Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 2, 2474 – 2486

# The Effectiveness of Test Justification Tables in Generating High-quality Assessments

**Faiz Zulkifli[1], Rozaimah Zainal Abidin[2], Zulkifley Mohamed[3]**

*[1,2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia*
*[3]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia*

***Corresponding Author:***
*Rozaimah binti Zainal Abidin*
*Faculty of Computer and Mathematical Sciences*
*Universiti Teknologi MARA, Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia*
*Email: rozai256@uitm.edu.my*

## ABSTRACT

The most significant phase in formulating assessment questions is the creation of the Test Specification Table (TST). A reliable TST can generate high-quality assessments that test students' comprehension of a topic. Most universities have begun requiring TST preparation for each course prior to the start of a teaching and learning session. Its goal is to assure consistency in the assessment questions provided by each lecturer, even if they are completed by various people. Despite the importance of TST preparation, no research has been conducted to determine the extent to which TST is effective in achieving learning objectives. As a result, descriptive statistical analysis will be used in this study to evaluate student achievement on an assessment created in accordance with revised Bloom's taxonomy. The data for statistics and probability courses offered to students at Malaysia's largest public universities is the focus of this case study. The study's findings revealed that TST is helpful in assisting in the development of high-quality assessment questions for assessing a student's progress. This research also offers some changes to the TST production process so that it can become a more comprehensive guide for future question drafters.

## 1. INTRODUCTION

Malaysia's assessment and evaluation system has experienced a metamorphosis, with components of high order thinking skills (HOTS) being included in the filling out assessment and exam questions (Abdullah et al., 2015). Every student will be exposed to various questions of varying degrees of difficulty, which will be graded based on the student's level of thinking. Before being delivered to students, university exam questions must go through several steps (Raus et a., 2014). The process begins by preparing a test specification table (TST) outlining the number of question items, question categories, total marks, the time necessary, and Bloom's taxonomy level.

TST for continuous assessment (also known as TSTC) and TST for final assessment are the two forms of TST that must be developed for a course. A well-designed TST can create high-quality assessments, which can assess students' understanding of a course (Zulkifli et al., 2018). To achieve learning objectives, it can also identify different difficulty levels of questions according to each course topic. The experienced lecturer will verify the content and format of exam questions or a final assessment if it follows the TST. Meanwhile, TSTC should guide the ongoing assessments offered by teaching lecturers, such as assignments, laboratories, quizzes, and tests, but there is no formal review procedure in place or no validation by experienced lecturers.

Although there are numerous benefits to preparing this TST, no research has looked into the extent to which TST accomplishes learning objectives by creating high-quality, comprehensive assessment questions. Students'

abilities at various levels of knowledge, from simple to HOTS levels, should be tested in assessments. Using a revised Bloom's taxonomy in evaluating the level of difficulty for a topic is critical (Grundspenkis, 2019). However, most instructors cannot differentiate between lower and HOTS when using taxonomies (Abdullah et al., 2017). According to Arievitch (2020), the taxonomy contains some conceptual flaws since it is incompatible with the concept of human mind function, teaching methods, and student learning. Philosophers regard it as ancient, and it is a mechanical model of human cognition known as "information processing".

Therefore, the purpose of this study is to see how effective TST is at producing high-quality question items by measuring student achievement in the courses chosen as study subjects. Descriptive analysis of the overall score and each assessment item mark can provide preliminary information on student achievement by topic. It can also classify the difficulty of a question using revised Bloom's taxonomy classification described in the syllabus. Finally, this research will recommend changes to the TST production technique to make it a more user-friendly and systematic guide for question drafters.

## 2. THEORETICAL BASIS AND TEST SPECIFICATION TABLE (TST)

The theory that underpins the assessment questions and the TST details used by the study subjects will be discussed in this section.

### 2.1. Revised Bloom's Taxonomy

A revised Bloom's taxonomy is required to improve the teaching, learning, and assessment of mathematics courses (Radmehr & Drake, 2018). Figure 1 shows the level differences between the original and revised Bloom's taxonomy (Sagala & Andriani, 2019).


Figure 1. Bloom's taxonomy, original vs revised

The changes between the two versions, as seen in Figure 1, begin early in the process. The 'remembering' aspect has taken the role of the former version's 'knowledge'. Students are encouraged to retain topics and learn them, which helps strengthen cognitive processes. Furthermore, the original version's 'synthesis' aspect has been incorporated into the new version's 'analysing'. In the new version, the parts of 'evaluating' and 'creating' have been relegated to levels five and six. The new version, which 'analysing', 'evaluating', and 'creating', has improved the quality of HOTS. The most challenging part of implementing Bloom's taxonomy is interpreting the level of the taxonomy in the context of cognitive processes that necessitate a comprehensive set of questions covering the whole course topic.

