Massification in Universities: Are Assessment Tools Still Reliable? A Reflection from Sokoine University of Agriculture, Tanzania

James George Mayeka* Ernest Simon Kira
School of Education, Sokoine University of Agriculture, P.O Box 3038, Morogoro, Tanzania

Abstract
A tremendous increase of the number of students in universities has been experienced by almost every country all over the world including Tanzania. The Increasing number of students has greatly affected the instructors' workload and general practices of student's assessment and evaluation. This study aimed at determining the reliability of the assessment tools at Sokoine University of Agriculture. Retrospective record review was done on education undergraduate students who sat for an EDP 100 in 2014/2015, 2015/2016 and 2017/2018 academic years where the course was selected through random procedures. A total of 214 scripts were systematically randomly sampled from each cohort. The results revealed a drop in internal consistency of the scores obtained from EDP 100 course across the three cohorts. Majority of the questions for the EDP 100 though were moderately difficulty, their discrimination powers were poor. However, the variation in difficulty and discrimination indices for the three cohorts was statistically not significant ($p>0.05$ for MCQ and MIQ) except the discrimination index for MIQ which shows significant variations ($p<0.05$). It is therefore recommended that similar studies should be done to determine both validity and reliability of the assessment tools for the other subjects at the University.

Keywords: Massification, Internal consistency, Difficulty Index, Discrimination Index
DOI: 10.7176/JEP/12-23-07
Publication date: August 31st 2021

1. Introduction

A tremendous increase of the number of students in universities has been experienced by almost every country all over the world. While, the global universities' enrolment has risen from 13.8% in 1990 to 29% in 2010, Sub-Saharan Africa has experienced a doubling of gross enrolment ratios from 3% in 1990 to 7% in 2010 (Hornsby & Osman, 2014). In Tanzania, the situation has become more evident in the recent past (Kapinga & Amani, 2016). According to Memba & Feng (2016), students' enrolments in Tanzanian universities increased from 98,915 to 354,430 between 2008/2009 and 2015/2016 academic years, respectively. Sokoine University of Agriculture which is one of the public universities in Tanzania was established in 1st July, 1984 (Sokoine University of Agriculture, 2007). Since its establishment, the university has also been experiencing the massive increase of the number of student just like other universities in the country. For example, the number of students raised almost four times from 2729 in 2008/2009 to 8296 in 2016/2017 academic years. Following this increase in number of students in universities, the instructor-student ratio has been greatly affected leading to ineffective provision of quality teaching and student assessments (Ntim, 2016). Large classes in education institutions affect much the interaction among instructors and students. Increase in numbers of students lead to poor communications among instructors with their students and the general practices of designing and using appropriate assessment tools (Alomari & Akour, 2014). Large classes hinder instructors to organize quizzes and regular class tests resulting into inefficient assessment of teaching and learning process (Yelkpieri, Namale, Esia-donkoh & Ofosu-dwamena, 2012). The increase in number of students in any education institution has turned the normal way of conducting assessment among students in universities. Regardless of the increasing number, universities would wish to maintain the quality of the programs offered. One of the means of maintaining quality of training is through effective evaluation of teaching and learning process. Effective evaluation requires valid and reliable assessment tools. Therefore, the need to check for internal consistency of the assessment tools used for teaching and learning in Tanzanian universities is one of the important aspects for effective assessment.

2. Statement of the Problem

Increasing number of students in universities which does not equally match recruitment of instructors has greatly affected the instructors' workload. With large classes, tutorials and practical sessions which were considered to be important element of learning has been replaced by examination papers or reports (Mohamedbhai, 2008). Examinations have been held more frequently and lecturers often repeat the same exams papers to different groups of students (Mohamedbhai, 2008). Furthermore, the nature of examinations questions have also changed greatly as most lecturers prefer multiple choices and short answers questions which are easier to mark and serve time (Chan, 2010). These objective questions are not necessarily bad as research shows that they can cover a wide range of content taught compared to essay questions. Also, such questions can measure even higher cognitive levels of learning when carefully constructed (Scully, 2017). Therefore, considering the situation of instructors' workload due to increased number of student's enrolment, one may not be certain on the attention required to ensure effectiveness in both teaching and assessment. This is the reason why determining the consistency of the tools used by instructors in assessing student learning outcomes created the desire for conduction of this study.

3. Objectives of the Study

The main objective of this study was to explore the reliability of the university examinations across years as the numbers of student's increases. Specifically, the study intended to:

i. Examine the internal consistency of the introduction to Educational Psychology (EDP 100) University examinations across three years at Sokoine University of Agriculture.
ii. Assess the difficulty and discrimination indices of the introduction to Educational Psychology (EDP 100) University Examination items across three years at Sokoine university of Agriculture.
iii. To determine whether difficulty and discrimination indices vary significantly across years.

4. Research Questions

i. What are the average values of the internal consistency of the EDP 100 University examinations at Sokoine University of Agriculture for a period of three years?
ii. What are the average values of difficulty and discrimination indices of the examination items used in EDP 100 across three years at Sokoine University of Agriculture?
iii. Do difficulty and discrimination indices vary significantly across years?

