ŠVIETIMAS: POLITIKA, VADYBA, KOKYBĖ
EDUCATION POLICY, MANAGEMENT AND QUALITY

POST EXAM ANALYSIS SELECTION OF HIGH AND LOW EXAM RESULT GROUPS BY NORMAL DISTRIBUTION CURVE

Faiz Marikar
General Sir John Kotelawala Defence University, Sri Lanka

Abstract

The key factor of an assessment is to minimize the errors by having a good reliability and validity of the assessment yardstick. To achieve high score in the test examinee must be aware about assessment cycle and use it in appropriate way in post exam analysis. Outcome of the results can be utilized as a constructive feedback in any given program. This cross-sectional study was conducted at department of Biochemistry, University of Rajarata. Multiple choice questions, structured essay type questions, objective structured practical examination, and continuous assessment was used in this study. Total number of students are 180 and was assessed for difficulty index, discrimination index, reliability, and standard error of measurement. In this study sample for analysis was used basically the examiner divides students into two groups ('high' and 'low') according to the score sheet of each student. Most of them are doing in a wrong way basically they divide high and low clusters as 25% each and considered upper quartile and lower quartile. In this study we compared it with the standard normal distribution curve where high and low groups are considered as 16% where is the standard. There is no significant difference among both clusters, and we recommend using the standard 16% as the high and low groups in post examination analysis.

Keywords: *difficulty index, post examination analysis, reliability of the examination, standard error of measurement*

Introduction

The reason for using post-exam analysis techniques is to improve the quality and reliability of the assessments and to select the most appropriate questions to assess students to gauge the proficiency level of underperforming students. Article analysis is an important step in developing a testing program. This phase uses statistical methods to identify test objects that are not working properly. If an item is too easy or too difficult, shows no difference between qualified and unskilled candidates, or is even rated incorrectly, an item analysis will indicate it. The two most common statistics reported in an article analysis are the difficulty of the article, which is a measure of the proportion of candidates who answered an article correctly, and the article discrimination, which is a measure of how the article was between candidates distinguishes between those who have knowledge of the article's content area and those who are not. In analyzing elements from test results, quantitative methods are used to assess which questions to accept, which questions to review, and which to reject.

In different professional examinations use of multiple-choice questions (MCQ), structured essay type question (SEQ), and practical assessment is frequently increasing to assess the knowledge of students (Fortun & Tempest, 2020). Well-constructed MCQ is a useful examination tool that can cover the wide area of subject with objectivity across all cognitive levels (Steinborn et al., 2021). It also lessens the evaluator's bias by minimizing

https://doi.org/10.48127/spvk-epmq/21.13.39

individual's judgement during scoring. Development of standardized MCQ is a can be easier or more difficult to be attempted by students as required. If the options given in MCQ are not according to standardized criteria, it will reduce the student recalling, comprehension or problem-solving skills and will direct the students towards guessing (Kikas et al., 2020; Fuchs et al., 2020; Feller et al., 2020). In medical colleges it is very important to give adequate and accurate knowledge to students and to improve their practical skills by objective specific practical examination (OSPE). A medical student should be more inquisitive and more analytical to develop appropriate professional attitude. The purpose of assessment taken during teaching and learning practice is multifold. It not only assures the students capability to grasp the knowledge given but also to observe that how much our teaching strategies are effective. Therefore, process of assessment should be effective and trustworthy (Weiskittel et al., 2021). To improve the students' knowledge and to enhance the quality of examination, continuous analyses of student's assessment methodologies should be a key step by Continuous Assessment (CA). There are previously defined pre-validation and post-validation assessment methods to analyze the formulated questions. In the process of pre-validation, before conduction of assessment a group of subject specialists should evaluate the applicability of topics covered in paper and appropriateness of structure of MCQs including stem and options and SEQ and OSPE also. Post validation process is basically a statistical method that is also called as item analysis. This is a valuable, relatively simple but an effective process to check the reliability and validity of MCQs, SEQs, CAs and OSPEs (Arooj et al., 2021). This is helpful in three aspects. First of all it tells that questions given to student is difficult or easy to attempt that is called the difficulty index (DIF). Secondly it can discriminate the students having good knowledge about subject assessed from those not performing well. It is called as discrimination index (DI). Thirdly it helps the subject specialist to assess the credibility of incorrect options (distractors). Furthermore, Reliability gives overall examination is well set according to the validity and finally Standard Error Measurement (SEM) gives the error rate of the examination. Overall, this analysis gives guidelines to evaluator to amend the questionss before next examination to make it more appropriate (James & Pattison, 2021; Morin-Chassé & Lachapelle, 2020).