The revised Bloom theory has a two-dimensional structure, one of its advantages (Radmehr & Drake, 2018). These two aspects have independent cognitive processes and information that can be used alone or in combination. In addition, the theory incorporates metacognitive knowledge while rejecting formal hierarchies. The study's findings are likely to aid people involved in mathematics education in enhancing the quality of teaching, learning, and assessment in a subject where Bloom's theory is revised less frequently than in other fields.

However, there have been studies that look back at the impact of Bloom's taxonomy on educational goals. Based on Piotr Galperin's study, Arievitch (2020) has investigated general themes in current psychology and education regarding teaching and learning developmental framework (TLDF). Bloom's taxonomy includes a range of conceptual issues, according to the TLDF approach. The issue is linked to a pervasive misperception of how the human mind works, how pupils learn, and how teachers are intended to teach. Piotr Galperin developed some hypotheses about developing an intellectual activity that might be used in education.

### 2.2. TST and TSTC

The ability of students to answer questions and the quality of a question item are the two factors that determine their performance on an assessment. The substance of a question item and the range of difficulty levels are used to evaluate its quality. The questions must be written by the guidelines that have been established in advance, notably TST and TSTC. Both tables are prepared by a resource person and then presented as a matrix that shows the number of questions, types of questions, total marks, time distribution, and taxonomy level of the topic to be examined.

As a case study, this research looked at courses in the discipline of statistics, namely statistics and probability. The mapping of each topic to the learning objectives is shown in Tables 1 and 2.

Table 1. TST

| **Topics** | | Continuous Random Variables | The Normal Distribution | Special Continuous Distribution | Multivariate Distribution | Distribution of Functions of Random Variables | **Total** |
|---|---|---|---|---|---|---|---|
| **Learning Outcomes (LO)** | | LO1, LO2 | LO1, LO2, LO3 | LO1, LO2, LO3 | LO1, LO2, LO3 | LO1, LO2, LO3 | |
| **Time (Hour)** | | 12 | 7.5 | 6.5 | 12 | 14 | **52** |
| **Time (%)** | | 23% | 14% | 13% | 23% | 27% | **100%** |
| **Exam (%)** | | 23% | 14% | 13% | 23% | 27% | **100%** |
| **Remembering** | **Question** | 1c | 7a | 6a | 2b | 2a | **16** |
| | **Marks** | 4 | 3 | 3 | 3 | 3 | |
| **Understanding** | **Question** | 1b, 9 | 7c | 4 | 2c | 8b | **18** |
| | **Marks** | 6 | 3 | 3 | 3 | 3 | |
| **Applying** | **Question** | 1a | 3b | 5b | 5a | 10b | **23** |
| | **Marks** | 5 | 4 | 2 | 5 | 7 | |
| **Analysing** | **Question** | 3a | 7b | 6b | 8a | 5c | **23** |
| | **Marks** | 8 | 4 | 2 | 5 | 4 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Evaluating** | **Question** | | | 10a | 8c | 10c | **20** |
| | **Marks** | | | 3 | 7 | 10 | |
| **Total Questions** | | **4** | **4** | **6** | **5** | **5** | **10 main questions** |
| **Total Marks** | | **23** | **14** | **13** | **23** | **27** | **100** |

Note: LO1- Explain the fundamental concepts of probability and distribution, LO2- Ability to explain the different types of random variables and their distributions, LO3- Apply the fundamental concepts of probability and probability distribution in various fields.

Table 2. TSTC

| **Assessments** | Quiz 1 | Quiz 2 | Test 1 | Test 2 | Assignment 1 | Assignment 2 | **Total** |
|---|---|---|---|---|---|---|---|
| **Learning Outcomes (LO)** | LO1, LO2, LO3 | LO1, LO2, LO3 | LO1, LO2, LO3 | LO1, LO2, LO3 | LO1, LO2, LO3 | LO1, LO2, LO3 | |
| **Time (Hour)** | 0.5 | 0.5 | 1.5 | 1.5 | 1 | 1 | **6** |
| **Time (%)** | 8% | 8% | 25% | 25% | 17% | 17% | **100 %** |
| **Marks (%)** | 8% | 8% | 25% | 25% | 17% | 17% | **100 %** |
| **Remembering** | 1% | 1% | 3% | 3% | 3% | 4% | **15%** |
| **Understanding** | 1% | 1% | 3% | 3% | 3% | 4% | **15%** |
| **Applying** | 1% | 1% | 8% | 8% | 8% | 4% | **30%** |
| **Analysing** | 4% | 3% | 8% | 8% | 3% | 4% | **30%** |
| **Evaluating** | 1% | 2% | 3% | 3% | | 1% | **10%** |
| **Total** | 8% | 8% | 25% | 25% | 17% | 17% | **100 %** |