5. Literature Review

5.1 Reliability of an Assessment Tool

Assessment of learning outcomes in any education institutions is a crucial thing due to its diagnosis role, improving teaching process and student leaning (Tremblay, Lalancette & Roseveare, 2012). Assessments are acknowledged as the most powerful educational tools for promoting effective student learning and that is what instructors can do to help their students to learn (Rahman & Majumder, 2014). The assessment of learning outcomes involves various assessment tools that have been used. Assessment of learning outcomes in the classes and in education institution in general can be achieved through the use of various types of testy items and techniques such as multiple choice, short answer response, true or false, essay questions, portfolio, tutorial, practical, observation, checklist, anecdotal, assignment and projects (Miller, Linn & Gronlund, 2009; Omari, 2006). In order for these assessment tools to be valid, they must also be reliable so as bring out the desired outcomes. Assessment tool reliability is concerned with the ability of a tool to measure consistently the desired learning outcomes (Tavakol & Dennick, 2011). It is the consistency of a measurement (Miller, Linn, & Gronlund, 2009). Reliable assessment tool should ensure that test scores are stable and free from measurement errors (Ghazali, 2016). Reliability exists in several forms such as test-retest, inter-rater, equivalent forms and internal consistency (Ursachi, Horodnic & Zait, 2015); (Oliveira *et al*., 2016). While, test-retest checks what happens with instrument in time by the assumption that there are no substantial changes in the construct being measured between two different occasions, inter-rater tells about the consistency of different investigators to obtain the same results using the same tool (Ursachi, Horodnic & Zait, 2015). Equivalent form involves the concurrent administration of two parallel or alternate forms of the assessment tool to the same students and obtains the correlation coefficient (Ajayi, 2013). Internal consistency reliability evaluates the consistency of results across factors within a test (Hajjar, 2018). It indicates whether items on a test that are intended measure the same construct and produces consistent score (Tang, Cui & Babenko, 2014). Furthermore, the examinations of individual scale items for deviation from particular factors are ensured by internal consistency (Harms & Biocca, 2004).

Reliability of an assessment tools are affected by various factors. Zhu and Han (2011) observed three factors that affect the reliability of the test. Firstly, change of candidates and testing process. This is attributed by either the change of true score due to change of candidates language ability or misleading test results of which is due to affected real language level of a candidate. Secondly, testing features; these include things like the length and the difficulty of the test paper. The longer the paper always shows more reliability than shorter ones. This is due to the fact that, the more the contents are in the paper, the bigger scale there is in it. It follows that, if there are more representative content in the paper, the reliability of the paper will be more complete. Also the degree of testing difficulty and division will also affect test reliability. This is due to the fact that, if in the test there are questions that are either very difficult or very easy, the reliability of the test will be influenced by both aspects. Thirdly, methods of going over the test paper of which are influenced by mistakes during the process of going over the test paper tend to lower the reliability of the test. Objective questions do not require any subjective judgment so that it can achieve high reliability. This is contrary to subjective questions that need people's subjective judgment and hence affect the reliability of the paper. Also, Kinyua and Okunya (2014) added that, the improper use of bloom's taxonomy in test construction, ambiguity of the test items and poorly written questions prompt students guessing, and hence in turn tend to lower the reliability of the test paper. Furthermore, reliability of the items may be affected by expressions attributed by insufficient information and use of terms that led to misunderstanding and biases in composing the question items (Ercan, Yazici & Sigirli, 2007).

5.2 Difficulty and discrimination Indices of the Test Item

Educators perform what is called item analysis after administering an examination on students (Khoshalm & Rashid, 2016). Item analysis examines student's responses to individual test item question in order to assess the quality of those items and of the test as whole (Khoshalm & Rashid, 2016). Item analysis focuses to identify the item problematic. According to Varma (2014) as cited by Adegoke (2014), poorly written items; pictures, graphs and diagrams or lack of clear information may lead to absence of the correct response on the test item, item containing default distracters and bias for or against ethnic groups constitutes the reason for item problematic response that must be resolved. Difficulty and discrimination indices are among the parameters in item analysis that ensure standards of items in examinations (Pande, Pande & Parate, 2013). According to Aron (2006) as cited in Johari *et al*., (2011), difficulty index serves four purposes. Firstly, it identifies the concept that needs to be taught again, upon discovering that students cannot answer some particular questions. Secondly, identification and reporting the strengths and weaknesses of curriculum parts, which can and cannot be dominated by students. Thirdly, giving feedback to students regarding their strengths and weaknesses on topics assessed; and finally, identification of the questions that are content biased, like the contents that may have been highlighted during the teaching sessions. Thus, difficulty index is crucial for all educators, regardless of their level. On the other hand, discrimination index compares the number of people with high test scores who answered the item correctly with the number of people with low scores who answered the same item correctly. This index is considered as a basic indicator of an item quality (Cornachione, 2005). Difficulty and discrimination indices of the test item are affected by some factors. According to Olatunji (2009), Oyejide (1991) and Mehrens and Lehmann (1973) as cited in Ngung'u (2015), three factors affects the difficulty and discrimination indices of an item. Firstly, number of objectives indicating that the numbers of options provided in the test have either a positive or negative effects on both indices. Secondly, student level of understanding on a particular concept which might lead to inappropriate response. Thirdly, teachers training on the item development of which is necessary in enhancing well formulated questions. In support of these factors, Sung, Lin and Hung (2015) pointed out that, phonetic discrimination, number of plausible distracters, heterogeneity of sentence patterns in options, necessity for inference, lexical overlap, content familiarity, redundancy of necessary information have influence on difficulty and discrimination level of the test items.

