Tanzania Journal of Science 49(4): 921-929, 2023
ISSN 0856-1761, e-ISSN 2507-7961

Digitisation of Classical Exercise Practices with STACK: Management of Large Mathematics Classes in Higher Education Institutions

Idrissa S. Amour[1*], Fatuma Simba[2], Septimi Kitta[3] and Abdi T. Abdalla[4]

[1]*Department of Mathematics, University of Dar es Salaam, P.O. Box 35062, Dar es Salaam, Tanzania.*

[2]*Department of Computer Sciences, University of Dar es Salaam, P.O. Box 33335, Dar es Salaam, Tanzania.*

[3]*Department of Educational Psychology and Curriculum Studies, University of Dar es Salaam, P.O. Box 35048, Dar es Salaam, Tanzania.*

[4]*Department of Electronics and Telecommunications Engineering, University of Dar es Salaam, P.O. Box 33335, Dar es Salaam, Tanzania.*

[*]*Corresponding author, e-mail: Idrissa.Amour@gmail.com*

Co-authors' e-mails: fatmasimba@gmail.com; sekitta@yahoo.com; abdit@udsm.ac.tz

Received May 2023, Revised 30 July 2023, Accepted 29 Aug 2023 Published Oct 2023

DOI: https://dx.doi.org/10.4314/tjs.v49i4.13

Abstract

Management of large classes' tutorials is a known problem in Mathematics. Engineering mathematics classes enrol around 1000 students at the University of Dar es Salaam. For effective learning, each tutorial session should register not more than 40 students. This requires at least 25 sessions of tutorials per week, which may not be feasible due to both scarce human resources and venues. In this work, we developed online Mathematics exercises based on the System for Teaching and Assessment using Computer Algebra Kernel (STACK). About 300 STACK questions in linear algebra, calculus, complex numbers, and numerical analysis were constructed for weekly tutorials and quizzes. Students were given an unlimited number of attempts in tutorials and only one attempt for quizzes. The quality of the questions was analysed by examining their facility indices and discriminative efficiencies. Majority of the questions (87%) were within acceptable region. The questions, therefore, provided reasonable insight as appropriate alternative to classical practice. Competent authoring of STACK questions can improve the quality of teaching and learning of Mathematics and save scarce human and material resources required to serve large classes. This can also address the issue of running online programmes in Mathematics and computational subjects to support distance learning.

Keywords: Discriminative efficiency, Facility index, Mathematics large classes, Mathematics digital assessment, STACK.

Introduction

Tutorials are a vital part of the teaching and learning process at higher learning institutions where students practice for mastering the subject content. Each course in Mathematics at the University of Dar es Salaam is allocated at least one mandatory session for tutorial per week (UDSM 2021). When the class is large, however, the number of sessions increases proportionally. The number of students in each session must be reasonable for effective discussion and learning. Students' achievements in Mathematics are related to what they weekly practice in small groups alongside the lectures (van Veggel and Amory 2014). E-learning

platforms like Moodle are well known for their capability in handling objective type questions, like Multiple Choice Questions (MCQ), Matching Items Questions (MIQ), True False Questions (TFQ), Short Answers Questions (SAQ) and the most high level type Pattern Match Questions (PMQ), where the pattern of words and its alternatives are defined as answers for matching with the students' inputs. All these question types are not suitable for mathematics questions which require testing students' answers using mathematics rules (commutative, associative, and distributive properties). There exist a number of mathematics engines which can be embedded into e-learning platforms like Moodle, for example, System for Teaching and Assessments using Computer Algebra Kernel (STACK) (Sangwin 2013), WeBWorK, and MyOpenMath (Gage 2017). Developing quality mathematics STACK questions require strong command of computer programming with Maxima; a system for the manipulation of symbolic and numerical expressions, including differentiation, integration, Taylor series, Laplace transforms, ordinary differential equations, systems of linear equations, polynomials, sets, lists, vectors, matrices and tensors (Maxima 2023). This is what makes some instructors develop questions with no or less programming efforts, which can be easily graded; very similar to SAQ or traditional MCQ (Gage 2017). This practice may degrade the competence of the students and the society as it deviates from the best teaching and learning practices.