The chosen course has five primary topics, each corresponding to one or more of the three learning objectives. Quizzes, tests, assignments and final exam are four assessments that students must complete during the teaching and learning sessions. Students will be exposed to a broader range of structured problems that are more cognitive. The difficulty levels of the questions should be divided into three categories: low, moderate, and high, with percentages of 20-35%, 40-60%, and 20-35%, respectively. Revised Bloom's taxonomy level, which includes six aspects, will lead to the cognitive level. Low-level aspects of 'remembering' and 'understanding' are classified. 'Applying' and 'analysing' aspects are rated as moderate. The challenging level focuses on 'evaluating' and 'creating'. In

comparison to TSTC, only TST reveals the number of questions and mark distribution for each cognitive level in detail for the final exam.

## 3. METHOD

This research focuses on statistics courses at Malaysia's most prominent public universities, which have fully implemented scoring based on cognitive questions and course learning outcomes. Furthermore, students chosen as respondents must be able to comprehend the types of questions that will be asked during the assessment, which will range in difficulty. Students who fail to complete the course or are unable to take the assessment for any reason will be removed from the study sample. Statistics and probability was the statistical course chosen for this research. Every student majoring in science and technology must take this course as a pre-requisite.

The data for the study was gathered using cluster sampling. The iCGPA (integrated Cumulative Grade Point Average) score reporting system, which details the marks for each question item and the difficulty level of the question, was used to collect data for the final exam. Each set of final exam questions must adhere to the TST and be content-checked by an expert lecturer. Throughout the semester, marks for continuous assessment will be gathered from lecturers who deliver the course. The TSTC directed lecturers in developing continuous assessment questions for quizzes, tests, and assignments, but experienced lecturers undertook no corrections.

A quantitative research approach is presented in determining student performance and the quality of questions for each assessment. Statistical descriptive methods will summarise information on the central location, dispersion, and shape of each assessment and question item. The central location describes the location of the majority of scores for a variable. The method can also determine whether the question items are appropriate for the cognitive levels of the TST. The dispersion measurement determines how close a score is to the mean in terms of dispersing. The shape measurement statistics can determine the normality of the study data.

A total of 326 students enrolled in statistics and probability courses were included in the study. Male students accounted for 38.7% of the overall student population, while female students accounted for 61.3%. Before proceeding with the analysis, the raw data must be processed and cleansed. Incomplete data should be disregarded and will not be included in the sample for the study. Meanwhile, each assessment item's score will be standardised to a maximum of five points using the following formula:

$$\frac{score\ value - minimum\ mark}{maximum\ mark} \times 5$$

As shown in Table 3, students' overall performance on each assessment will be graded. Each assessment should receive a 100-point total. As a result, conversion of the measurement scale from ratio to ordinal is required for quizzes, tests, and assignments. The following are the steps to coding the size:-

First: The mark obtained is converted into 100 counts using the formula,

$$\frac{mark\ obtained}{full\ mark} \times 100$$

Second, the score count of 100 will be recoded using the following indicators:

Table 3. Assessment grade re-coding

| Mark | 90-100 | 80-89 | 75-79 | 70-74 | 65-69 | 60-64 | 55-59 | 50-54 | 47-49 | 44-46 | 40-43 | 30-39 | 0-29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | A+ | A | A- | B+ | B | B- | C+ | C | C- | D+ | D | E | F |
| Code | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |

All data in grade form will be translated to numerical form to make it easier to customise the software, as seen in row three.

## 4. RESULTS AND DISCUSSION

The descriptive analysis of each assessment and each of its items will be discussed in this session. The central location, dispersion, and shape statistics were the three types of measurement used.

### 4.1. Results of descriptive statistics

In order to pass the statistics and probability course, students must pass seven separate assessments. Table 4 shows the statistics of central tendency and dispersion measurements for each assessment.

Table 4. The measure of central location and dispersion for each assessment

| Assessment | Central Location Statistics | | | Dispersion Statistics | | | |
|---|---|---|---|---|---|---|---|
| | Mean | Median | Mode | Standard Deviation | Minimum | Maximum | Range |
| Quiz 1 | 6.23 | 6 | *1 | 4.20 | 1 | 12 | 11 |
| Quiz 2 | 5.64 | 6 | 1 | 3.65 | 0 | 12 | 12 |
| Test 1 | 8.40 | 9 | 12 | 3.25 | 0 | 12 | 12 |
| Test 2 | 8.37 | 9 | 12 | 2.90 | 2 | 12 | 10 |
| Assignment 1 | 10.40 | 10 | 9 | 1.158 | 9 | 12 | 3 |
| Assignment 2 | 8.92 | 9 | 6 | 2.416 | 5 | 12 | 7 |
| Final exam | 9.19 | 9 | 8 | 1.91 | 2 | 12 | 10 |

*The least mode value for which a multiple modes exist is specified.