Item analysis studies for a long time have been conducted worldwide to check for the reliabilities of the assessment tools in education institutions. In Malaysia, a retrospective study was done to reveal the competency assessment to medical undergraduate students who had undertaken the end of posting examinations after completing pediatric rotation. The study involved two cohorts and the results showed that the difficulty and discrimination level of multiple choice and long case questions were varying (Taib & Yusoff, 2014). Similarly, about 50% only of the test items for research in teaching beginners among music students in public universities in Malaysia were reported to have moderate difficulty and discrimination indices through the use of Kuder Richardson 20 and 21 (Sabri, 2013). Study done by Mukherjee and Lahiri (2015) reported on the acceptance of multiple choice questions to be used for further assessment after attaining the p-value of 20% to 90% and discrimination index of ≥ 0.3 in a medical college of Kolkata in Bengal, India. On the other hand, the test items analysis for an achievement test in the history subject to Indian standard 11$^{\text{th}}$, led to rejection of some of the items (Gowdhaman & Nachimuthu, 2013). Furthermore, the internal consistency and reliability of the networked minds as a measure of social pretence were studied in Nepal. The results obtained using cronbach alpha indicated that, the subscales factors were consistent (Harms & Biocca, 2004). According to Boopathiraj and Chellamani (2013) analysis of researcher made test items is of great importance as some of the items made by postgraduate students in Tamilnadu were found to be defective in both difficulty and discrimination levels. In Indonesia, the analysis of the difficulty level of the subjective English test was done during the mid semester, 2013 at SMA Negeri 1 Pendole where students answer sheets of the tenth grade were used during the data collection. The findings showed that, most of the test items were moderately easier and the test made by the teachers could be qualified in good test (Lebagi & Darmawan, 2014).

Study by Adegoke (2014) on the role of item analysis in detecting and improving faulty of physics objective test items involved 900 sample sizes among senior secondary school was conducted in ibadan, Nigeria. The results showed that some of the items were extremely easy and some failed to discriminate among higher and lower achievers. In South Africa, a quality assurance study for tools used in assessment was conducted. The study investigated the difficulty and discrimination ability of examinations in an undergraduate pharmacy programme at Medunsa campus of the University of Limpopo. The difficulty and discrimination indices were calculated for each True/False and constructed response questions from a total of 15 summative examinations in 1st and 4th year's level. The results found that most of the items had an acceptable level of difficulty and discrimination indices, though some were detected to have some weaknesses. They finally recommended more educators to carry out more items analysis for their test-writing and communicate their finding (Fourie, Summers & Zweygarth, 2010). In Kenya, the study on item analysis concentrated on the investigation of factors affecting reliability, difficulty and discrimination indices of science test items in commercial paper to class eight students (Ngung'u, 2015). In Tanzania, studies have reported on little capacity of secondary school teachers in assessment practices including computing both difficulty and discrimination indices for their constructed test items (Byabato & Kisamo, 2014). On the other hand, HakiElimu (2012) reported on the reliability of the examinations administered by the national examination council of Tanzania to secondary school students.

In the light of what has been pointed out in the consulted literature, it can be argued that, more on the item analysis in different examinations concentrated on intensively lower levels of education. Furthermore, the analysis practices in higher education has been observed more on formative assessment and little on summative assessment in south Africa. In Tanzania, through national examination council of Tanzania, there is evidence of little capacity by teachers on assessment practices in secondary schools and less is known in higher learning institutions. Thus, this study aimed at determining the reliability of the assessment tools in Tanzanian Universities.

6. Methodology

The target population comprised of first year undergraduate education students at Solomon Mahlangu College of Science and Education, Sokoine University of Agriculture (SUA). The college has five departments where the department of education was randomly selected. Also, the researchers selected one of the courses from the sampled department randomly where Introduction to Education Psychology (EDP 100) was selected. Retrospective record review was done on education undergraduate students who sat for an EDP 100 in 2014/2015, 2015/2016 and 2017/2018 academic year. 214 scripts were systematically randomly sampled from each academic year and hence a total of 642 scripts were obtained and include in the study.

6.1 Introduction to Educational Psychology university format and Assessment Components

Each of the university examination contained multiple choice (MCQ), matching items (MIQ) and essay type questions (ETQ) as assessment tools for summative evaluation developed from the department of education. Before end of the Semester University examination administration students were required to participate fully during the lectures and take all the required continuous assessment tests, seminars and assignments as part of formative assessment. The MCQ, MIQ and ETQ were constructed by the course instructors and later moderated by the department of education. MCQ and MIQ were found to range from 17 to 30 of which, the first 17 or 20 questions were included in the study. ETQ were found to be two or three of which they were all included in the study.

6.2 Statistical Analysis

Three statistical tests were done namely internal consistency reliability coefficient, difficulty index and discrimination index. The calculations were done on excel sheet.

Determination of Internal consistency reliability coefficient: This was measured by split-half method as proposed by (Boyle, 2017). According to Webb, Shavelson and Haertel (2006) split-half method is done by dividing the test items into half parts and a host of split-half reliability coefficients is derived. The correlation between two halves, which is odd score (X) and even score (Y), was estimated by Pearson product moment correlation formula shown in equation 1 (Mukaka, 2012).

