The Impact of Item Writer Training on Item Statistics of Multiple-Choice Items for Medical Student Examination

Cherdsak Iramaneerat, M.D., Ph.D.
Department of Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand.

ABSTRACT

Background: Multiple-choice question (MCQ) item writer training is an important step towards the production of high quality tests. Educators generally suggested faculty development programs of long duration, which are not practical for many clinical teachers. This study explored an alternative approach, using a series of short workshops.

Objective: This study examined (1) the perception of clinical teachers on item writer training workshops, and (2) the impact of training on item difficulty and discrimination.

Methods: A series of three short workshops on MCQ item development and item analysis were administered. Participating clinical teachers were asked to provide satisfactory ratings of the workshops. Items developed for comprehensive examinations of fifth-year medical students were analyzed, comparing item difficulty and item discrimination between those written by workshop participants and non-participants, both before and after the workshops.

Results: Participants were very satisfied with the workshops. The items developed by participants before the workshop tended to be inappropriately too easy or difficult than those developed by non-participants, but showed similar discriminating power with items of non-participants. After the workshops, participants developed test items that were similarly appropriate in the difficulty level as compared with items of non-participants. Post-workshop items of participants had higher discriminating power than those of non-participants.

Conclusion: A series of three short workshops on item writer training is an effective and acceptable program for clinical teachers. It has a positive impact on test statistics, improving item quality both in their difficulty and discrimination.

Keywords: Item writer training, multiple-choice test items, item statistics, item analysis

Siriraj Med J 2012;64:178-182
E-journal: http://www.sirirajmedj.com

INTRODUCTION

Multiple-choice question (MCQ) tests are a widely used assessment method in medical education. Its popularity is based on its testing efficiency, scoring objectivity, high internal consistency reliability, along with strong research evidence supporting its validity.[1] A properly written MCQ test items can assess not only low level recall knowledge, but also understanding and application of knowledge.[2-3] Unfortunately, researchers have demonstrated that the quality of MCQ test items that individual medical schools developed in-house were relatively low.[4-5]

One critical piece of validity evidence for MCQ tests is an item writer training.[6-8] Many researchers have shown that proper training substantially enhanced item quality.[4,9] However, not all item writer training programs are equally effective. Faculty workshops are one commonly used method.[9] The effectiveness of these item writer training workshops is debatable. Some researchers believed that isolated short-term workshops are inadequate to improved teachers' performance and proposed that faculty development workshops should span over several days.[9-11] However, given a busy schedule of clinical teachers, setting up an item writer training workshop that spans several days is a format that would not be popular in many medical schools. Only few clinical teachers would be able to attend such workshop.

An alternative item writer training program in a series of short workshops has been proposed. A series of five short MCQ item writer training workshops had been done with success.[12] In this study, I studied the impact of a series of three MCQ item writer training workshops at

Correspondence to: Cherdsak Iramaneerat
E-mail: sicir@mahidol.ac.th, CherdsakIramaneerat@gmail.com
Received 14 December 2011
Revised 10 April 2012
Accepted 11 April 2012

Faculty of Medicine Siriraj Hospital. The primary objective was to examine the impact of these workshops on the quality of MCQ items for medical students. The secondary objective was to evaluate the perception of clinical teachers on the appropriateness of these workshops.

MATERIALS AND METHODS

I carried out this study in two steps. In the first step, I ran a series of MCQ item development and item analysis workshops for clinical teachers at Faculty of Medicine Siriraj Hospital. In the second step, I evaluated the item statistics of MCQ items that faculty members developed for a comprehensive examination for fifth-year medical students.

I administered a series of three workshops. The first one was a three-hour session of MCQ item development guidelines and common item flaws. The second and third workshops were each two-hour teaching of classical item analysis and how to use the results of item analysis to improve item quality. The first workshop was scheduled three months prior to the second and third workshops. Participation in these workshops was voluntary. At the end of each workshop, participants were asked to evaluate the quality of the workshop using questionnaires. All questionnaire items were satisfaction ratings on a five-point Likert scale, where one referred to very unsatisfied and five referred to very satisfied.