The median score for each test, assignment 2, and the final exam were nine, as shown in Table 4. This means that half of the student's mark will be below B+ and the other half will be above it. At six (grade C+), the median value for both quizzes was somewhat lower. In contrast, assignment 1 had a relatively high median value of ten. The frequency distribution curve for quizzes, assignments, and final exams was skewed to the right because the median value surpassed the mode value. Meanwhile, the frequency distribution curve for both tests is skewed to the left.

Grade E had the highest frequency or mode of quizzes, with 20.6 % and 19.3% for quiz 1 and quiz 2, respectively. Quiz 1 contains two modes, the second of which is grade A+. The quiz's low median value is attributable to students' lack of preparation for the assessment. However, students' test-taking performance has improved significantly, with most students receiving an A+ grade. There was also a rise in the final exam, with most students receiving a B grade (20.9%). Most students received a B+ for assignment 1 and a C+ for assignment 2. The assignments are guided questions that must be completed and answered outside of class time within the time frame specified.

Meanwhile, the quiz, test, and final exam assessments all had values ranging from ten to twelve in dispersion statistics. In contrast, assignments had a range of three or seven. Smaller range values indicate lower score dispersion and more consistent data. Furthermore, the assignment's minimal grade is a C or B+. Students who attended statistics and probability classes arguably performed exceptionally well academically.

Next, a discussion on a central location and dispersion measurement statistics will be conducted on each question item. In addition, shape measurement statistics will be calculated to determine the normality of the study data. Cohen et al. (2018) have suggested that skewness and kurtosis values should be in the value-added range less than twice its standard error. Information related to measurement statistics is included in the Appendix section.

The quiz question items were created to fulfil the three lowest levels of difficulty: remembering, understanding, and applying. The measurement statistics for quiz 1 item revealed that item 'a' had the lowest mean value and item 'c' had the highest. This is consistent with item 'a' (applying) having a higher difficulty level than item 'c' (remembering). Similarly, items 'b' and 'e' had a lower mean score for understanding difficulty than item 'd' (remembering). When central location statistics values are compared, the frequency distribution curve for items 'c' and 'd' is skewed to the left because the mean, median, or both values exceed the mode values. Whereas items 'a', 'b', 'e' and 'f' have a frequency distribution curve skewed to the right.

The dispersion statistics for all items quiz 1 revealed a five-point range. Items 'b' have the lowest standard deviation values, followed by items 'd', 'e', 'a', 'f', and 'c' in ascending order. The score has a narrow dispersion when the standard deviation is low. The skewness statistic for items quiz 1 should be in the range of -0.135 to 0.135. While the accepted range for kurtosis is -0.269 to 0.269. Nevertheless, the kurtosis value for quiz 1 was outside of the acceptable range. In addition, only item 'e' falls inside the skewness range. This signifies that the data normality standards for item quiz 1 are not met.

Items for quiz 2 had the greatest mean value for questions with the difficulty level of remembering, especially items 'c' and 'a'. Then level of understanding, such as items 'e' and 'f', followed by applying level questions, such as items 'd' and 'b'. The frequency distribution curves for items' b', 'c', 'd', and 'e' are skewed to the left. Meanwhile, the frequency distribution curve for items 'a' and 'f' is skewed to the right. All items have a range of five, with a standard deviation of 1.3 to 1.6. Furthermore, only item 'a' had a kurtosis value within the acceptable range. However, only items 'e' and 'f' fall inside the skewness range.

The test consisted of 12 items that assessed remembering, understanding, applying, and analysing difficulties. At the remembering level, the highest and lowest mean values for items test 1 were 3.79 and 3.64, respectively. Meanwhile, the highest and lowest mean values for each item at the understanding level were 3.54 and 3.26, respectively. There was no crossover between the degrees of remembering and understanding in the mean range. This implies a clear distinction between items of both difficulty levels. Furthermore, the greatest and lowest mean values for item test 1 at the applying level, respectively, were 3.39 and 3.04. The mean value of item 'f' at the applying level is higher than the mean value of item 'j' at the understanding level. Finally, the analysing level item's mean value was lower than the applying level item's.