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\,(\sum(Y-\bar{Y})^2}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Whereby: r= correlation coefficient of a half length test, X= odd score, Y= even score, $\bar{X}$= mean of X scores, $\bar{Y}$= mean of Y scores

Furthermore, the reliability of full length test was pressed using the spearman-Brown formula shown in equation 2 as adopted from (Webb *et al.*, 2006).

$$r^2 = \frac{2r}{1+r} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

Whereby: r^2= reliability coefficient of a full length test, r = correlation coefficient of a half length test.

Difficulty index (P): This is known as p-value or easy value describing the percentages of students who correctly answered the item, and that it ranges from 0 to 100% or 0 to 1 (Hingorjo & Jaleel, 2012). According to Bichi (2015), difficulty index is denoted as P and for objective questions is symbolically given as;

$$P = \frac{R}{N} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots . (3)$$

Whereby: P= difficulty index, R= number of examinees who get that item correctly, N= Total number of examinee who sat for a test. The extension of this formula was given by Boopathiraj and Chellamani (2013) who stated that;

$$P = \frac{Ru+Rl}{Nu+Nl} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots . (4)$$

Whereby: P = difficulty index, Ru= number of students in the upper group who responded correctly, Rl = number of students in the lower group who responded correctly, Nu = Numbers of students in the upper group, Nl = Number of students in the lower group.

On the other hand, the difficulty index for subjective questions is expressed according to Nitko (2004) of which a formula was given as;

$$P = \frac{Average\ Score}{Range\ of\ full\ mark} \dots\dots\dots\dots\dots\dots\dots (5)$$

Difficulty index determines the difficulty levels in examinations questions by classifying into easy, moderate and hard (Johari *et al*., 2011). Difficulty index further compares the difficulty in answering the same examination question by a group of students (Johari *et al*., 2011). Test items are classified as easy, moderate difficulty or difficulty if their difficulty indices are >70, 0.31≤0.7 or ≤0.3, respectively (Bichi, 2015)

Discrimination index (D): Refer to the ability of an item to distinguish high and low scoring learners (Koçdar, Karadag & Sahin, 2016). According to Fourie *et al*., (2010), discrimination index is denoted as D and is mathematically expressed as;

$$D = \frac{Ru-Rl}{Nu\ or\ Nl} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots . (6)$$

Whereby: D= discrimination index, Ru= number of students in the upper group who responded correctly, Rl = number of students in the lower group who responded correctly, Nu = Numbers of students in the upper group, Nl = Number of students in the lower group.

The procedure followed was adapted from Boopathiraj and Chellamani (2013) as follows;

i. The sample test papers were obtained from the administered university examinations of which, 214 examination papers from each cohort were drawn through systematic random sampling.
ii. The upper 27% and lower 27% examinees were obtained with the highest and lowest ranking order respectively on the total test of which, 58 examinees were obtained for each upper and lower groups from all three cohorts.
iii. Calculations for each item were done correctly using the relevant formula as shown above.

The index ranges from -1.00 to +1.00 and classified as satisfactory, acceptable, marginal or poor items if their discrimination indices are ≥0.40, 0.30 to ≤0.39, 0.2 to ≤0.29 or ≤0.2, respectively (Bichi, 2015; Hingorjo & Jaleel, 2012). This index reflects the degree to which an item and the test as a whole are measuring a unitary ability and thus, values of the coefficient will tend to be lower for tests measuring a wide range of content area than more homogenous tests (Quaigrain & Arhin, 2017). Furthermore, it is expected that the higher performing students selects the correct answer for each item more often than the lower performing students (Hingorjo & Jaleel, 2012). Positive discrimination index (0.00 to +1.00) and negative discrimination index (-1.00 to 0.00), entails that higher achievers got correct answer for specific item more than lower achievers and vice versa, respectively (Hingorjo & Jaleel, 2012).

7. Findings and Discussion

7.1 Internal Consistency Reliability of an Assessment Tool

Table 1 shows the correlation coefficient r, expressing a reliability of an introduction to educational psychology (EDP 100) university examinations administered to all first years students who are pursuing a bachelor degree of science with education at Solomon Mahlangu College of Science and Education, Sokoine University of Agriculture. The results showed that, internal consistency reliabilities of the three examinations ranged from 0.49 to 0.74. Typical classroom test displays an internal consistency reliability of between 0.60 and 0.80, implying that, on average of between 20% and 40% of the variations in the students' scores is a result of measurement of errors (Blerkom, 2009). Results in table 1 showed that, the correlation coefficients of the examinations administered during the 2014/2015 and 2015/2016 academic years had the internal consistency values that fall within the provided limits and hence are considered to have acceptable reliabilities. On the other hand, the reliability coefficient of the exams administered during the 2017/2018 academic year had low internal consistency compared to the proposed limits. Furthermore, the results shown in table 1 indicated the decrease in reliabilities across the three years. Sokoine University of Agriculture has experienced a substantial increase of students in almost all the offered programs including bachelor degrees of education since its establishment. The decrease in reliabilities as observed can be attributed with this increase in number of students. Associating these results with increase in the number of students is supported with the observation that class size has an effect on the test reliability (Alomari & Akour, 2014). Blerkom, (2009) specifies that lack of time for teachers and their incompetency during classroom test construction tends to lower the reliability of the test. Therefore, it can be established that the increase of the numbers of students in universities overloads instructors and hence run shortage of time for proper test construction.

