Measuring test item quality: a case study in mathematics course

Gede Pramudya [1,2*], Muhammad Suyanto[5], Zuraida Abal Abas[1,2], Siti Nur Azrreen Ruslan[1], Aliza Che Amran[3,4]

[1)] Faculty of Information and Communication Technology (FTMK),
[2)] Centre for Advanced Computing and Technology (C-ACT),
[3)] Centre for Robotics and Industrial Automation (CERIA),
[4)] Faculty of Electrical and Electronics Engineering Technology (FTKEE),
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.
[5)] Universitas AMIKOM Yogjakarta, Ringroad Utara, Yogjakarta, Indonesia.

[*]Corresponding e-mail: gedepramudya@utem.edu.my

Keywords: Assessment, item quality, item analysis

ABSTRACT – Implementation of Outcome-Based Education (OBE) in Higher Education (HE) institutions requires quality items for summative students' learning. In order to meet the quality of test items, there have been attempts in such institutions assessment such as rigorous top-down practices, peer reviews or vetting or using item external examiners. However, these procedures have not been able to well-justify the quality of items. Some measures of item analysis were implemented on a set of test item sample of Calculus and Numerical Methods (BITI 1223) course, which was offered in session two for academic year 2018/2019 in the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). The analysis resulted in a recommendation of list of can-be-sustained, should-be-improved, and must-be-discarded items.

1. INTRODUCTION

The implementation of national-level OBE in HE institutions is a must as stated in Code of Practise for Programme Accreditation (COPPA) and Malaysia Qualification Framework (MQF) [2,3,4]. This requires education programme providers to prepare with high quality of students' learning assessment methods and tools. One of the most common assessment tools used for assessing lower students' cognitive achievement or attainment on Course Learning Outcomes (CLO) is instructor-made test items in forms of forced-choice objective structures. In practices, the item set was prepared by a group of lectures who plan, implement and evaluate the courses.

In order to meet a standard of quality for the set of items, several approaches or procedures have been practised. Peer reviews within departments were the most common approach as well as guidance, control and monitoring from the heads of departments or related vice deans. Further improvement on the practiced was initiated by hiring external item vetters. This item vetting is a lengthy and tedious process. Those approaches are able to review the quality of items in terms of their alignment with pre-set CLO or format related issues. However, they were unable to assess the item quality in terms of some basic issues on item validity nor reliability. There have been no significant measures taken for addressing issues on validity or reliability indicators such as item difficulty indices, item discrimination indices, nor coefficients of discrimination.

The difficulty index of an item measures how easy or difficult the item for the respondents. It is the ratio of the number of respondents that answer the item correctly. Lower indices signify more difficult items. Item discrimination index represents on how the item differentiate students into lower or higher achievers. Higher discrimination indices of items indicate better indicators that the items differentiate students well. The coefficients of discrimination of an item indicate contribution of the items' score to the total score. Higher the coefficient indicates higher contribution [1,5,6].

2. METHODOLOGY

For the study, a set of lecturer-made question problems consist of thirty (30) multiple-choice items were trailed in ninety-eight (98) undergraduate students undertaking Calculus and Numerical Methods (BITI 1223) course in the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, session two, for academic year 2018/2019. The items are aimed at assessing attainment CLO on area of Calculus, particularly for subareas of Real Numbers, Real Valued Functions and Graphs, Limits and Continuity, and Derivatives.

The students' responses on the items were analysed for their indices of difficulty (μ), indices of discrimination (α), and coefficients of discrimination (r_φ) by using the following formulas.

$$\mu = \frac{number\ of\ correct\ responses}{number\ of\ responses} \quad (1)$$

$$\alpha = \frac{c_u - c_l}{u} \quad (2)$$

where

c_u = number of upper performers that respond correctly;
c_l = number of lower performers that respond correctly;
u = number of upper or lower performers.

$$r_\varphi = \frac{\bar{x_1} - \bar{x_0}}{s_x}\sqrt{\frac{n_0 n_1}{n(n-1)}} \quad (3)$$

where

r_ϕ = the biserial coefficient of correlation of an item;
$\bar{x}_1$ = median of the total scores of those who

answered an item correctly;
$\bar{x}_0$ = median of the total scores of those who answered an item incorrectly;
s_x = standard deviation of the scores;
n_1 = number of those who answered an item correctly;
n_0 = number of those who answered an item incorrectly;
$n = n_1 + n_1$.