When the mean, median, and mode values were compared, it was discovered that four of items test 1, namely 'c', 'd', 'e', and 'i' had a frequency distribution curve skewed to the left. Whereas, items 'a', 'b', 'f', 'g', 'h', 'j', 'k' and 'l' have a frequency distribution curve that is skewed to the right. Dispersion statistic reveals that all items were recorded in a four-point range. A standard deviation result less than 1.3 supports a modest range value compared to the quiz assessment. According to skewness statistics, four items, namely items 'a', 'b', 'h', and 'j', meet the condition of normality. None of the items, however, met the criteria for kurtosis.

Based on TSTC, test 2 had isolated items 'a', 'c' and 'i' as remembering level items and items 'b', 'e', 'f', 'g' and 'k' as understanding level items. Items 'd' and 'l' applying-level items, whereas objects 'h' and 'j' were analysing-level items. The mean value corroborates the isolation for each item, which has a greater mean value at the lower level. However, at a lower level than it, item 'e' had a slightly greater mean value than item 'i'. The frequency distribution curve for items 'a', 'b', 'c', 'd', 'e', 'g', 'i', 'j', and 'l' had skewed to the left. Whereas items 'f', 'h' and 'k' had a frequency distribution curve skewed to the right. The standard deviation value for item test 2 is less than 1.2. Items 'd', 'j', 'k', and 'l' have also met the skewness range based on shape statistics. Items 'a', 'g', 'h', 'j' and 'l' in turn met the kurtosis range.

The final exam consisted of 24 question items separated into five difficulty levels:

remembering, understanding, applying, analysing, and evaluating. According to the mean statistics, the lower the statistic's value, the more complex the question item is. However, three items, '3b', '7c', and '8a', are slightly out of the mean range for items of similar complexity. The frequency distribution curve for items '1a', '1c', '2a', '2b', '2c', '4', '5a', '5b', '6a', '7a', '7c', '8b', '8c', '9', and '10a' is skewed to the left, according to the central location statistics. Furthermore, the frequency distribution curves for items '3a' and '10c' are symmetrical since the mean, median, and mode values are nearly identical.

The majority of the final exam items have a range of three, except for item '1b', which has a range of three. The low range also results in a low standard deviation. On the other hand, large standard deviations generate high range values when no outlier data is available. According to the skewness statistic, there are 10 items that match the normality conditions: '1b', '3a', '3b', '5b', '5c', '6b', '7b', '10a', '10b', and '10c'. However, only the items '2a' and '5a' were inside the kurtosis range.

### 4.2. Discussion

Continuous assessment scores have a positive impact on final exam results. Students will do well in final exams if they perform well on assignments, quizzes, and tests. However, only hypothesis testing can determine the significance of such beneficial impacts. Several previous research has demonstrated that tests (Sikder et al., 2016), quizzes (Joyce et al., 2015), and assignments (Meier et al., 2016) have all been found to contribute to students' final assessment performance.

Furthermore, the results of this study show that the difficulty of a question can be determined by students' scores on an assessment item. Easy questions will have a higher mean score than the harder ones. In contrast, the mean score for the most challenging questions was the lowest. The technique only gives a general idea of a question's difficulty level based on measurements taken in a central location. This study's findings align with (Zainudin et al., 2012), who discovered that an item is classified as very difficult if the index value generated using the mean score is the lowest.

This study also discovered that the primary assessments of the statistics and probability courses, such as quizzes, tests, assignments, and final exam, satisfied the revised Bloom's taxonomy assumptions. Numerous forms of assessments can be used to divide the difficulty level into five categories: remembering, understanding, applying, analysing, and evaluating. Easy, medium and HOTS questions are described at each of these levels.

The ability of students to answer questions and the quality of a question item determine their performance on an assessment. The substance of a question item and the range of difficulty levels are used to determine its quality. Both tables are laid out as a matrix, with the number of questions, total marks, and taxonomy level of the topic to be assessed listed. The questions to be asked must adhere to the guidelines that have been established ahead of time, particularly TST and TSTC. However, there is still potential for improvement to reinforce the schedule more. The following are some ideas:

a. The details of the subtopics should be done as the main topic at TST and subsequently the distribution of question number, total marks, and taxonomy level of the subtopic.
b. Matters on TST should be extended to TSTC so that the questions produced to meet the same standards as the final exam.
c. In ensuring that no identical or nearly identical questions are asked at each assessment, creating a section that displays the total frequency of questions referring to each subtopic is proposed. Question providers should focus more on questions that were less asked in the previous semester. This can also prevent students from simply memorising the answer rather than understanding what is being asked.
d. In most cases, the test justification table merely corresponds to the topic and cognitive level. It is also possible to incorporate psychomotor and affective factors into TST and TSTC so that the questions generated do not only test students in terms of cognitive only. On the other hand, question providers must know how to construct more comprehensive questions, such as those based on real-world case studies.
e. The generation of a TST or TSTC can be done using a systematic software system that can generate tables according to each topic, subtopic, marks, question number, and

taxonomy level at random. Regular and rapid schedule changes can help resource persons and question providers produce better quality and standardised questions. However, there needs to be an individual who should review and approve the resulting schedule each time before it is distributed to the question provider.

