/kuey.v30i4.1557

[85] Sayegh, S. (n.d). Direct & indirect assessment: What's the difference? University of Portland

[86] Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2012). *General overview of the theories used in assessment.* Amee Guide

[87] Shou, Y., Sellbom, M., & Chen, H. -F. (2022). Fundamentals of measurement in clinical psychology. *Comprehensive Clinical Psychology,* 4, 13-35. https://doi.org/10.1016/B978-0-12-818697-8.00110-2

[88] Sullivan, G., & Artino Jr., A. R. (2013). Analyzing and interpreting data from likert-type scales. Journal of Graduate Medical Education. 5(4), pp. 541-542. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3886444/

[89] Suresh, K. (2017). How to standardize the research tool correctly?–The Answer. *Shanlax International Journal of Education,* 6(1), 149-153. https://www.researchgate.net/publication/346416336_How_to_Standardize_the_Research_Tool_Correctly_-_The_Answer

[90] Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., Silver, E., Ufer, S., & Pauline Vos, P. (2016). Assessment in mathematics education. In: Assessment in mathematics education. ICME-13 Topical Surveys. Springer, Cham. https://doi.org/10.1007/978-3-319-32394-7_1

[91] Taherdoost, H. (2016). Validity and reliability of the research instrument; How to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management (IJARM),* 5(3), 28-36. DOI:10.2139/ssrn.3205040

[92] UNESCO. (2018). Self-report indirect and simple direct assessment tools for reporting: Concept paper for GAML 5. https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/12/4.6.1_11_Self-Report-Indirect-and-Simple-Direct-Assessment-Tools-for-Reporting.pdf

[93] Vos, P. (2005). Measuring mathematics achievement: A need for quantitative methodology literacy. In J. Adler &amp; M. Kazima (Eds.), *Proceedings of the 1st African regional congress of the international commission on mathematical instruction (ICMI). Johannesburg, South Africa: University of the Witwatersrand (PP. 264-271).*

[94] Willem, E., Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research (2nd ed.).* John Wiley & Sons, Inc.

[95] Willis, G. B. (2019). Questionnaire design, development, evaluation, and testing: Where are we, and where are we headed? https://doi.org/10.1002/9781119263685.ch1

[96] Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Arch Psychiatry*, 26(3), 171-7. doi:10.3969/j.issn.1002-0829.2014.03.010.

[97] Yantini, N. T. M. I., Suarjana, M., Sumantri, M. (2021). Assessment of mathematics learning outcomes of fourth-grade elementary school students. *Jurnal Ilmiah Sekolah Dasar*, 5(3), 452-460. https://ejournal.undiksha.ac.id/index.php/JISD/index

[98] Yudha, R. P., Anggara, D. S., & Zulaeha, O. (2019). Authentic assessment instruments for performance in mathematics learning in elementary schools. *Journal of Physics: Conference Series,* 1321, 1-13. doi:10.1088/1742-6596/1321/3/032012

[99] Yusoff, M. S. B., Arifin, W. N., Hadie, S. N. H. (2021). ABC of questionnaire development and validation for survey research. *Education in Medicine Journal,* 13(1), 97–108. https://doi.org/10.21315/eimj2021.13.1.10

[100] Zhao, N., Valcke, M., Desoete, A., Sang, G., & Verhaeghe, J. P. (2012). A holistic model to infer mathematics performance: the interrelated impact of student, family and school context variables. *Scandinavian Journal of Educational Research*, 1–20.