Determining whether a question is appropriate to reflect the objective of the course is an important aspect of any assessment procedure. Johari et al. (2011) studied the achievement of a University programme by examining the difficulty of the tests' questions from the Departments of Mechanical and Material Engineering. However, the authors did not examine whether the questions can distinguish students of different learning abilities. Creating meaningful assessment tools is not obvious; Khairani and Shamsuddin (2016) reported that a number of teachers failed to create appropriate questions. It was not possible to establish a correlation between the difficulty of the questions and ability to distinguish learners of different abilities (Khairani and Shamsuddin 2016). The use of psychometric analysis; facility index and discrimination index embedded within Moodle has been studied and used to analyse the Moodle objective type questions' qualities (Gamage et al. 2019). Based on the quality measures, the authors made decision on whether to keep current questions or improve them (Gamage et al. 2019). To the best of our knowledge, the Psychometric analysis for the STACK questions was however yet to be reported.

The use of STACK questions in teaching and learning Mathematics around the world has received increased attention over the past decade. It has been reported that a number of authors have successfully employed STACK in their teaching practice of Mathematics; for example, Calculus and Linear Algebra courses (Ellis et al. 2015, Pauna 2017, Kinnear 2020, Davies et al. 2022). The only reported setback in using STACK is the difficulty of authoring quality questions (Gage 2017), as they require good commands in both Maxima programming and mathematics properties, especially when randomisation of question parameters is chosen. However, this difficulty is tolerable as a successful setup of course questions will save much time and resources in the future. It is considered as one-time capital investment.

In this work, we developed quality mathematics STACK questions for two Engineering Mathematics undergraduate courses offered by the Department of Mathematics of the University of Dar es Salaam. The questions setting focuses on having a similar effect on classical paper-and-pencil questions. The quality of the questions was analysed based on the difficulty of each question as well as its ability to distinguish students of different learning abilities. Appropriate measures of difficulty level and ability to distinguish students of different learning abilities were analysed and discussed.

Conceptual Framework

Behind the STACK, there is Maxima; a Computer Algebra System (CAS) (Sangwin 2013). In STACK question, a teacher can test students' answers using any mathematics property based on the requirement of the question (Lowe et al. 2019). The power of STACK is associated with its Potential Response Tree (PRT), which allows feedback provision on specific or anticipated students' mistakes, but also provides a graphical branching programming mechanism, for example in grading of multi-method questions, as shown in Figure 1 of Cauchy Euler differential equation's question, whose preview is shown in Figure 3. Node 1 in Figure 3 is not used to grade anything but to test the choice of method which is controlled by the Boolean variable (as seen in Code 1) associated with the two methods (Variation of parameters (VP) and method of undetermined coefficients (UC)).

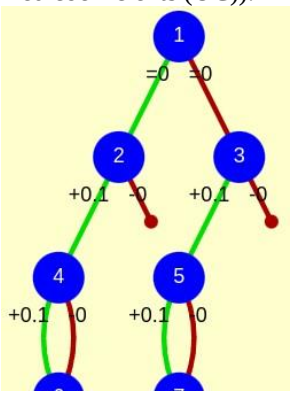

Figure 1: Section of PRT used for grading between Variation of Parameters or Undetermined Coefficients methods, determined by Boolean flag at node 1.

Code 1: The Boolean flag uc to detect student's choice of solution method (UC or VP) imposed at node 1 of Figure 1

```
uc: if is(ans5 = 1) then true
else false;
```

Single PRT is also used to grade multipart questions, where parts of the questions are related; automatic grading can be stopped at any node where necessary. Figure 2 shows the grading of multipart questions at nodes 1 and 2, while nodes 3 and 4 are used to give feedback for anticipated mistakes. This makes the practice to be more relevant as assessment for learning; as defined in Black et al. (2003).