## 5. CONCLUSION

The course of statistics and probability was chosen as the study sample since it is one of the courses in statistics. According to the findings of the descriptive study, students who took the course performed well academically. In order to ensure the quality of the question, it is subjected to expert review in the field. However, the review procedure is limited to the final exam assessment. Thus, TSTC contribution to developing high-quality continuous assessment questions is vital.

Several more actions have been suggested in this study to improve the generation of test justification tables. Subtopic-related features, question frequency information, and the application of psychomotor and affective aspects were among the elements. Finally, implementing a systematic software system can aid in the timely and accurate creation of schedules.

This suggestion can be used to ensure that assessment questions are of a high enough quality to meet the established standards. Furthermore, the set of assessment questions should include a variety of difficulty levels, ranging from easy to challenging. Too difficult question items should be removed so that students in the vulnerable group can answer more questions at the medium and manageable levels. Students in the excellent category, on the other hand, must be tested with rather tricky questions. According to the level of study, the proportion of each level of difficulty for an assessment is typically specified in the TST or TSTC. Students' academic performance can be improved without interfering with their motivation to learn through assessments that have been developed utilising quality questions.

In addition to some highlighted suggestions to improve the existing TST, this study also wants to suggest further research based on two aspects: research methods and analytical methods. In terms of research methods, this study only targets one statistics course taken from one of the largest public universities in Malaysia. Therefore, it is proposed that data analysis be extended to all statistics courses for all universities in Malaysia. Following that, this research focuses solely on a descriptive analysis for measurement. More complicated statistical analyses, such as inference statistics and item response theory, should be considered in the future. The suggested extension should provide students' actual academic performance in the statistics course and the quality of assessment questions produced using the test justification table.

## ACKNOWLEDGEMENTS

The authors would like to thank UiTM and UPSI for making it possible for this study to be completed successfully.

## APPENDIX

**Measurement Statistics for Quiz 1**

| Statistic | Item | | | | | |
|---|---|---|---|---|---|---|
| | **a** | **b** | **c** | **d** | **e** | **f** |
| Mean | 3.64 | 1.80 | 2.85 | 3.24 | 3.30 | 2.35 |
| Median | 4.00 | 1.00 | 3.00 | 5.00 | 4.00 | 2.00 |
| Mode | 4 | 0 | 3 | 5 | 5 | 2 |
| Standard deviation | 1.254 | 1.877 | 1.093 | 2.075 | 1.701 | 1.610 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Skewness | -.747 | .643 | -.474 | -.595 | -.587 | .331 |
| Standard error for skewness | .135 | .135 | .135 | .135 | .135 | .135 |
| Kurtosis | -.140 | -1.068 | .955 | -1.385 | -.995 | -1.008 |
| Standard error for kurtosis | .269 | .269 | .269 | .269 | .269 | .269 |
| Range | 5 | 5 | 5 | 5 | 5 | 5 |
| Minimum | 0 | 0 | 0 | 0 | 0 | 0 |
| Maximum | 5 | 5 | 5 | 5 | 5 | 5 |

**Measurement Statistics for Quiz 2**

| **Statistic** | **Item** | | | | | |
|---|---|---|---|---|---|---|
| | **a** | **b** | **c** | **d** | **e** | **f** |
| Mean | 2.85 | 2.47 | 1.63 | 3.40 | 2.90 | 2.50 |
| Median | 3.00 | 3.00 | 1.00 | 4.00 | 3.00 | 2.00 |
| Mode | 3 | 3 | 1 | 4 | 4 | 5 |
| Standard deviation | 1.248 | 1.576 | 1.181 | 1.483 | 1.156 | 2.142 |
| Skewness | -.614 | .022 | .953 | -.839 | -.608 | .020 |
| Standard error for skewness | .135 | .135 | .135 | .135 | .135 | .135 |
| Kurtosis | .228 | -.803 | .708 | -.289 | -.522 | -1.711 |
| Standard error for kurtosis | .269 | .269 | .269 | .269 | .269 | .269 |
| Range | 5 | 5 | 5 | 5 | 5 | 5 |
| Minimum | 0 | 0 | 0 | 0 | 0 | 0 |
| Maximum | 5 | 5 | 5 | 5 | 5 | 5 |