Table 1: Internal Consistent Reliability of the EDP 100 across three years

Cohort	2014/2015	2015/2016	2017/2018
Correlation Coefficient, r	0.74	0.67	0.49

7.2 Difficulty and Discrimination Indices of the Test Items

7.2.1 Multiple Choices Items (MCQ)

Table 2 indicates the results difficulty and discrimination levels of the multiple choices questions (MCQ) of the three administered university examinations. 20 examination items were considered for analysis for all examinations. Basing on the classification made by Bichi (2015) and Hingorjo & Jaleel (2012) an item is considered as satisfactory, acceptable, marginal or poor items if its discrimination index is ≥0.40, 0.30 to ≤0.39, and 0.2 to ≤0.29 or ≤0.2, respectively. On the other hand, test items are considered to be easy, moderately difficulty or difficulty if their difficulty indices are >70, 0.31≤0.7 or ≤0.3, respectively (Bichi, 2015). Thus, from the table, MCQ 4 and 6 were easy, 1,2,3,7,8,9,11,12,13,14,16,17,18, and 20 were moderate difficulty while, 5,10,15 and 19 were considered as difficulty items in 2014/2015, MCQ 3,9 and 20 were easy items, 1,2,5,6,7,8,9,10,11,12,13,14,15,16,17,18 and 19 were moderately difficulty items and 4 was considered as difficulty items during the 2015/2016 academic years, while MCQ 1,4,5,7,8,9,10,11,12,13,14,15 and 16 were moderately difficulty items , MCQ 2,3 and 6 were difficulty items and there were no easy items during the 2017/2018 academic year. This indicated that, most of the administered items across the three years were moderately difficulty, though some of the items needed some improvements. The discrimination indices for the 2014/2015 cohort indicated that, MCQ 1,3 and 4 had satisfactory discrimination indices, MCQ 2,7,8,16,17 and 20 had an acceptable discrimination indices, MCQ 9,12 and 13 had marginal discrimination indices and the MCQ 4,5,6,10,11,14,15,18 and 19 had poor discrimination indices. For 2015/2016 cohort MCQ 1,2,6,10 and 16; MCQ 9,12 and 17; and MCQ 3,4,7,8,12,14,18,19 and 20 had acceptable, marginal and poor discrimination indices, respectively with the absence of satisfactory items. Furthermore, the discrimination indices for MCQ 9; MCQ 8; MCQ1,4,15 and 16; and MCQ 2,3,5,6,7,10,11,12,13,14,17, 18,19 and 20 for 2017/2018 cohort were satisfactory, acceptable, marginal and poor, respectively. The obtained results informed that 53.33% of the items administered across the three years had poor discrimination indices failing to discriminate higher and lower achievers. According to Mahrens and Lehman (1991) pointed out the reasons for discrimination indices being poor as firstly, the items being more difficulty or easy and hence lowering their discrimination power; and secondly, the purpose of the items in relation to the total test of which influence the magnitude of their discrimination power. These delineations demonstrate that the obtained results in this study indicated more moderately difficulty with either fewer easy or difficult items. Furthermore, negative discrimination indices were observed in MCQ 6, 15 and 19 of the 2014/2015 cohort and MCQ 17 of the 2017/2018 cohort. There are reasons for items to have negative discrimination indices. Quaigrain and Arhin (2017) argued that, wrong and ambiguous key in framing a question contributes to the negative discrimination power. Furthermore, Matlock-hetzel (1997) pointed out that items with negative discrimination indices are useless and that tends to lower the validity of the test. Therefore, so long as the items with negative discrimination indices are observed in the test, they should be examined to determine why a negative value was obtained (Quaigrain & Arhin, 2017).

Table 2: Difficulty and Discrimination Indices of Multiple Choice Questions

Cohort	2014/2015		2015/2016		2017/2018	
Item	**P**	**D**	**P**	**D**	**P**	**D**
1	0.68	0.43	0.55	0.39	0.68	0.29
2	0.67	0.34	0.58	0.39	0.18	0.02
3	0.50	0.40	0.72	0.15	0.22	0.14
4	0.81	0.10	0.26	0.07	0.41	0.24
5	0.04	0.02	0.67	0.07	0.40	0.20
6	0.82	-0.02	0.67	0.33	0.28	0.19
7	0.60	0.30	0.30	0.02	0.35	0.01
8	0.66	0.30	0.32	0.17	0.65	0.36
9	0.51	0.20	0.77	0.29	0.50	0.42
10	0.26	0.10	0.51	0.32	0.52	0.17
11	0.49	0.12	0.60	0.09	0.50	0.10
12	0.70	0.22	0.39	0.26	0.35	0.02
13	0.70	0.26	0.41	0.15	0.64	0.10
14	0.30	0.05	0.57	0.02	0.82	0.16
15	0.24	-0.10	0.44	0.30	0.48	0.24
16	0.66	0.31	0.60	0.32	0.52	0.21
17	0.60	0.36	0.34	0.28	0.28	-0.05
18	0.70	0.09	0.64	0.20	0.61	0.12
19	0.25	-0.10	0.68	0.20	0.50	0.10
20	0.60	0.40	0.90	0.20	0.50	0.10