STACK has its own way of managing objective type questions; MCQ, TFQ and MIQ. The advantage of using these types of questions over the default Moodle's questions is that they provide the benefits of CAS integration. Just to mention examples of the advantage, one can have a MCQ, TFQ or MIQ with curve plots as choices or having question variables' values defined as random variables. The STACK objective questions are very useful to grade for example intermediate steps of a comprehensive solution or testing students reasoning along the way of the question solution. The objective type Moodle questions must be used as standalone questions.

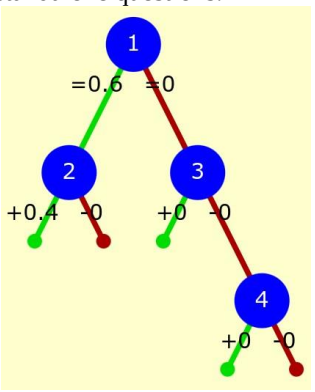

Figure 2: Grading of multipart related questions at nodes 1 and 2 alongside feedback provisions at nodes 3 and 4.

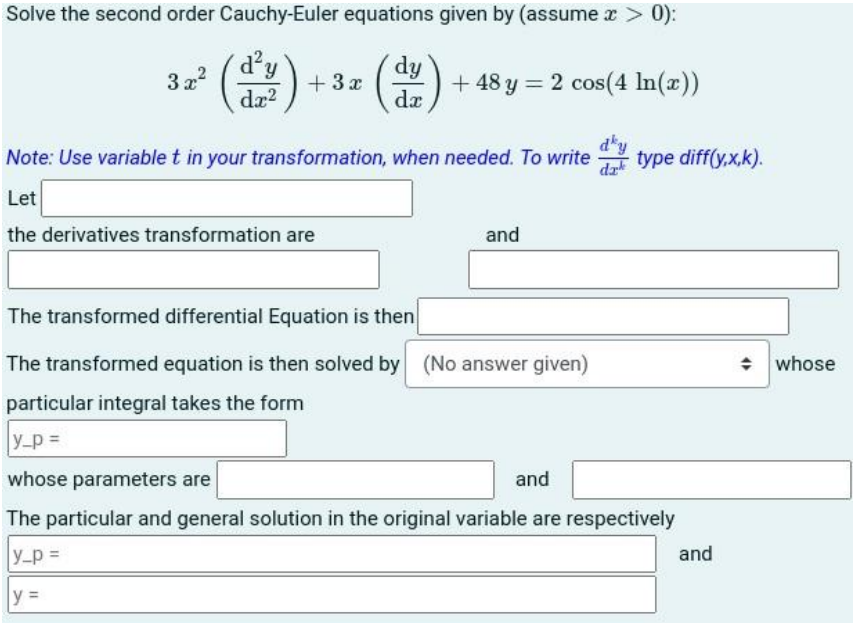

Figure 3: STACK question's preview of the Cauchy-Euler differential equation to be solved by either VP or UC method.

Methodology

In this work, we investigated management of online tutorials of undergraduate engineering Mathematics courses. Two courses from the University of Dar es Salaam: One Variable Calculus and Differential Equations for Non-Majors (MT171); and Matrices and Basic Calculus for Non-Majors (MT161) were used. The contents of the courses cover the topics in calculus, linear algebra, complex numbers, and numerical analysis. Each course enrolled around 1000 students. The courses were run in two different semesters; MT171 ran for semester 2 of 2021/2022 academic year, while MT161 ran for semester 1 of the 2022/2023 academic year.

The practice had two activities per week: a tutorial where students had an unlimited number of attempts, for mastering subject matter, and only one attempt quiz as assessment of learning, which plays a central role in the teaching and learning process. To set ourselves free from the accusation of authoring easily gradable questions, we have adapted and digitised the questions of classical tutorials of the same course from previous years. The practice involved solving the problems with variables and randomise the variables for different question variants, as smartly supported by STACK (Sangwin 2013). Because tutorials were unlimitedly attempted, it would make no sense to analyse the quality of the questions based on students' performance. That is why this study analysed the quality of STACK questions from the quizzes only.