**Measurement Statistics for Test 1**

| **Statistic** | **Item** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **a** | **b** | **c** | **d** | **e** | **f** | **g** | **h** | **i** | **j** | **k** | **l** |
| Mean | 3.33 | 2.46 | 2.97 | 3.13 | 3.17 | 2.72 | 2.74 | 2.15 | 2.49 | 2.64 | 2.65 | 2.35 |
| Median | 3.00 | 2.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 2.00 | 3.00 | 3.00 | 3.00 | 2.00 |
| Mode | 3 | 2 | 3 | 4 | 4 | 2 | 3 | 2 | 3 | 3 | 3 | 2 |
| Standard deviation | .766 | 1.039 | .679 | 1.079 | .923 | .941 | .614 | .770 | .586 | .818 | .707 | .698 |
| Skewness | -.361 | .347 | -.319 | -.493 | -.378 | .204 | -.266 | .310 | -.183 | -.400 | -.322 | .111 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard error for skewness | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 |
| Kurtosis | -.100 | -.856 | .890 | -.845 | -.737 | -.587 | .187 | -.219 | -.504 | -.298 | -.007 | -.167 |
| Standard error for kurtosis | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 |
| Range | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 |
| Minimum | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Maximum | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 |

**Measurement Statistics for Test 2**

| **Statistic** | **Item** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **a** | **b** | **c** | **d** | **e** | **f** | **g** | **h** | **i** | **j** | **k** | **l** |
| Mean | 2.84 | 2.67 | 2.34 | 3.10 | 2.90 | 2.68 | 3.44 | 3.18 | 2.70 | 3.82 | 3.53 | 3.21 |
| Median | 3.00 | 3.00 | 2.00 | 3.00 | 3.00 | 3.00 | 4.00 | 3.00 | 3.00 | 4.00 | 4.00 | 3.00 |
| Mode | 3 | 3 | 2 | 3 | 3 | 2 | 4 | 3 | 2 | 4 | 4 | 5 |
| Standard deviation | .724 | .838 | .687 | .770 | .710 | .996 | .945 | 1.103 | .892 | 1.020 | .890 | 1.467 |
| Skewness | -.330 | -.196 | .245 | -.536 | -.112 | -.066 | -.379 | -.067 | .602 | -.593 | -.331 | -.089 |
| Standard error for skewness | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 | .135 |
| Kurtosis | .044 | -.509 | -.022 | -.145 | -.412 | -1.118 | -.152 | -.687 | .118 | -.520 | -.325 | -1.461 |
| Standard error for kurtosis | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 | .269 |
| Range | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 |
| Minimum | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Maximum | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 |

**Measurement Statistics for Final Exam**

| **Item** | **Statistics** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Mean** | **Median** | **Mode** | **Standard deviation** | **Skewness** | **Standard error for skewness** | **Kurtosis** | **Standard error for kurtosis** | **Range** | **Minimum** |
| **1a** | 3.97 | 4 | 4 | 0.75 | -0.351 | 0.135 | -0.204 | 0.269 | 3 | 2 |
| **1b** | 3.29 | 3 | 3 | 0.945 | 0.128 | 0.135 | -0.575 | 0.269 | 4 | 1 |
| **1c** | 3.67 | 4 | 4 | 0.754 | -0.186 | 0.135 | -0.007 | 0.269 | 4 | 1 |
| **2a** | 3.82 | 4 | 4 | 0.982 | -0.479 | 0.135 | -0.512 | 0.269 | 4 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **2b** | 3.83 | 4 | 4 | 0.924 | -0.364 | 0.135 | -0.622 | 0.269 | 4 | 1 |
| **2c** | 3.51 | 4 | 4 | 0.914 | -0.243 | 0.135 | -0.359 | 0.269 | 4 | 1 |
| **3a** | 3.52 | 4 | 4 | 0.726 | -0.126 | 0.135 | 0.015 | 0.269 | 4 | 1 |
| **3b** | 3.06 | 3 | 3 | 0.786 | 0.159 | 0.135 | 0.022 | 0.269 | 4 | 1 |
| **4** | 3.31 | 3 | 3 | 0.718 | -0.138 | 0.135 | -0.486 | 0.269 | 3 | 2 |
| **5a** | 3.43 | 3 | 3 | 0.841 | 0.042 | 0.135 | -0.429 | 0.269 | 4 | 1 |
| **5b** | 3.44 | 3 | 3 | 0.789 | -0.042 | 0.135 | -0.445 | 0.269 | 3 | 2 |
| **5c** | 3.22 | 3 | 3 | 0.789 | -0.072 | 0.135 | -0.206 | 0.269 | 4 | 1 |
| **6a** | 3.4 | 3 | 3 | 0.741 | -0.171 | 0.135 | -0.161 | 0.269 | 4 | 1 |
| **6b** | 3.3 | 3 | 3 | 0.838 | -0.133 | 0.135 | -0.234 | 0.269 | 4 | 1 |
| **7a** | 3.08 | 3 | 3 | 0.763 | -0.046 | 0.135 | -0.191 | 0.269 | 4 | 1 |
| **7b** | 3.59 | 4 | 4 | 0.782 | -0.347 | 0.135 | 0.14 | 0.269 | 4 | 1 |
| **7c** | 3.45 | 3.5 | 4 | 0.742 | -0.263 | 0.135 | -0.126 | 0.269 | 4 | 1 |
| **8a** | 3.3 | 3 | 4 | 0.949 | -0.22 | 0.135 | -0.313 | 0.269 | 5 | 0 |
| **8b** | 3.83 | 4 | 4 | 0.885 | -0.676 | 0.135 | 1.12 | 0.269 | 5 | 0 |
| **8c** | 3.59 | 4 | 4 | 1.048 | -0.689 | 0.135 | 0.824 | 0.269 | 5 | 0 |
| **9** | 3.29 | 3 | 3 | 0.897 | -0.427 | 0.135 | 1.129 | 0.269 | 5 | 0 |
| **10a** | 3.99 | 4 | 4 | 1.084 | -1.448 | 0.135 | 2.755 | 0.269 | 5 | 0 |
| **10b** | 3.82 | 4 | 4 | 0.994 | -1.422 | 0.135 | 3.588 | 0.269 | 5 | 0 |
| **10c** | 3.55 | 4 | 4 | 1.268 | -0.651 | 0.135 | -0.169 | 0.269 | 5 | 0 |