7.2.2 Difficulty and Discrimination Indices of Matching Items

Table 3 showed the difficulty and discrimination indices of the matching items questions, (MIQ) for the selected three cohorts. During 2014/2015 cohort, 17 items were administered. 20 items were administered for both 2015/2016 and 2017/2018 cohorts. The results indicated that, MIQ 4; MIQ 1,2,3,5,6,7,8,9,10,11,12 and 16; and MIQ 13,14,15 and 17 were easy, moderately difficulty and difficulty items during the 2014/2015 cohort. MIQ 2,5,6,7,8,9,10,11,14,15,17,18,19 and 20 were moderately difficulty while MIQ 1,3,4,10,12,13 and 16 were difficulty items during the 2015/2016 cohort with no easy items. MIQ 12,17 and 18 were easy items, MIQ 1,2,3,4,5,6,8,9,10,11,13,14,15,16 and 20 were moderately difficulty items while MIQ 7 and 19 were difficulty items during the 2017/2018 cohort. On the other hand, discrimination indices for MIQ 1,2,3,4,8,11,12,14 and 16; MIQ 10; MIQ 5,6, and 9; and MIQ 7,13, 15 and 17 indicated that, the items were satisfactory, acceptable, marginal and poor, respectively during the 2014/2015 cohort. MIQ 7, 14 and 20; MIQ 5,9 and 17; MIQ 2,13, 18 and 19; and MIQ 1,3,4,6,8,10,11,12,15, and 16 had satisfactory, acceptable, marginal and poor discrimination indices, respectively during the 2015/2016 cohort and MIQ 3,9 and 20; MIQ 1,2,5,6, 8, 12, 15 and 16; MIQ 10,11 and 16; and MIQ 4,7,17,18 and 19 had satisfactory, acceptable, marginal and poor discrimination indices, respectively during the 2017/2018 cohort.

Table 3: Difficulty and Discrimination Indices of Matching Items

Cohort	2014/2015		2015/2016		2017/2018	
Items	P	D	P	D	P	D
1	0.63	0.43	0.10	0.04	0.69	0.31
2	0.36	0.41	0.32	0.22	0.60	0.30
3	0.60	0.52	0.22	0.15	0.50	0.50
4	0.48	0.45	0.17	0.19	0.62	0.10
5	0.85	0.26	0.48	0.30	0.34	0.36
6	0.54	0.22	0.34	0.13	0.34	0.34
7	0.70	0.20	0.50	0.40	0.00	0.00
8	0.44	0.64	0.50	0.11	0.38	0.34
9	0.35	0.22	0.50	0.30	0.56	0.40
10	0.62	0.38	0.24	0.04	0.46	0.29
11	0.53	0.55	0.30	0.07	0.57	0.24
12	0.47	0.46	0.10	0.00	0.79	0.34
13	0.20	0.19	0.28	0.22	0.34	0.05
14	0.27	0.43	0.47	0.46	0.50	0.21
15	0.13	0.19	0.52	0.11	0.40	0.30
16	0.47	0.52	0.27	0.17	0.70	0.30
17	0.11	0.10	0.39	0.33	0.90	0.10
18	----	---	0.37	0.26	0.90	0.17
19	----	---	0.64	0.28	0.00	0.00
20	----	---	0.40	0.50	0.40	0.40

7.2.3 Difficulty and Discrimination Indices for Essay items

Table 4 showed the difficulty and discrimination indices for the essay type questions (ETQ) for the three cohorts. The results indicated that both ETQ 1and 2 2014/2015 cohort has moderate difficulty indices, ETQ 1 was easy while ETQ 2 and 3 were moderately difficulty in the 2015/2016 cohort. ETQ 1 and 2 during the 2017/2018 cohort were moderately difficulty and easy items, respectively. Both ETQ in 2014/2015 had acceptable discrimination indices; ETQ 1 had marginal discrimination power, while ETQ 2 and 3 were considered to have poor discrimination power during the 2015/2016 cohort. On the other hand, ETQ 1had high and satisfactory discrimination power while ETQ 2 was observed to have poor discrimination index.

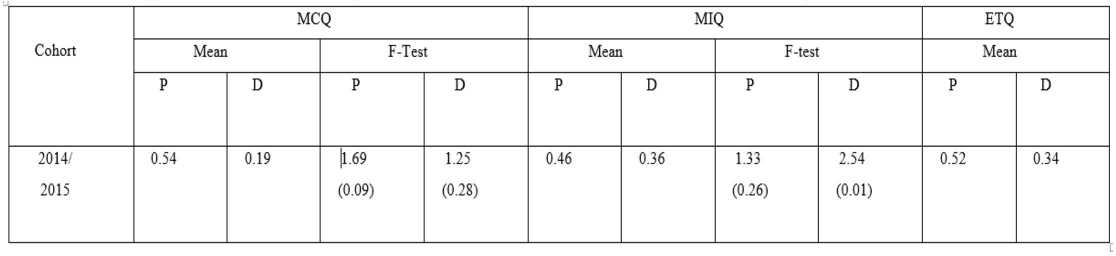
Table 4: Difficulty and Discrimination Indices of Essay items

Cohort	2014/2015		2015/2016		2017/2018	
Item	P	D	P	D	P	D
1	0.65	0.31	0.77	0.28	0.50	0.85
2	0.38	0.36	0.31	0.02	0.79	0.14
3	---	---	0.37	0.04	---	---

7.3 Comparison of Difficulty and Discrimination Indices across three years

Table 5 compared the means of difficulty and discrimination indices for the university examinations administered to the three cohorts. Though it appears that the difficulty indices were moderately difficulty across the three years, the discrimination indices for MCQ were on average poor, marginal poor and satisfactory items during 2014/2015, 2015/2016 and 2017/2018 academic years, respectively. MIQ were on average acceptable, marginal poor and poor items during 2014/2015, 2015/2016 and 2017/2018 academic years, respectively. On the other hand, the ETQ were on average accepted, marginal poor and satisfactory items during 2014/2015, 2015/2016 and 2017/2018 academic years, respectively. The variation in difficulty and discrimination indices for all examinations administered to the three cohorts was statistically not significant for the two types of questions ($p>0.05$ for MCQ and MIQ) with exception to the discrimination index for MIQ which shows significant variation across the three years ($p<0.05$) as shown in Table 5.

