Educational Administration: Theory and Practice
2024, 30(6), 3035-3040
ISSN: 2148-2403
https://kuey.net/

Research Article

Utilizing High-Stake Online Assessment Reports To Improve Attainment Of Student Learning Outcomes

Anita Dani*

*Manipal Global-nxt University, Malaysia. Anita.Dani@campus.globalnxt.edu.my

Citation: Anita Dani, et al (2024), Utilizing High-Stake Online Assessment Reports To Improve Attainment Of Student Learning Outcomes, *Educational Administration: Theory and Practice, 30(6), 3035-3040*
Doi: 10.53555/kuey.v30i6.5956

ARTICLE INFO

ABSTRACT

Due to digitization in the education field, learning management systems (LMS) are widely used for teaching and conducting assessments. These LMS allow creation of randomized question pools, provide a secure examination environment, and generate reports post-assessments, such as the item analysis report. Item analysis report calculates two summary measures: difficulty level indicating the percentage of students who answered the question correctly and discrimination index indicating the quality of each question. These summary measures serve as the basis for improving test questions' quality and assessing the attainment of course learning outcomes. These system-generated reports help identify topics that are well-understood by most students. Appropriate remedial actions can be taken to ensure all course learning outcomes are attained by most students.

This research aims to compare students' performance in the final examination using item analysis reports spanning over three consecutive semesters. Descriptive and inferential analysis is used to examine differences in difficulty levels of the same questions measured in all three semesters.

Conclusion: The percentage of students, who solved multi-part questions on data analysis correctly, was higher than those in the preceding semester and this increase was statistically significant. Remedial teaching strategies, such as providing formative assessments using instant feedback, were useful in improving the attainment level of the course learning outcome.

Teaching Implications: Closely monitor students' performance using post-assessment reports. Provide more support using formative assessments to improve students' performance. It is recommended to keep the format and structure of formative and summative assessments consistent.

Keywords: Item analysis, online assessments, mathematics, data analysis

I.INTRODUCTION

The final exam is a major component of the assessment strategy, and it is administered at the end of the semester to measure the outcome of student learning. They are high stakes for students whose success is assessed and for teachers also as it can indicate teaching effectiveness [13]. Final exams are carefully developed to ensure coverage of all course learning outcomes and validity of test questions [11].

In the years 2020 and 2021, all educational institutes had to switch to online mode for course delivery including the high-stake summative assessments. Most institutes preferred to conduct their assessments within learning management systems, such as Blackboard, which allows a secure environment for exams. Within such systems, there is a provision for maintaining randomized question pools and the tests can be auto-graded. These systems also provide a facility of presenting questions in different formats so that it is also possible to set questions assessing different cognitive demands.

Another major advantage of conducting online assessments is that these systems generate detailed reports of students' performance in the assessments, such as the item analysis report. A detailed description item analysis is presented in the next section. Many researchers have recently demonstrated the effectiveness of such reports in improving the quality of assessments [2,3,9,11,15]. The focus of their research is to assess and improve the quality of test questions. Whereas, in this research, reports of item analysis are examined to

assess and improve students learning. The aim of this research is to compare the results of item analysis over a period of three semesters, during which classes were conducted online.
The rest of the paper is organized as follows: a review of existing research explaining terms used in item analysis is presented in the next section, which is then followed by a description of the research, data analysis, and findings.

II. LITERATURE REVIEW

Rash model is a commonly used statistical technique to analyze test data and evaluate question item banks. It is also used to evaluate development in longitudinal studies [5,6]. There are two assumptions of Rasch Analysis; one is the independence of items, that is, the probability of answering one item correctly should not be dependent on the answers to other questions and the second one is that the test is set to examine one type of ability. Due to the first assumption, Rasch analysis may be most appropriate for tests containing multiple-choice questions [6].

Another commonly used analysis technique is item analysis. Similar to Rasch's analysis, this analysis is performed post-assessment using data from students' responses to each test item (referred to as question in the rest of the paper) [16]. It calculates two measures of each question: difficulty level and discrimination index (denoted by letter d).
The difficulty level of an item is the ratio of the number of students, who answered it correctly to the total number of students who took the test. It is expressed as a percentage. Difficulty level should be between 30% and 80% for a good question [11,16]. A question which is found difficult by most students will have a difficulty level below 30% and it is above 80% for a question that is too easy for most students. Such questions do not provide authentic information about students' learning.