The analyses of the questions were based on the psychometric analysis of the test questions; namely the difficulty index and discrimination index. The difficulty level (P) of a question (or easiness or facility index or P-value) is defined as the percentage of students who answered the question correctly (Rezigalla 2022);

$$P = \frac{R}{T} \times 100 \qquad (1)$$

where, R is the number of students who answered the question correctly, and T is the number of students who attempted the question (Mahjabeen et al. 2017). One interpretation of the facility index values, as documented in Mahjabeen et al. (2017) and Rezigalla (2022), is shown in Table 1.

Table 1: Facility indices regions

High (Difficult)	Medium (Moderate)	Low (Easy)
$P < 30\%$	$30\% < P < 80\%$	$P > 80\%$

The discrimination index (D) is the ability of a question to discriminate between students of different learning abilities. It is defined as a measure that tells the degree to which a question distinguishes the students who performed well from those who performed poorly (Rezigalla 2022);

$$D = \frac{R_U - R_L}{\frac{1}{2}T} \qquad (2)$$

where, R_U is the number of students in the upper group who answered the question correctly, and R_L is the number of students in the lower group who answered the question correctly (Mahjabeen et al. 2017). The discrimination index of a question is calculated by categorising the students into upper 27% group and lower 27% group according to their total test score (Mahjabeen et al. 2017, Rezigalla 2022). The discrimination index value lies between -1.0 and $+1.0$, i.e $-1.0 \leq D \leq +1.0$. The positive value indicates that high performers answer the question correctly more than those the lower ones, which is acceptable. The negative value tells that lower performers students answer the question more correctly than the upper performers, this is not desired information. The zero value tells equal numbers of students in the upper and lower groups who answered the question correctly (Rezigalla 2022). Negative discrimination could possibly be due to question flaws or question ambiguity. A question with poor discriminating index will never provide an appropriate interpretation of the student's actual ability (Sugianto 2020). The Discriminative Efficiency, (d), is another term which suggests how good is the discrimination index relative to the difficulty of the question (Hofmann 1975). The values of d ranges between 0 and 1, i.e., $0 \leq d \leq 1$, therefore, shall provide additional information for question analyses using the same scale as in the facility index.

In this work, the facility indices together with discriminative efficiencies were analysed as reported having positive relation for a meaningful question (Fozzard et al. 2018, Ramzan et al. 2020). Both values of facility indices and discriminative efficiencies of each question were extracted from Moodle quiz reports' statistics.

Results and Discussion

In this work, about 300 STACK questions were developed to cover the contents of the two courses (MT161 and MT171). The quality of the questions was analysed based on their facility indices and discriminative efficiencies, as also used by Gamage et al. (2019).

Tables 2 and 3 show the total number of questions in each category and their intersection regions, respectively, as suggested in Gamage et al. (2019) and Fozzard et al. (2018) too. It can be observed that 96% and 80% of the questions from MT171 and MT161, respectively, were within acceptable region and therefore require no attention. But, 4% and 20% of the questions (marked with asterisks) from MT171 and MT161, respectively, require revisions; to make them suitable to achieve and test intended learning.

Table 2: Number of questions in each region defined by the facility index and discriminative efficiency regions for MT161

Discriminative Efficiency, d	Facility Index		
	High	Medium	Low
$d \leq 30\%$	$0*$	$1*$	$3*$
$30\% \leq d \leq 50\%$	0	$14*$	8
$d \geq 50\%$	2	42	16

Table 3: Number of questions in each region defined by the facility index and discriminative efficiency regions for MT171

Discriminative Efficiency, d	Facility Index		
	High	Medium	Low
$d \leq 30\%$	$1*$	$2*$	$0*$
$30\% \leq d \leq 50\%$	7	$0*$	0
$d \geq 50\%$	14	46	0

Table 4 shows the total number of questions for both courses in each category and their intersection regions. It can be observed from Table 4 that a total of 135 out of 156 (approximately 87%) questions are of good quality, while the remaining questions require revision (marked with asterisks).