The analysis was computer-assisted with the use of Microsoft Excel spreadsheet. For classification and summary, then the indicators were categorised by use of internationally recognised standards as the followings [1,5].

- For norm-referenced tests, five choices items require 0.6 at the maximum for their indices of difficulty (μ). Criterion-referenced tests require lower value.
- If item discrimination index (α) is more or equal to 0.4, then the item is good, but if it is less than 0.2, then it is poor.
- If the item coefficient of discrimination (r_φ) is equal or more than 0.35, then the item can be retained. Else, the item should be revised or omitted.

3. RESULTS AND DISCUSSION

The following set of tables summarises the indices of difficulty (μ), indices of discrimination (α) and coefficients of discrimination (r_φ) of the thirty items.

Table 1 Values of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{r_\varphi}$

	Item 1	Item 2	Item 3	Item 4	Item 5	Item6	Item 7	Item 8	Item 9	Item 10
μ	0.65	0.86	0.84	0.65	0.68	0.31	0.64	0.49	0.85	0.57
α	0.63	0.41	0.71	0.74	0.83	0.35	0.81	0.44	0.52	0.64
r	0.47	0.40	0.53	0.53	0.62	0.38	0.65	0.39	0.31	0.68

	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20
μ	0.71	0.73	0.86	0.70	0.87	0.73	0.82	0.78	0.38	0.73
α	0.56	0.56	0.58	0.43	0.68	0.60	0.58	0.46	0.22	0.75
r	0.47	0.30	0.44	0.24	0.38	0.38	0.41	0.45	0.13	0.57

	Item 21	Item 22	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28	Item 29	Item 30
μ	0.78	0.51	0.66	0.26	0.55	0.30	0.82	0.68	0.58	0.53
α	0.47	0.47	0.58	0.44	0.54	0.39	0.57	0.27	0.48	0.37
r	0.44	0.37	0.39	0.54	0.42	0.52	0.42	0.22	0.33	0.29

From the tables above, elaboration and discussion can be proposed as follows. From the first-two ten groups of items, only items 6, 8, 10 and 19 can be accepted or retained since all standard are surpassed. However, item 10 is a little bit too easy for a criterion-referenced assessment item. Other items in the group are too easy, even for a norm-referenced test, even though they well discriminate higher or lower performers as well as all items are powerful in responders' performance prediction and score contribution. Therefore, they may be subjected to revision or omission for future uses.

Slightly different finding can be observed from the last ten item set. There are five (5) items that may be retained, which are item 22, 24, 25, 26, and 29 since their properties approximately met the standards. For an example, item 24 is able to discriminate students well as well as its consistence with other items and prediction ability of students' scores, even though it is somewhat difficult to be attempted by the students. In opposite, item 30 is sufficient in difficulty and in ability to differentiate good performers but is unable to show its predictive power.

4. CONCLUSIONS

Implementation of OBE in learning and teaching requires criterion-referenced testing for students' summative learning assessment [2]. For this approach, quality assessment tools are needed. The study on some properties for such quality of a set of lecturer-made multiple choice items showed that the processes to uncover the measures such as indices of difficulty, indices of discrimination, and coefficients of discrimination of test items are affordable. It also found that only thirty (30) per cents of the set met the recognised standards.

5. ACKNOWLEDGMENT

The authors would like to address deep gratitude and appreciation to any member of Pervasive Computing & Educational Technology (PET) research group, Centre for Advanced Computing Technology (C-ACT) of Universiti Teknikal Malaysia Melaka (UTeM) and Universitas AMIKOM Yogjakarta, for any valuable contribution made. Also, to MESTECC and National STEM Movement for the grant to sponsor this paper.

REFERENCES

[1] Ebel, R.L. & Frisbie, D.A. (1991). Essentials of Educational Measurement. 5th Edition, Prentice-Hall, Englewood Cliffs.
[2] Malaysian Qualification Agency. (2014). Guidelines to Good Practices: Assessment of Students.
[3] Malaysian Qualification Agency. (2017). Malaysian Qualification Framework (MQF). 2nd Edition.
[4] Malaysian Qualification Agency. (2018). Code of Practice for Program Accreditation. 2nd Edition.
[5] McCowan, R.J. & McCowan, S.C. (1999). Item Analysis for Criteria-Referenced Test. Center for Development of Human Services, Buffalo State College (SUNY) 1695 Elmwood Avenue.
[6] Quaigrain, K., Arhin, A.K. & Hui, S.K.F. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation, Cogent Education, 4:1.