The discrimination index (d) of a question represents how well that question can distinguish between a strong and a weak student [16]. While examining a discrimination index of any question, it is expected that a student who has answered most questions correctly would answer this question also correctly. Similarly, a student who has not answered this question correctly would have answered other questions also incorrectly. The values of d can be between -1 to +1. Questions with a discrimination index higher than 0.3 are good quality questions. Whereas questions with a negative or a very low value for d do not meet the validity criteria and these questions should be reviewed. Such questions can be replaced or rewritten to maintain a good quality of the assessment instrument.
In this paper, the scope of analysis is set to examine the difficulty levels of chosen questions over a time frame of three consecutive semesters and assess if there is a positive trend of improving the difficulty levels of selected questions.

III.RESEARCH CONTEXT

The present research presents a comparative study of students' performance in Quantitative Reasoning course. This course is taken by students from non-technical and non-business majors. Although this course is not a direct pre-requisite for another mathematics course, its overall aim is to develop students' understanding of important concepts, such as analysis and interpretation of quantitative data. These skills are important in any profession. Data analysis is one of the five course learning outcomes of this course. The context of this research is set to in-depth analysis of students' understanding of data analysis. Concepts from this course learning outcomes include organization, analysis, and interpretation of primary and secondary data. These are assessed in the project assignment, a low-stake summative quiz and in the final exam.

Final exams were conducted online, but on-campus, within the learning management system under safe browsing conditions. In order to avoid any possibility of cheating, randomized question pools were developed. Each question was randomly presented to each student out of a pool of eight versions. All eight versions were of same for-mat having the same expected difficulty and they all assessed the same level of thinking order as per Bloom's taxonomy.

A. Research aim

The aim of this research is to examine the results of item analysis and use them to improve students' attainment of learning outcomes.

B. Research approach

A quantitative approach is taken in this research, where students' grade data is collected from the learning management system. This data was collected over three semesters, during which the course content and assessment strategy were the same. The core course delivery model and methods of instruction also remained the same, but a more rigorous remedial strategy was used in the third semester.

IV. DATA ANALYSIS AND INTERPRETATION

As stated above, the final exam was conducted within a learning management system, which has an in-built tool for performing item analysis. The item analysis report was generated using the system-built tool. Out of two questions chosen from the course learning outcome of data analysis, the first question assessed students' understanding of grouped frequency distribution. A group frequency distribution was presented, and students were asked to calculate measures of central tendency and variation. It was not a multiple-choice question, but students were expected to choose from a pool of options for each sub-question. Answers of sub-questions were independent of each other. Partial credits were given for correct answers to each sub-part.

The second question from data analysis was presented in a similar format but it was about comparing two raw data sets. Students were asked to choose the best measure of central tendency and comment on the consistency of each data set. According to Bloom's taxonomy, both questions assessed higher-order thinking skills, such as applying and analyzing. Discriminant index levels for each question were found to be 0.3 or higher, indicating it was not required to rewrite those questions.

It can be seen from table -1, that two out of eight versions of question 1 (grouped frequency distribution) had a difficulty level below 30% in the first two semesters, while only one version out of eight had a difficulty level below 30%. These versions were presented to 45 students randomly and they found it difficult. In question 2 (analysis of raw data sets), all eight versions had a difficulty level above 30%. None of the versions of both questions had a difficulty level above 60% indicating these were not very easy questions.

Table 1. Summary of difficulty level – Question on frequency distribution

	Semester		
Version of the question	Fall 2020	Spring 2020	Fall 2021
A	21%	14%	40%
B	33%	36%	39%
C	23%	40%	27%
D	45%	24%	38%
E	45%	39%	50%
F	46%	31%	68%
G	33%	50%	60%
H	41%	47%	46%
number of difficult questions	2	2	1
number of easy questions	0	0	0
Range of discriminant index	**0.3 to 0.8**	**0.3 to 0.8**	**0.3 to 0.9**

Table 2. Summary of difficulty level – Question on comparison of two data sets

Version of the question	**Fall 2020**	**Spring 2020**	**Fall 2021**
A	36%	55%	50%
B	32%	48%	52%
C	35%	44%	39%
D	49%	40%	33%
E	38%	55%	50%
F	47%	50%	48%
G	37%	46%	48%
H	39%	58%	58%
number of difficult questions	0	0	0
number of easy questions	0	0	
Range of discriminant index	**0.3 to 0.7**	**0.3 to 0.7**	**0.3 to 0.7**

After the first semester analysis, teachers took remedial actions to improve students' performance, such as frequent revisions using formative assessments with instant feedback. Although the effect of such remedial actions was not visible immediately in the next semester, but there is an overall trend of improvement in students' performance as shown in Figure 1.