In this study setup, mostly medical teachers are not able to assess the quality of their MCQs, SEQs, OSPEs, and CAs through item analysis. As a result, many unstandardized questionss can be added in examinations. Standard method of analysis of DIF, DI, R, and SEM is used High and Low scoring students were selected by quartile method. High performance was taken higher quartile (25% of the top score students) and Low performance students were selected from lower quartile (25% of the low score students) (Teltemann & Schunck, 2020). According to the normal distribution curve High and Low end defined as 16% each from High score and low score by student population. In this study we compared both categories and shown that there is no significant difference among both groups, and we want to emphasize that better use the standard distribution norm by selecting 16% than quartile. There is a statistical evidence for the analysis of the item analysis. Purpose of the study to examine the existence system with a proper statistical method to assess the use of proper method in analyzing exam results.

https://doi.org/10.48127/spvk-epmq/21.13.39

Research Methodology

This research was carried out at Department of Biochemistry, Faculty of Medicine University of Rajarata, Sri Lanka. Study design is cross-sectional study and data was collected from academic session 2013-2019. Total 180 students of 2nd year MBBS appeared in Biochemistry continuous and final examination. Before assessment paper was evaluated by a subject specialist. Paper was comprised of 40 MCQs, each having a single stem with five options including one correct answer and four distractors (incorrect answers). Each MCQ was assigned one mark. Maximum marks possible to score were 100 and minimum was zero, with no negative marking for this study. Continuous assessment three Structured Essay Question (SEQ) were given and final examination it was five questions. For item analysis, results of all papers were ranked in descending order, from highest marks to lowest marks. Then papers were divided into quartiles. Upper quartile or high scored (n=45) and lower quartile or low scored (n=45) groups were included into the analysis according to the standard where quartile considered as 25% in accordance with Cohens (Cohen and Swerdlik (2010)). Upper and lower groups in standard distribution curve is not 25% and its 16%. Compared the results in standard format where we compared the results in our standard format where 16% and it was compared with 25% as what Cohen said (Cohen and Swerdlik (2010)). The Difficulty index, discrimination index, reliability and Standard error of measurements were measured in both sections as follows.

The item-difficulty index (DIF)

If all students answer a question correctly or incorrectly, that question is not a good question and should be tested. It's too easy or too difficult.

$$DIF= [(H+L)/N] \times 100$$

H= Number of students gave correct options in high score group
L=Number of students gave correct options in low score group
T=Total number of students in both groups
Criteria of categorization in DIF is: DIF>70%=Too easy, DIF b/w 30-70%=Average, DIF b/w 50-60%= Good, DIF<30%=Too difficult

The item-discrimination index (DI)

The item discrimination index indicates the extent to which a question can distinguish between good and poor performers or between “strong” and “weak” students. Its interval is 0-1. The formula used to calculate the DI is

$$DI= 2\times[(H-L)/N]$$

H= Number of students gave correct options in high score group
L=Number of students gave correct options in low score group
T=Total number of students in both groups

https://doi.org/10.48127/spvk-epmq/21.13.39

DI is categorized as:
DI≤0.2= Poor, DI b/w 0.21-0.24= Acceptable, DI b/w 0.25-0.35= Good,
DI≥0.36=Excellent

Reliability of the examination

The traditional way of explaining and defining reliability is to look at the reproducibility, stability and internal consistency of a classification. The Kuder-Richardson 20 (KR-20) formula allows medical educators to estimate reliability between departments. It provides a confidence coefficient for the entire exam.

$$rKR20 = (\frac{k}{k-1})(1 - \frac{\sum pq}{\sigma^2})$$

K is the number of questions/stations,
σ^2 the variance of total station scores,
p is the probability of students who pass the test
q is the probability of students who fail the test
If the reliability coefficient is low it suggests that some stations do not share equally in the common core clinical performance and need to be revised or discarded.

The standard Error of measurement of the examination (SEM)

A final useful concept for post-examination analysis is the standard error of measurement (SEM). SEM provides an estimate of the amount of error inherent in a person's test result

$$\mathrm{SEM} = \mathrm{SD}\sqrt{1 - \mathrm{r}}$$

SD is the standard deviation; r is equal to the reliability coefficient of the test.