Table 4: Number of questions in each region defined by the facility index and discriminative efficiency regions for both MT161 and MT171

Discriminative Efficiency, d	Facility Index		
	High	Medium	Low
$d \leq 30\%$	$1*$	$3*$	$3*$
$30\% \leq d \leq 50\%$	7	14^*	8
$d \geq 50\%$	16	88	16

The proportions of the number of questions in each quality measure category are visualized by the plots in Figures 4 and 5 for MT161 and MT171, respectively. It can be seen that MT171 has more difficult questions (bars with square mesh in Figures 5) than MT161 (bars with square mesh in Figure 4). However, those questions have good discriminative efficiencies, hence the difficulty is not a problem and justified.

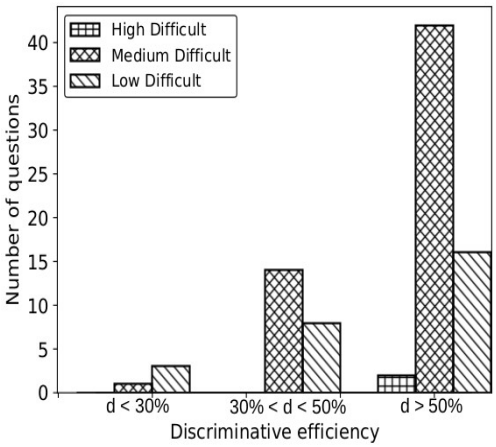

Figure 4: Number of developed questions with respect to each category for MT161.

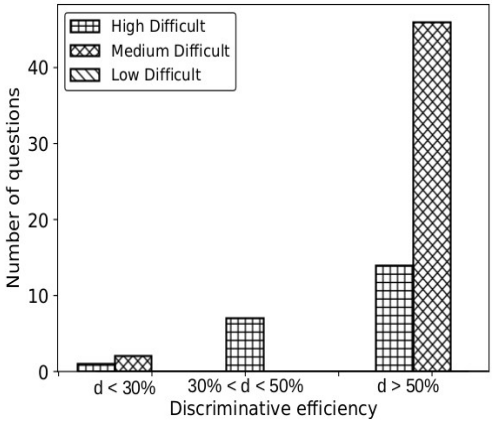

Figure 5: Number of developed questions with respect to each category for MT171.

Figure 6 shows the quality measures in one plot. It can be seen, from the first three bars in Figure 6, that very few number of questions with $d \leq 30\%$ were developed. The medium difficulty questions with discriminative efficiency $30\% \leq d \leq 50\%$ in Figure 6 do not appropriately reflect learning, and therefore not a desired property, as discussed in Gamage et al. (2019).

Figures 7 and 8 show the facility index and discriminative efficiency plots for the quizzes' questions of MT171 of academic year years 2021/2022 and MT161 of the academic year 2022/2023, respectively. Figure 7 depicts very good number of questions with acceptable facility indices and discriminative efficiencies. The acceptable minimum value of each measure is shown by a dashed line border. The horizontal continuous line border shows the lower boundary of the acceptable discriminative efficiency and the vertical continuous border shows the upper boundary of moderate acceptable values of facility index. On the other side, Figure 8 shows the improved quality of the questions in terms of both facility index and discriminative efficiency.

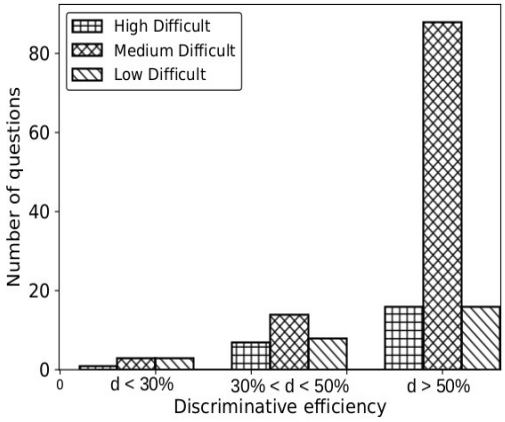

Figure 6: Number of developed questions with respect to each category for both courses.