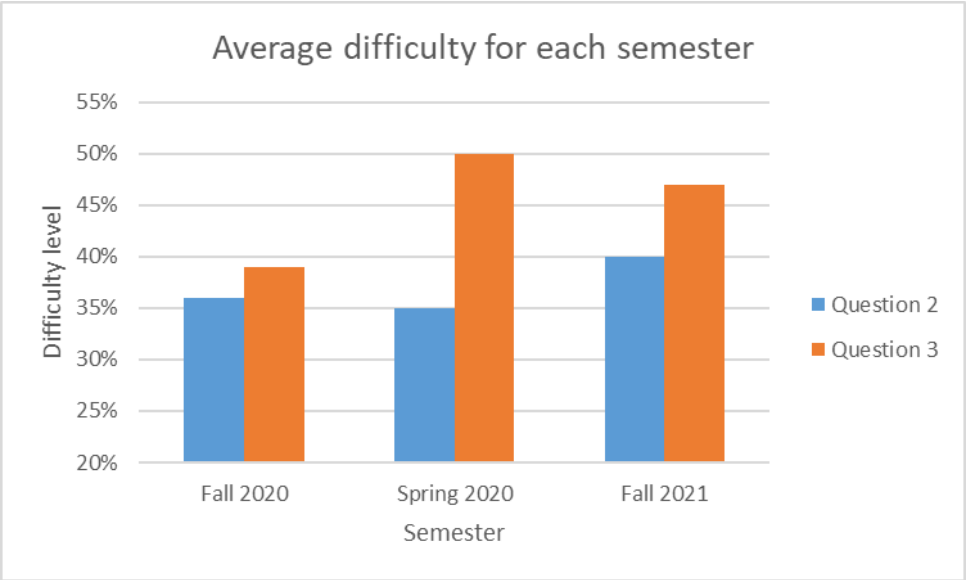

Fig. 1. – Average difficulty levels

Looking at the increase in average difficulty level from Spring 2020 to Fall 2021, further investigation was carried out to check if this difference is statistically significant.
The results of the independent samples t-test are given below. There is a piece of statistical evidence to support the claim that the difference is not by chance as the p-value is less than 0.05. See the Table -3 below.

Table 3. – Summary of independent samples t-test

	Spring 2020	***Fall 2021***
Mean	0.34875	0.46
Variance	0.013698	0.017229
Observations	8	8
Pearson Correlation	0.22876	
Hypothesized Mean Difference	0	
df	7	
t Stat	-2.03546	
P(T<=t) one-tail	0.040628	

V. DISCUSSION

Although the positive trend looks promising, the difficulty levels are still below 60%, which indicates that there can be other contextual factors affecting overall students' performance.
It is important to consider the effect of the external cognitive load caused due to irrelevant or out-of-context information included in the question [7]. The aspect of reduction in possible cognitive load had been considered while setting the exam and the description of the question included contextual information, such as, local currency and units of measurement. Besides the difficulty level of each question, the discrimination index reveals the validity of the question. In order to investigate and check for evidence from the item analysis report, we looked at other statistics, such as the discriminant index, mean score, and standard deviation for versions C and D of question 1, in the Fall 2021 data. These two questions have the lowest difficulty levels (27% and 38% respectively). The following values are found for each of these.

Table 4. – Other statistics from Fall 2021 data

Version	Discriminant Index	Number of students	Mean	Standard Deviation	Coefficient of variation
C	0.91	12	2.17	2.89	133%
D	0.79	7	3	3.30	110%

These two questions have excellent discriminant index, but the coefficient of variation is much higher. These numbers indicate that there is a large variation within these groups of students. Those who could not get these answers correct might have got other answers also wrong. These are students who did not perform well in the course which indicates that either they did not understand the core data analysis concepts or they did not review rigorously to ensure retention of their knowledge.

The topics chosen for this study are covered in the first three weeks of the semester. One possible reason for lower values of difficulty levels is this time lapse between the teaching weeks and the final exam period.

Although periodic cumulative review sessions are offered, they are not equally effective in motivating all students to re-call all topics or may be too demanding for weaker students. It is also an indication that most students are not able to retain their knowledge until the last week of the semester.

Lack of motivation and interest in studying Statistics is another reason why many students found these questions difficult. This has been reported in similar studies. Students in developing countries studying Statistics in English may find word problems difficult and may not be able to see the value of Statistics in their chosen career [1,10]. At the post-secondary level, it is expected that students demonstrate an ability to solve real-life problems. In such a situation, poor language comprehension skills can lead to poor performance in mathematics. As suggested in [13, 14], students' performance can be improved by giving them a lot of practice exercises which are presented in a similar format as expected in the summative assessments. A short-term solution to this issue is to provide a lot of examples to students and help them get familiar with the format and expectations of the final exam. This will ease out the cognitive load faced by students while reading the final exam questions which can subsequently reduce anxiety during exams. However, a more robust solution is to make students realize the value of developing data analysis skills in their future careers, which can motivate them to engage in learning.

After reviewing the results of Fall 2020 and Spring 2021, it was found that although self-paced review sessions were offered, only motivated students benefited from them. Less motivated and weaker students needed rigorous and instructor-guided review sessions. During the Fall 2021, more such review sessions were conducted.