The MCQ test items that I examined were test items developed for comprehensive examination administered to fifth-year medical students. All medical students at Faculty of Medicine Siriraj Hospital are required to take this examination at the end of their fifth year, as a partial fulfillment towards the MD degree. I analyzed two cohorts of test items. The first cohort was test items employed in a comprehensive examination given about one year prior to the workshop. The analysis of these items served as a baseline of item quality before an intervention. The second cohort was test items employed in a comprehensive examination given about nine months after the first workshop (six months after the second and third workshops). The analysis of these test items should demonstrate the impact of workshops on item quality.

In each test item cohort, I carried out classical test theory item analysis and checked for two item statistics, item difficulty (p-value), and item discrimination (point-biserial correlation). Item difficulty is the proportion of examinees answering the item correctly. The values range from 0 to 1. Higher values indicate easier test items. Testing experts suggested categorizing items according to their difficulty into four classes: class 1 (excellent) items have p-values range from 0.45 to 0.75, class 2 (good) items have p-values range from 0.76 to 0.91, class 3 (acceptable) items have p-values range from 0.25 to 0.44, and class 4 (poor) items have p-values below 0.25 or above 0.91.[2,13-14]

Item discrimination indicates the ability of test items to separate high and low scorers. The most popular item discrimination index is a point-biserial correlation (r).[13-15] The values range from -1 to 1. Higher values indicate better discriminating items, those that can differentiate high and low scorers well. Testing experts suggested categorizing items according to their point-biserial correlations into four classes: class 1 (excellent) items have r values higher than 0.20, class 2 (good) items have r values range from 0.15 to 0.19, class 3 (acceptable) items have r values range from 0.10 to 0.14, and class 4 (poor) items have r values lower than 0.10.[2,13-14]

In each test item cohort, I classified items according to their p and r values. I then compared the item distribution between those developed by workshop participants and those developed by non-participants. I evaluated the statistical significance of the difference in item distributions between the two teacher groups using a contingency-table Chi-square test. I also compared the mean p and r values between the two teacher groups using independent-samples t test. All statistical analyses were carried out with the assumption of Type I error rate of 0.05, using SPSS version 11.5.

This research was exempted from IRB review because the study was a commonly accepted educational practice in an established educational setting and involved only data obtained from standard educational tests and survey questionnaires that were recorded in a manner that human subjects who responded to questionnaires could not be identified.

RESULTS

1. Clinical teachers' perception of the workshops

The three item writer training workshops were well received by clinical teachers. The numbers of participants of the first, second, and third workshops were 68, 51, and 51, respectively. All workshop sessions had lively discussion about MCQ item development and item analysis. The numbers of returned questionnaires obtained from the first, second, and third workshops were 43 (63%), 44 (86%), and 43 (84%), respectively. Clinical teachers had good perception of these workshops as demonstrated in Table 1. Items related to workshop content, teaching quality, learning materials, and achievement of their objectives got average ratings higher than four. The only two items that had average ratings below four were the effectiveness of public relation activity, and the appropriateness of the classroom.

TABLE 1. Medical teachers' satisfaction ratings of MCQ item writer training workshops.
(1 = very unsatisfied, 2 = unsatisfied, 3 = neutral, 4 = satisfied, 5 = very satisfied).

	Workshop 1		Workshop 2		Workshop 3	
Items	Average	SEM	Average	SEM	Average	SEM
1. Appropriateness of content	4.22	0.11	4.57	0.08	4.57	0.09
2. Appropriateness of time period	4.24	0.11	4.23	0.11	4.35	0.10
3. Instructor's teaching effectiveness	4.33	0.10	4.69	0.07	4.75	0.07
4. Achievement of objectives	4.15	0.11	4.40	0.08	4.48	0.09
5. Appropriateness of the classroom	4.19	0.10	3.66	0.17	4.42	0.09
6. The audiovisual and learning materials	4.16	0.10	4.35	0.10	4.40	0.10
7. The effectiveness of public relation activity	3.90	0.13	3.89	0.11	3.77	0.14

TABLE 2. Item classification based on item difficulty and item discrimination from the analysis of the comprehensive examination administered prior to the workshops.