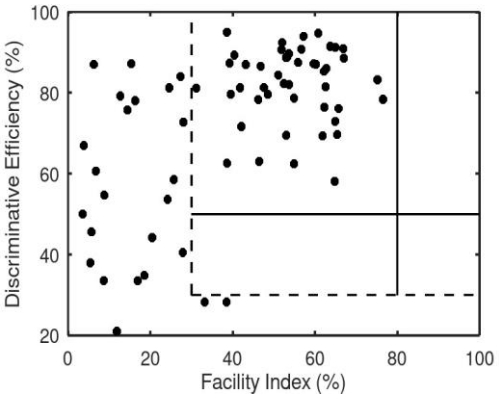

Figure 7: Facility index and discriminative efficiency for MT171.

Figure 9 shows the facility indices and discriminative efficiencies for all courses altogether. The figure shows few questions were falling outside the acceptable region. Most of these questions were from MT171. Some questions of MT171 look more difficult and not able to distinguish students of different learning abilities. This can be justified by the fact that MT161 (taught in semester 1 2022/2023) followed after MT171 (taught in semester 2 2021/2022), whose experience changed the way of developing the questions.

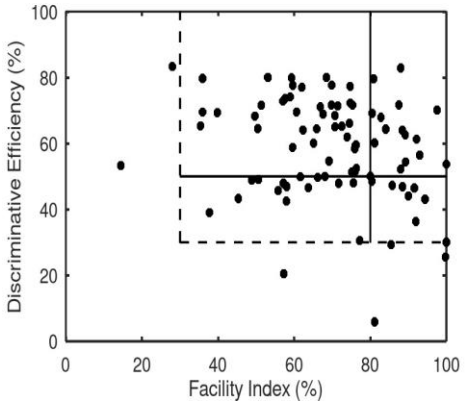

Figure 8: Facility index and discriminative efficiency for MT161.

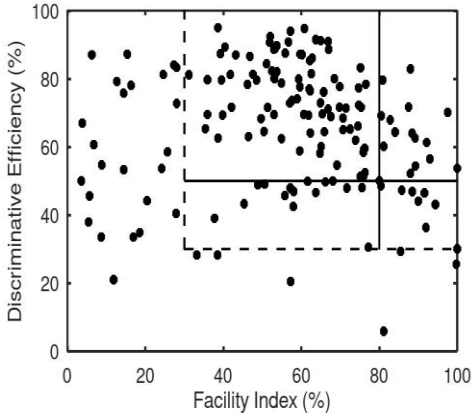

Figure 9: Facility index against discriminative efficiency for the two courses.

Analysis of questions helps in determining the competence of instructors in creating quality questions, as demonstrated in Khairani and Shamsuddin (2016). The correlation between the difficulty index and discriminative efficiency could not be established, as evident from Figures 7–9. This finding is also supported by Khairani and Shamsuddin (2016). This can be explained by the fact that these measures depend solely on instructor's competence on setting-up the questions.

The appropriateness of each question can be decided based on the region of its facility index and discriminative efficiency. Thirteen percent (13%) of the questions developed in this work should be revised, this amount is

reasonable and was expected, similar report is found in Khairani and Shamsuddin (2016).

Development of high quality STACK questions (as opposed to easily gradable questions) required significant amount of time and effort resources (Gage 2017). In this study, this would not be realistic if we were not timely and resourcefully prepared. This high development cost is tolerable and justified by the fact that the questions resources will be used for coming years and also create a base for other courses' development.

Conclusion

About 300 STACK questions have been developed to manage online tutorials and quizzes for the two courses. The quality of the questions was analysed using the difficulty index and discriminative efficiency. The results suggest revision on only 13% of the questions. This suggests that 87% of the developed questions were of good quality in terms of difficulty level as well as ability to distinguish students of different learning abilities as seen in Figures 6 and 9. These questions are worth keeping and can serve as a benchmark for future questions developments.

Development of high quality STACK questions shall give us a mandate of replacing classical tutorial sessions with online sessions, which require less human and material resources. Large classes' tutorials management shall, therefore, be effective and less expensive to run. This shall also open the doors to curriculum developers at higher educational institutions to work on the possibility of establishing online programmes in mathematics, science and engineering, where the barrier of conducting online assessments is currently relaxed.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgment

This work is funded by the University of Dar es Salaam under competitive research grant awards; registered as project number CoNAS-MT22044.