## REFERENCES


[1] Abdullah, A. A. (2015). Analysis of Students' Errors in Solving Higher Order Thinking Skills (HOTS) Problems for the Topic of Fraction. *Asian Social Science, 11*(21), 133-142.

[2] Abdullah, A., Mokhtar, M., Halim, N., Ali, D., Tahir, L., & Kohar, U. (2017). Mathematics Teachers' Level of Knowledge and Practice on the Implementation of Higher-order Thinking Skills (HOTS). *Eurasia Journal of Mathematics, Science and Technology Education, 13*(1), 3-17.

[3] Arievitch, I. (2020). The Vision of Developmental Teaching and Learning and Bloom's Taxonomy of Educational Objectives. *Learning, Culture and Social Interaction, 25*, 100274.

[4] Cohen, L., Manion, L., & Morrison, K. (2018). *Research Methods in Education.* Routledge.

[5] Grundspenkis, J. (2019). Intelligent Knowledge Assessment Systems: Myth or Reality. *Frontiers in Artificial Intelligence and Applications*, 31-46.

[6] Joyce, T., Crockett, S., Jaeger, D., Altindag, O., & O'Connell, S. (2015). Does Classroom Time Matter? *Economics of Education Review, 46*, 64-77.

[7] Meier, Y., Xu, J., Atan, O., & Van Der Schaar, M. (2016). Predicting Grades. *IEEE Transactions on Signal Processing, 64*(4), 959-972.

[8] Radmehr, F., & Drake, M. (2018). Revised Bloom's Taxonomy and Major Theories and Frameworks That Influence the Teaching, Learning, and Assessment of Mathematics: a Comparison. *International Journal of Mathematical Education in Science and Technology*, 1-26.

[9] Raus, R., Janor, R., Sadjirin, R., & Sahri, Z. (2014). The Development of I-qubes for Uitm: From Feasibility Study to the Design Phase. *Proceedings - 2014 5th IEEE Control and System Graduate Research Colloquium, ICSGRC 2014*, (pp. 96-101).

[10] Sagala, P., & Andriani, A. (2019). Development of Higher-Order Thinking Skills (HOTS) Questions of Probability Theory Subject Based on Bloom's Taxonomy. *Journal of Physics: Conference Series*, *1188*, pp. 1-13.

[11] Sikder, M., Uddin, M., & Halder, S. (2016). Predicting Students Yearly Performance Using Neural Network: a Case Study of BSMRSTU. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, (pp. 524-529).

[12] Zainudin, S., Ahmad, K., Ali, N., & Zainal, N. (2012). Determining Course Outcomes Achievement Through Examination Difficulty Index Measurement. *Procedia - Social and Behavioral Sciences, 59*, 270-276.

[13] Zulkifli, F., Abidin, R., Razi, N., Mohammad, N., Ahmad, R., & Azmi, A. (2018). Evaluating quality and reliability of final exam questions for probability and statistics course using rasch model. *International Journal of Engineering and Technology(UAE), 7*(4), 32-36.