Table 5: Comparison of Difficulty and Discrimination Indices across three years



Cohort	MCQ				MIQ				ETQ	
	Mean		F-Test		Mean		F-test		Mean	
	P	D	P	D	P	D	P	D	P	D
2014/ 2015	0.54	0.19	1.69 (0.09)	1.25 (0.28)	0.46	0.36	1.33 (0.26)	2.54 (0.01)	0.52	0.34

Figures in the brackets represent calculated p-value

8. Conclusion

This study revealed that majority of the questions for the EDP 100 though were moderately difficulty, their discrimination powers were poor. However, the variation in difficulty and discrimination indices for the three cohorts was statistically not significant with exception to the discrimination index for MIQ which vary significantly across years. Also, there is a drop in internal consistency across the three cohorts. This could be partly associated with the increase in numbers of students in universities leading to increased instructors' workload that may limit instructor's time for concentrating on test construction effectively. Therefore, to tackle these challenges instructors need to be conversant with the knowledge of item analysis and apply it frequently during formative assessment. Such analysis will enhance identification of strong and weak test items as early as possible before they are included in the summative assessments tools. It is also recommended that similar studies should be done to determine both validity and reliability of the assessment tools for the other subject at the University.

References

Adegoke, B. A. (2014). The role of item analysis in detecting and improving faulty physics objective test items. *Journal of Education and Practice*, **5**: 110–121.

Ajayi, B. K. (2013). A comparative analysis of test re-test and equivalent reliability methods. *International Journal of Education and Research*, **1**: 209–216.

Alomari, H. & Akour, M. (2014). The effect of class size on reliability estimates of college-students course grades. *Journal of Educational & Psychological Sciences*, **15**: 541-556.

Bichi, A. A. (2015). Item Analysis using a derived science achievement test data. *International Journal of Science and Research*, **4**: 1656-1662.

Blerkom, M. L. V. (2009). *Measurement and Statistics for Teachers*. New York, NY, USA: Routledge.

Boopathiraj, C. & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisclipinary Research*, **2**: 189–193.

Boyle, G. J. (2017). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales ? Does item homogeneity indicate internal consistency or item redundancy in psychometric scales ?. *Retrieved from http://www. researchgate.net*. On 03/03/2019.

Byabato, S. & Kisamo, K. (2014). Implementation of school based continuous assessment (ca) in Tanzania ordinary secondary schools and its implications on the quality of education. *Developing Countries Studies*, **4**: 55–62.

Chan, C. (2010). Assessment: assessing large class, assessment resouces @HKU, University of Hong Kong. Retrieved from *Http://Ar.Cetl.Hkuhk/Large-Class.Htm.* Accessed on March, 03, 2019.

Cornachione Junior, E. B. (2005). Objective tests and their discriminating power in business courses: a case study. BAR. *Brazilian Administration Review*, **2**: 63–78.

Ercan, I., Yazici, B. & Sigirli, D. (2007). Review of reliability and factors affecting the reliability. *InterStat*, 2007: 1-22.

Fourie, S., Summers, B. & Zweygarth, M. (2010). Difficulty and discrimination indices as quality assurance tools for assessments in a South African problem-based pharmacy programme. *Pharmacy Education*, **10**: 119–128.

Ghazali, N. (2016). A reliability and validity of an instrument to evaluate the school-based assessment system : A Pilot Study. *International Journal of Evaluation and Research in Education*, **5**: 148–157.

Gowdhaman, K. & Nachimuthu, K. (2013). Item analysis of history achievement test on difference index (DI) in the criterion referenced measurement. *Research Journal of Educational Science*, 1: 1–8.

Hajjar, S. T. EL. (2018). Statistical Analysis: Internal-Consistency Reliability And Construct Validity . *International Journal of Quantitative and Qualitative Research Methods*, 6: 27–38.

HakiElimu. (2012). *School Children and National Examinations : Who Fails Who ? A research Report on the*

Relationship between Examination Practice and Curriculum Objectives in Tanzania. Dar es Salaam. Hakielimu.

Harms, C. & Biocca, F. (2004). Internal consistency and reliability of the networked minds measure of social presence measure. In Seventh Annual International Workshop: Presence 2004, valencia: *Universidad Politecnica de valencia*.

Hingorjo, M. R. & Jaleel, F. (2012). Original article analysis of one-best MCQs : the difficulty index , discrimination index and distractor efficiency. *Journal of Pakistan Medical Association*, **62**: 142–147.

Hornsby, D. & Osman, R. (2014). Massification in higher education : large classes and student learning. *High Education*, 2014: 1-9.