Also, during the Fall 2021 semester, lots of formative assessments were provided to students for independent learning. These assessments were developed using interactive digital tools, such as Bookwidget worksheets, and Nearpod lessons. It was possible to provide more scaffolding and instant feedback and multiple attempts to give students hands-on practice. Instant and interactive feedback is a relevant formative assessment practice, which provides a personalized learning experience and enhances students' interaction with the contents [4]. This remedial treatment is found to be effective. Formative assessment practice is found to have a positive impact on students' self-regulation of learning. Our findings are concurrent with the results established in [8].

VI. CONCLUSION

Going beyond the analysis of overall pass rates, a detailed investigation was carried out using item analysis reports. This analysis was possible due to digitization in the education field, which has facilitated the generation of post-assessment reports. Results from item analysis reports highlighted the gap between expected and actual attainment of learning outcomes. Consequently, remedial actions were taken with the help of digital tools available which include providing formative assessments. These remedial actions improved difficulty levels significantly which implies that these can be continued to achieve consistent improvements in the following semesters.

Digitization in education is certainly adding value and more resources for empowering teachers to assess the effectiveness of their teaching. The use of item analysis reports is one such example. It has enabled a finer analysis of assessment quality as well as students' learning based on authentic quantitative data generated post-assessment.

VII. LIMITATIONS OF THIS STUDY AND FUTURE DIRECTION

The scope of this study was limited to the analysis of questions on one of the five learning outcomes, which is a noted limitation. In order to overcome this limitation, this study will be enhanced by including questions from other learning outcomes and by taking data from the next two semesters. Although a recommendation is provided to improve students' learning of Statistics, it is not statistically proven to be applicable in a different academic context. More research is needed to draw robust conclusions.

REFERENCES:

1. Abdelbasit, K. M. (2010). Teaching statistics in a language other than the students'. In Data and Context in Statistics Education: Towards an Evidence-based Society. Proceedings of the Eighth International Conference on Teaching Statistics. Voorburg, The Netherlands: International Statistical Institute. http://iase-web.org/documents/papers/icots8/ICOTS8_C215_ABDELBASIT.pdf.
2. Alias, M. (2005). Assessment of learning outcomes: Validity and reliability of class-room tests. World Transaction on Engineering and Technology Education, 4(2), 234-238.
3. Bai, X., & Ola, A. (2017). A Tool for Performing Item Analysis to Enhance Teaching and Learning Experiences. Issues in Information Systems, 18(1).

4. Barana, A., Marchisio, M., & Sacchet, M. (2021). Interactive feedback for learning mathematics in a digital learning environment. Education Sciences, 11(6), 279.
5. Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models.
6. Edwards, A., & Alcock, L. (2010). Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. Teaching Mathematics and its Applications: An International Journal of the IMA, 29(4), 165-175.
7. Gillmor, S. C., Poggio, J., & Embretson, S. (2015). Effects of Reducing the Cognitive Load of Mathematics Test Items on Student Performance. Numeracy: Advancing Edu-cation in Quantitative Literacy, 8(1).
8. Granberg, C., Palm, T., & Palmberg, B. (2021). A case study of a formative assessment practice and the effects on students' self-regulated learning. Studies in Educational Evaluation, 68, 100955.
9. Gugiu, M. R., & Gugiu, P. C. (2013). Utilizing item analysis to improve the evaluation of student performance. Journal of Political Science Education, 9(3), 345-361.
10. Hijazi, R., & Alfaki, I. (2020). Reforming undergraduate statistics education in the Arab world in the Era of information. Journal of Statistics Education, 28(1), 75-88.
11. Khoshaim, H. B., & Rashid, S. (2016). Assessment of the Assessment Tool: Analysis of Items in a Non-MCQ Mathematics Exam. International Journal of Instruction, 9(1), 119-132.
12. Kausar S., Dani, A. (2020). Do The Reading Skills Of Emirati Students Impact Their Problem-Solving Skills? International Journal of Education and Knowledge Manage-ment (IJEKM) 3(2): 1-10 (2020) Print ISSN: 2616-5198, Online ISSN: 2616-4698.
13. Kibble, J. D. (2017). Best practices in summative assessment. Advances in physiology education, 41(1), 110-119.
14. Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why Are the Mathematics National Examination Items Difficult and What Is Teachers' Strategy to Overcome It?. International Journal of Instruction, 10(3), 257-276.
15. Sharma, L. R. (2021). Analysis of difficulty index, discrimination index and distractor efficiency of multiple choice questions of speech sounds of English. International Research Journal of MMC, 2(1), 15-28.
16. Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. Medical Teacher, 33(6), 447-458.