Research Results

Out of 180 students, total 90 were categorized as high performers and low performers by considering as quartile method (25%) where most of the studies were done. In the second group we have selected total of 58 from 180 students were selected as sixteen percent (16%) with respect to normal distribution curve. Therefore, two groups with 25% quartile and 16% as normal distribution were two groups where comparison were done in all SEQ, MCQ, OSPEs and CAs. Structed essay question analysis were done with 12 sets of papers each with 180 students and sample size in 25% as 90 and 16% as 58. It shows that the test reliability and SEM there is no difference in the significant by statistical method among both categories. Slight change was observed in Difficulty index and PQ where no change in significancy was observed among both parties as well (Table 1).

https://doi.org/10.48127/spvk-epmq/21.13.39

Table 1
Structured Essay Question analysis

SEQ Final	25% each in H and L	16% each in H and L
DI Average	74.44	70.83
PQ	2.04	2.27
Variance	7.22	7.22
Reliability	0.72	0.69
SEM	1.42	1.50
Question sets	12	12

Regarding DI, out of total 90, and 58 majority of MCQs were in the acceptable category (Table 2). Among these acceptable category average fall under the category of having good DI. Also, examination reliability in both groups its similar values and SEM also follows (Table 2)

Table 2
Multiple Choice Question Analysis

MCQ Final	25% each in H and L	16% each in H and L
DI Average	54.44	50.00
PQ	0.97	1.00
Variance	2.42	2.42
Reliability	0.60	0.59
SEM	0.98	1.00
Question sets	4	4

Three replication of continuous assessment analysis also confirmed that all DI in both groups there is no significant difference, and its DI reflects an excellently set the papers. However, Reliability and SEM is little varying among the groups and there is no significant difference (Table 3)

Table 3
Continuous Assessment Analysis

Continuous Assessment	25% each in H and L			16% each in H and L		
	A1	A2	A3	A1	A2	A3
DI Average	73.33	79.63	79.63	66.67	76.44	76.44
PQ	0.55	0.37	0.48	0.62	0.45	0.52
Variance	0.61	0.40	0.39	0.61	0.40	0.39
Reliability	0.10	0.07	0.24	0.03	0.11	0.33
SEM	0.74	0.61	0.69	0.79	0.67	0.72
Replication	3	3	3	3	3	3

Table 4 reflect the analysis of objective structured practical examination analysis is well set with both groups. PQ means P-probability of corrected answers and Q means probability of wrongly answered. Reliability and SEM also in both groups par with each other.

Table 4
Objective Structured Practical Examination Analysis

OSPE	25% each in H and L	16% each in H and L
DI Average	64.44	62.64
PQ	0.65	0.70
Variance	0.69	0.69
Reliability	0.05	0.04
SEM	0.81	0.83
Replication	4	4

In this study it was highlighted to use the normal distribution phenomenon due to the it has a meaningful way of doing the analysis. There is no significant difference among both groups where most of the people do the analysis with respect to quartile method, in this study it can be conclude and recommend following the standard method of 16% of each group consisted with high and low accordance with the normal distribution curve (Gaussian Distribution).

Discussion

A question with a single correct answer is an effective way to assess a student's cognitive knowledge. According to Blooms Taxonomy, a well-constructed question is an effective tool for quickly assessing different levels of knowledge, such as comprehension, application, analysis, and synthesis among students (Graham et al., 2021). However, the first mandatory step for quality assessment is standardization of question. Frequent evaluation of questions through item and test analysis is an active approach to make the valid.

In the post-exam analysis, most examiners did not use a standard method. Basically the examiner divides the students into two groups ("high" and "low") according to the evaluation sheet for each student. Based on this classification, 27% of the students are classified as a strong group and 27% as a weak group. Some methods prefer an "upper third" and a "lower third", but studies have shown that the sensitivity and precision of the d-value is increased when the students are divided into two groups to 27% (Cohen & Swerdlik, 2010). It is evident that 46% of students who received the average grade are excluded from the calculation of the disaggregated Discrimination Index.

Some other groups used quartile method for the analysis high and low groups are consist with quartiles it means 25% in a group. Obviously 50% group consist with middle scoring students. In our study we compared all the methods and used the normal distribution curve phenomena. It clearly says divides students into two groups (‘high’ and ‘low’) according to the marks with 16% in each group. Where 68% of the population in the middle scoring students. Clearly results shows there is no significant difference among the results and theoretically we used the correct way of selecting. Therefore we recommend using the standard way of using the post examination analysis.