Statistics	Class	Non-participants	Participants	Total
Item difficulty (p)	1: Excellent	122 (44.36%)	3 (12%)	125
	2: Good	51 (18.55%)	6 (24%)	57
	3: Acceptable	43 (15.64%)	8 (32%)	51
	4: Poor	59 (21.45%)	8 (32%)	67
	Total	275 (100%)	25 (100%)	300
Item discrimination (r)	1: Excellent	99 (36%)	11 (44%)	110
	2: Good	51 (18.55%)	6 (24%)	57
	3: Acceptable	49 (17.82%)	1 (4%)	50
	4: Poor	76 (27.64%)	7 (28%)	83
	Total	275 (100%)	25 (100%)	300

TABLE 3. Item classification based on item difficulty and item discrimination from the analysis of the comprehensive examination administered after the workshops.

Statistics	Class	Non-participants	Participants	Total
Item difficulty (p)	1: Excellent	91 (41.94%)	26 (31.71%)	117
	2: Good	46 (21.20%)	22 (26.83%)	68
	3: Acceptable	39 (17.97%)	19 (23.17%)	58
	4: Poor	41 (18.89%)	15 (18.29%)	56
	Total	217 (100%)	82 (100%)	299
Item discrimination (r)	1: Excellent	62 (28.57%)	29 (35.37%)	91
	2: Good	49 (22.58%)	17 (20.73%)	66
	3: Acceptable	36 (16.59%)	22 (26.83%)	58
	4: Poor	70 (32.26%)	14 (17.07%)	84
	Total	217 (100%)	82 (100%)	299

TABLE 4. Item classification based on item difficulty and item discrimination of items developed before the workshops compared to items developed after the workshops.

Statistics	Class	Before	After	Total
		Workshops	Workshops	
Item difficulty (p)	1: Excellent	125 (41.67%)	117 (39.13%)	242
	2: Good	57 (19%)	68 (22.74%)	125
	3: Acceptable	51 (17%)	58 (19.40%)	109
	4: Poor	67 (22.33%)	56 (18.73%)	123
	Total	300 (100%)	299 (100%)	599
Item discrimination (r)	1: Excellent	110 (36.67%)	91 (30.43%)	201
	2: Good	57 (19%)	66 (22.07%)	123
	3: Acceptable	50 (16.67%)	58 (19.40%)	108
	4: Poor	83 (27.67%)	84 (28.09%)	167
	Total	300 (100%)	299 (100%)	599

2. The impact of workshops on item statistics

2.1 Item statistics prior to the workshops

There were 300 scored test items in the comprehensive examination administered prior to the workshops. Among these 300 items, 275 items were developed by non-participants, and 25 items were developed by workshop participants. Two hundred and forty-two students took this exam.

a.) Item difficulty

Item difficulty classification revealed that prior to the training, workshop participants had significantly higher proportion of classes 3 and 4 items and lower proportion of classes 1 and 2 items (Table 2), χ^2 (3, N=300) = 10.87, p = 0.01. However, there was no significant difference in item difficulty levels between the two groups of teachers, t (26.9) = 1.19, p = 0.24. This indicates that before training, clinical teachers who participated in the workshops tended to write inappropriately easy or difficult test items more than workshop non-participants.

b.) Item discrimination

Item discrimination classification revealed that prior to the training, workshop participants and non-participants had similar distribution of item classes (Table 2), χ^2 (3, N=300) = 3.39, p = 0.34. Point-biserial correlation coefficients of the two groups of clinical teachers had no significant difference, t(298) = 0.01, p = 0.99. This demonstrated the equivalence of both groups in their item discrimination prior to training.