Johari, J., Sahari, J., Abd Wahab, D., Abdullah, S., Abdullah, S., Omar, M. Z. & Muhamad, N. (2011). Difficulty index of examinations and their relation to the achievement of programme outcomes. *Procedia - Social and Behavioral Sciences*, **18**: 71–80.

Kapinga, O and Amani, J. (2016). Determinants of students ' academic performance in higher learning institutions in Tanzania. *Journal of Education and Human Development*, **5**: 78–86.

Khoshalm, H. B and Rashid, S. (2016). Assessment of the assessment tool : analysis of items in a non-MCQ. *International Journal of Instruction*, **9**: 119-132.

Kinyua, K. & Okunya, L. O. (2014). Validity and reliability of teacher-made tests : Case study of year 11 physics in Nyahururu District of Kenya. *African Educational Research Journal*, **2**: 61–71.

Koçdar, S., Karadag, N. & Sahin, M. . (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context. *The Turkish Online Journal of Educational Technology*, **15**: 16–24.

Lebagi, D. & Darmawan, N. (2014). Analyzing difficulty level of subjective test. *E-Journal of English Language Teaching Society*, **2**: 1–14.

Matlock-hetzel, S. (1997). *Basic concepts in item and test analysis*. A paper Presented at the Annual Meeting of the Southwest Educational Research association, Austin, 23-25, January, 1997.

Mehrens, W.A. & Lehmann, I.J. (1991). *Measurement and Evaluation in Educational Psychology*. 2nd Edition. New York, NY. Houghton Mifflin Company.

Miller, M.D., Linn, R.L. & Gronlund, N. E. (2009). *Measurement and Assessment in Teaching.* 10th Edition. New Jersey, USA: Pearson Educational. Inc.

Mohamedbhai, G. (2008). Massification in higher education institutions in africa : causes , consequences, and responses. *International Journal of African Higher Education*, **1**: 61-83.

Mukaka, M. M. (2012). Statistics corner : a guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, **24**: 69–71.

Mukherjee, P. & Lahiri, S. K. (2015). Analysis of multiple choice questions (MCQs): item and test statistics from an assessment in a medical college of Kolkata, West Bengal. *Journal of Dental and Medical Sciences*, **14**: 2279–2861.

Ngung'u B. C. (2015). *Reliability, Item Difficulty And Discrimination Indices Of Science Test Items In Commercial Test Papers And Their Correlation To Students Kcpe Performance In Science In Limuru Subcount*y. Master's Thesis. University of Nairobi.

Nitko, A.J. (2004). *Educational Assessment of Students*. 4th Edition. Upper saddle River, NJ. Pearson/Merill. Prentice Hall.

Ntim, S. (2016). Massification in ghanaian higher education : implications for pedagogical massification in ghanaian higher education : implications for pedagogical quality , equity control and assessment. *International Research in Higher Education*, **1**: 160-169.

Oliveira, K., Santos, B., Martins, F., Ii, C., Maria, T. & Iii, D. A. (2016). Internal consistency of the self-reporting questionnaire-20 in occupational groups. *Revista de Saude Publica*, **1**: 1–10.

Omari, I.M. (2006). *Educational Psychology for Teachers*. Dar es Salaam. Dar es Salaam University Press.

Pande, S. S., Pande, S. R. & Parate, V. R. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in physiology. *South-East Asian Journal of Medical Education*, **7**: 45–50.

Quaigrain, K. & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, **4**: 1–11.

Rahman, S. & Majumder, A. A. (2014). Is it assessment of learning or assessment for learning ?. *South East Asia Journal of Public Health*, **4**: 72-74.

Sabri, S. (2013). Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. *International Journal of Education and Research*, **1**: 1–14.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research and Evaluation*, **22**: 1-11.

Sokoine University of Agriculture (SUA), (2007). *The sokoine University of Agriculture Charter*.

Morogoro,Tanzania.

Sung, P., Lin, S. & Hung, P. (2015). Factors affecting item difficulty in english listening comprehension tests. *Universal Journal of educational Research*, **3**: 451–459.

Taib, F. & Yusoff, M. S. B. (2014). Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. *Journal of Taibah University Medical Sciences*, **9**: 110–114.

Tang, W., Cui, Y. and Babenko, O. (2014). Internal consistency: do we really know what it is and how to assess it? *Journal of Psychology and Behaviour Science*, **2**: 205–220.

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach ' s alpha. *International Journal of Medical Education*, **2011**: 53–55.

Tremblay, K., Lalancette, D. & Roseveare, D. (2012). *Assessment of higher education learning outcomes*. AHELO Feasibility Study Report. OECD, Volume 1.

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales ? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, **20**: 679–686.

Webb, N. M., Shavelson, R. J. & Haertel, E. H. (2006). Introduction. *Handbook of Statistics*, **26**: 1–42.

Yelkpieri, D., Namale, M., Esia-donkoh, K. & Ofosu-dwamena, E. (2012). Effects of large class size on effective teaching and learning at the Winneba Campus of the UEW (University of Education, Winneba), Ghana. *US-China Educational Review,* **3**: 319–332.

Zephaniah, A., Zhao, M. & Feng, J. (2016). Significance of trends on enrolment , budget and actual expenditure in the examination of higher education financing. *Journalof Education and Practice*, **7**: 129–141.

Zhu, J. & Han, L. (2011). Analysis on the main factors affecting the reliability of test papers. *Journal of Language Teaching and Reserach*, **2**: 236–238.