Conclusion and Implications

The post examination analysis is an important exercise due to its feedback can go for a quality in all aspects in assessment. The item-difficulty index and the item- discrimination need to be calculated in standard way whereby considering the normal distribution curve. A large positive R is an indication of a good question or an assessment while a low positive or a negative R is an indication of a bad question or an assessment. SEM where is a good indicator of error of the test gives a good sign for validity of the examination. In this study we confirmed that use the normal distribution curve standards for analysis. The outcome itself have a meaningful thing to say in a statistical way. Therefore use 16% instead of 25% quartile where previous studied emphasis on it. Finally, for quality assessment process, conduction of faculty development program can be helpful to enhance the learning and performance of medical faculty for development of new standardized method.

References

Arooj, M., Mukhtar, K., Khan, R. A., & Azhar, T. (2021). Assessing the educational impact of cognitive level of MCQ and SEQ on learning approaches of dental students. *Pakistan Journal of Medical Sciences*, *37*(2), 445. https://doi.org/10.12669 /pjms.37.2.3475

Cohen, A. S., & Swerdlik, H. (2010). Audit quality, management ownership and the in formativeness of accounting earnings. *Journal of Accounting, Auditing and Finance*, *17*(1), 172. https://doi.org/10.1177%2F0148558X0201700102

Feller, D. P., Magliano, J., Sabatini, J., O'Reilly, T., & Kopatich, R. D. (2020). Relations between component reading skills, inferences, and comprehension performance in community college readers. *Discourse Processes*, *57*(5-6), 473-490. https://doi.org/10.1080/0163853X

Fortun, J., & Tempest, H. (2020). A case for written examinations in undergraduate medical education: Experiences with modified essay examinations. *Assessment & Evaluation in Higher Education, 45*(7), 926-939. https://doi.org/10.1080/02602938.2020.1714543

Fuchs, L. S., Fuchs, D., Seethaler, P. M., & Craddock, C. (2020). Improving language comprehension to enhance word-problem solving. *Reading & Writing Quarterly*, *36*(2), 142-156. https://psycnet.apa.org/doi/10.1037/edu0000467

Graham, L. M., Sahay, K. M., Rizo, C. F., Messing, J. T., & Macy, R. J. (2021). The validity and reliability of available intimate partner homicide and reassault risk assessment tools: A systematic review. *Trauma, Violence, & Abuse*, *22*(1), 18-40. https://doi.org/10.1177/1524838018821952

James, H. K., & Pattison, G. T. (2021). Disruption to surgical training during Covid-19 in the United States, United Kingdom, Canada, and Australasia: A rapid review of impact and mitigation efforts. *Journal of Surgical Education, 78*(1), 308-314. https://doi.org/10.1016/j.jsurg.2020.06.020

Kikas, E., Mädamürk, K., & Palu, A. (2020). What role do comprehension-oriented learning strategies have in solving math calculation and word problems at the end of middle school? *British Journal of Educational Psychology*, *90*, 105-123. https://doi.org/10.1111/bjep.12308

Morin-Chassé, A., & Lachapelle, E. (2020). Partisan strength and the politicization of global climate change: A re-examination of Schuldt, Roh, and Schwarz 2015. *Journal of Environmental Studies and Sciences*, *10*(1), 31-40. https://doi.org/10.1080/13669877.2010

Steinborn, A., Werner, S., März, M., & Brunk, I. (2021). Evaluation of a 3D-MC examination format in anatomy. *Annals of Anatomy-Anatomischer Anzeiger*, 151666. https://doi.org/10.1016/j.aanat.2020.151666

Teltemann, J., & Schunck, R. (2020). Standardized testing, use of assessment data, and low reading performance of immigrant and non-immigrant students in OECD countries. *Frontiers in Sociology, 5*, 544628. https://doi.org/10.3389/fsoc.2020.574811

Weiskittel, T. M., Lachman, N., Bhagra, A., Anderson, K., St. Jeor, J., & Pawlina, W. (2021). Team-based ultrasound objective structured practice examination (OSPE) in the anatomy course. *Anatomical Sciences Education*, *14*(1), 99-109. https://doi.org/10.1002/ase.2069

Received 14 February 2021; accepted 20 June 2021

Cite as: Marikar, F. (2021). Post exam analysis selection of high and low exam result groups by normal distribution curve. *Švietimas: politika, vadyba, kokybė / Education Policy, Management and Quality*, *13*(1), 39-46. https://doi.org/10.48127/spvk-epmq/21.13.39

Faiz Marikar
General Sir John Kotelawala Defence University, Kandawala Road, Dehiwala-Mount Lavinia 10390, Ratmalana, Sri Lanka.
E-mail: faiz@kdu.ac.lk
ORCID: https://orcid.org/0000-0003-4579-7263