2.2 Item statistics after the workshops

The original comprehensive exam that was administered after the workshops contained 300 items. However, one item was eliminated from final scoring due to the controversy in item content, which was discovered after test administration. Thus, there remained 299 scored test items. Among these, 217 items were developed by non-participants, and 82 items were developed by workshop participants. Three hundred and fifteen students took this exam.

a.) Item difficulty

The classification of item difficulty as shown in Table 3 demonstrated a similar distribution of items in both groups of teachers, χ^2 (3, N=299) = 3.26, p = 0.35. There was no significant difference in average difficulty levels of items developed by the two groups of clinical teachers, t(297) = -0.49, p = 0.62. This finding indicated the improvement of test item quality developed by teachers who attended the workshops. After participating in the item writer training workshops, clinical teachers reduced the number of items that were too easy or too difficult.

b.) Item discrimination

Item discrimination classification (Table 3) revealed that workshop participants generated significantly more test items in classes 1 and 2, and less items in classes 3 and 4, as compared with non-participants, χ^2 (3, N=299) = 9.1, p = 0.03. The workshop participants had a significantly higher value of point-biserial correlation coefficient than non-participants, t(297) = -2.48, p = 0.01. This finding indicated that workshop participants generated more discriminative test items than non-participants.

2.3 Comparison of item statistics before and after the workshops

Considering the findings that workshop participants significantly improved their item statistics, I asked if these improvements were enough to impact the quality of the whole examination. I carried out another analysis to compare item statistics of the two item cohorts.

a.) Item difficulty

The distribution of items based on their difficulty showed a similar pattern in the two item cohorts (Table 4), χ^2 (3, N=599) = 2.66, p = 0.45. There was no significant difference in average item difficulty of the two item cohorts, t (597) = -0.66, p = 0.51. This finding suggested that the number of items that participating clinical teachers provided in the comprehensive examination was not enough to impact the quality of the whole test.

b.) Item discrimination

The classification of items based on their point-biserial correlations revealed a similar pattern of item distribution of the two item cohorts (Table 4), χ^2 (3, N=599) = 3.05, p = 0.38. There was no significant difference in average item point-biserial correlation coefficients of the two item cohorts, t (597) = 0.83, p = 0.41. This finding suggested that the number of test items with class 1 or 2 discrimination power that was developed by workshop participants was not enough to impact the discrimination power of the whole test.

DISCUSSION

Item writer training is a critical step in the production of high quality test items for assessment of medical students. Although many researchers suggested that an effective faculty development program should require a protected time for days, many clinicians would not be able to attend such lengthy training program. In this study, I examined an alternative approach that employed a series of short workshops to train item writing skills for clinical teachers.

My results indicated that clinical teachers who attended the workshops were very satisfied with this training format. Their average ratings on satisfaction with teaching content, teaching effectiveness, teaching materials, teaching time period, and achievement of trainees' goals were higher than four (satisfied). The only two items that were the potential areas for improvement were the public relation activities and the classroom setting, which had average ratings lower than four (satisfied). In the future planning of item writer training workshops, we should make the announcement about the workshop earlier, using more media, and repeat the announcement more frequently. We should also carefully choose a classroom that is more convenient for a discussion; avoid using a classroom that was set up for a traditional one-way lecture.

The analysis of item statistics revealed that a series of short workshops was effective in improving the quality of test items as demonstrated by the improvement of both item difficulty and discrimination indices (Fig 1 and 2). An interesting finding that was also revealed from the analysis was that the number of items that workshop participants developed had increased from 25 items before the workshops to 82 items after the workshops. This may suggest that our workshops also help improve participants' confidence in their item writing skills, resulting in their increased contribution of items to the comprehensive exam. Nevertheless, their contribution of test items was not enough to impact the quality of the whole test. This suggested that more item writer training workshops should be needed to train more clinical teachers.

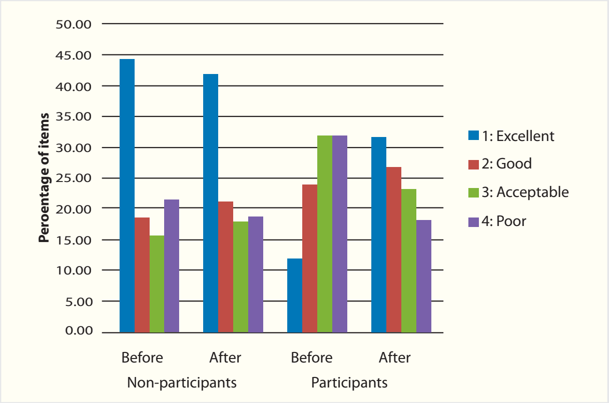

Fig 1. Distribution of test items according to their item difficulty classification, comparing between before and after the workshop.