References

Black P, Harrison C, Lee C, Marshall B and William D 2003 Assessment for Learning-putting it into practice. Open University Press, Maidenhead, UK.

Davies B, Smart T, Geraniou E and Crisan C 2022 STACKification: automating assessments in tertiary mathematics. In: *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education (CERME12)*, Feb 2022 ⟨hal-03750584⟩. European Society for Research in Mathematics Education. Bozen-Bolzano, Italy.

Ellis J, Hanson K, Nunez G and Rasmussen C 2015 Beyond plug and chug: an Analysis of Calculus I homework. *Int. J. Res. Undergrad. Math. Educ.* 1(2): 268–287.

Fozzard N, Pearson A, du Toit E, Naug H, Wen W and Peak IR 2018 Analysis of MCQ and distractor use in a large first year Health Faculty Foundation Program: Assessing the effects of changing from five to four options. *BMC Med. Educ.* 18: 252.

Gage ME 2017 Methods of Interoperability: Moodle and WeBWork. *J. Learn. Anal.* 4(2): 22–35.

Gamage SHPW, Ayres JR, Behrend MB and Smith EJ 2019 Optimising Moodle quizzes for online assessments. *Int. J. STEM Educ.* 6(1): 27.

Hofmann RJ 1975 The concept of efficiency in item analysis. *Educ. Psychol. Meas.* 35(3): 621–640.

Johari J, Sahari J, Abd Wahab D, Abdullah S, Abdullah S, Omar M and Muhamad N 2011 Difficulty index of examinations and their relation to the achievement of programme outcomes. *Procedia-Soc. Behav. Sci.* 18: 71–80.

Khairani AZ and Shamsuddin H 2016 Assessing item difficulty and discrimination indices of teacher-developed multiple-choice tests. In: *Assessment for Learning Within and Beyond the Classroom: Taylor's 8th Teaching and Learning Conference 2015*

Proceedings (pp. 417-426). Springer Singapore.

Kinnear G 2020 Using JSXGraph for diagrams and interactivity (Demonstration slides), School of Mathematics, The University of Edinburgh. URL https://eams.ncl.ac.uk/sessions/2020/using-jsxgraph-for-diagrams-and-interactivity/slides.pdf

Lowe T, Sangwin C and Jones I 2019 Getting started with STACK Loughborough University URL https://docs.stack-assessment.org/content/2019-STACK-Guide.pdf.

Mahjabeen W, Alam S, Hassan U, Zafar T, Butt R, Konain S and Rizvi M 2017 Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Ann. Pak. Inst. Med. Sci*. 13(4): 310-315.

Maxima 2023 A Computer Algebra System (Maxima Website). Accessed on April 14, 2023 URL https://maxima.sourceforge.io/

Pauna MJ 2017 Calculus course assessment data. *J. Learn. Anal.* 4(2): 12-21.

Ramzan M, Imran S, Bibi S, Khan K and Maqsood I 2020 Item analysis of multiple-choice questions at the Department of Community Medicine, Wah Medical College, Pakistan. *Life Sci.* 1(2): 60-63.

Rezigalla AA 2022 Item analysis: Concept and application. In: Firstenberg MS and Stawicki SP (Editors), Medical Education for the 21st Century chapter 9, IntechOpen, Rijeka URL https://doi.org/10.5772/intechopen.100138.

Sangwin C 2013 Computer Aided Assessment of Mathematics, University Press Oxford, UK ISBN 978-0-19-966035-3.

Sugianto A 2020 Item analysis of English summative test: EFL teacher-made test. *Indonesian EFL Research and Practices* 1(1): 35–54.

UDSM (University of Dar es Salaam) 2021 Undergraduate Prospectus 2021/2022, Dar es Salaam University Press.

van Veggel N and Amory J 2014 The impact of maths support tutorials on mathematics confidence and academic performance in a cohort of HE animal science students. *PeerJ* 2: e463.