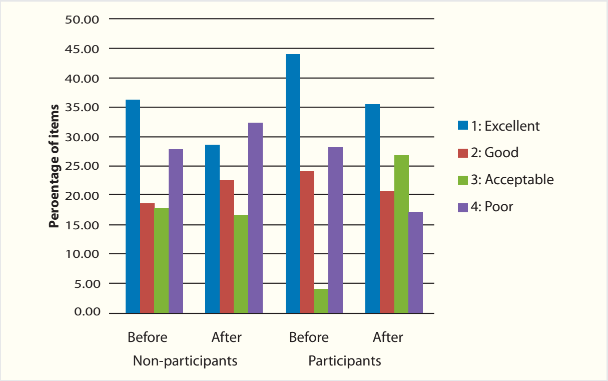

Fig 2. Distribution of test items according to their item discrimination classification, comparing between before and after the workshop.

There are some limitations of the generalizability of the findings from this study. First, due to the nature of the study in an educational setting which randomization of subjects is difficult to attain, this study recruited clinical teachers into these item writer training workshops on voluntary basis. Thus, there may be a selection bias. Clinical teachers who voluntarily attended the workshops are likely to be those who were interested in improving their item writing skills. This may explain why these teachers provided high satisfaction ratings on the questionnaire and why they were so quick in making changes in their item writing behavior. If the workshops were to be given to other groups of clinical teachers that unwillingly attend the activity, the results may be different.

Another noteworthy detail is the questionnaire response rates, which were 63 to 86%, which are typical numbers for survey research. There were other workshop participants who did not return the questionnaires. Those participants might have different perspectives about the appropriateness of the workshops. Thus, interpretation of the questionnaire analysis should be done with caution.

Because the study was not conducted in a controlled experimental environment, other factors besides the workshops may also influence the outcomes, considering a long period of time it took between the administrations of two cohorts of test items. Clinical teachers who attended the workshops might also seek advice from experts, study guidelines from the literature, or find other means to improve their item writing skills on their own as well.

Another important consideration is the decay of item writing skills. Like all cognitive skills, without regular practice, the skills could be forgotten over time. This study only examined the quality of test items developed by clinical teachers in less than six months after training. It would be interesting to check if the items these workshop participants developed one or two years after their training still possess item statistics that are better than items they developed before their participation in the workshops.

REFERENCES

1. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten C, Newble DI, editors. International handbook of research in medial education. Dordrecht: Kluwer Academic Pubishers, 2002. p. 647-72.
2. Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Assoicates, 2004.
3. Maatsch JL, Huang RR, Downing SM, Munger BS. The predictive validity of test formats and a psychometric theory of clinical competence. The 23rd Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1984.
4. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002; 77:156-61.
5. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ. 2008;42:198-206.
6. Downing SM, Haladyna TM. Test item development: Validity evidence from quality assurance procedures. Appl Meas Educ. 1997;10:61-82.
7. Downing SM. Twelve steps for effective test development. In: Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, 2006. p. 3-27.
8. Schmeiser CB, Welch CJ. Test development. In: Brennan RL, editor. Educational measurement, 4th ed. Washington, DC: The National Council on Measurement in Education and the American Council on Education, 2006. p. 307-53.
9. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ Theory Pract 2012;17:369-76.
10. Hill HC. Learning in the Teaching Workforce. Future Child. 2007;17:111-27.
11. Darling-Hammond L. School reform at the crossroads: Confronting the central issues of teaching. Educ Policy. 1997;11:151-66.
12. Ware J, Vik T. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Med Teach. 2009;31:238-43.
13. Haladyna TM. Writing test items to evaluate higher order thinking. Boston, MA: Allyn and Bacon, 1997.
14. Haladyna TM. Writing multiple choice items. Chicago, IL: CAT Inc, 2003.
15. Livingston SA. Item analysis. In: Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, 2006. p.421-41.