RESEARCH

Collaborative Examination Item Review Process in a Team-Taught, Self-Care Sequence

David J. Caldwell, PharmD, Laurel Sampognaro, PharmD, Adam N. Pate, PharmD

University of Louisiana at Monroe School of Pharmacy

Submitted August 14, 2014; accepted January 9, 2015; published August 25, 2015.

Objective. To improve examination item quality by educating and involving course instructors in evidence-based item review and encouraging use of this process in future courses.

Methods. A peer-review process was implemented in a 2-course sequence (intervention) that involved training and review sessions before each examination and was compared to the previous year's courses (control). Instructors completed a presurvey and postsurvey on training, experience, self-confidence, and self-rated success in multiple-choice item writing. Statistics were calculated for all items in the control and intervention sequences and compared using independent *t* tests. Items also were classified into levels based on difficulty and discrimination, and distribution into these levels was compared between sequences with independent *t* tests.

Results. No significant difference was found between control and intervention sequence items with regard to mean difficulty (86.3% and 84.4%) or discrimination (0.23- and 0.25), respectively, although item classification distribution did appear to change between the control and intervention sequences' subjective feelings of confidence, and success in item writing increased between presurvey and postsurvey. Confidence in ability to peer-review test items and to implement a formal item evaluation process also increased.

Conclusion. Item statistics did not change significantly, but reviewed and edited items distributed more favorably into item statistic-based categories. This method of review positively affected instructors' perceptions of their item-writing confidence and success and improved self-rated opinions of their ability to edit items and train others to do so.

Keywords: multiple-choice, item-writing, test construction, classroom assessment, peer review

INTRODUCTION

Examination item writing is a complex process that faculty members may find challenging. Texts are available to help item writers with formatting and techniques.¹⁻⁵ The Accreditation Council for Pharmacy Education (ACPE) Guideline 11 specifically addresses a need to "systematically improve the assessment process" and recommends having experts in educational methodology assist in this process.⁶ Although expert assistance may be ideal, use of these resources may not be feasible in many pharmacy programs for various reasons, such as monetary constraints and availability of experts.

Significant variation in examination quality exists among programs, departments, and individual faculty members based on item writing training.⁷ A systematic process for examination development and item review could help address these issues and lead to higher quality

examinations, for which nonknowledge-related variability is limited by attempting to decrease deviation from item-writing guidelines. Instead of employing an educational methodology expert as recommended by ACPE, faculty members themselves could create a process for improving assessment materials in the curricula. Faculty development activities in item writing may be beneficial in this process, but one-time trainings without application may not be sufficient. A better method to produce tangible improvements may involve developing a formal item-writing process including faculty training and examination review executed at the course level.

Examination items not adhering to item-writing guidelines may affect student performance on examinations.^{3,8,9} Education literature indicates that intense training in item writing can improve the quality of faculty members' examination items.¹⁰ However, whether individuals may improve overall quality of examinations during a semester course has not been evaluated. In this study, the authors implemented a course-level examination improvement process through a peer-led item-review

Corresponding Author: Adam Pate, PharmD, ULM School of Pharmacy, 1800 Bienville Drive, Monroe LA 71201. Tel: 318-342-1903. Fax: 318-372-5290. E-mail: apate@ulm.edu

program. The primary goal of this project was to improve examination quality through a faculty development program, followed by a longitudinal item review occurring before examination administration; the study also was intended to improve faculty members' self-rated confidence and success and to measure changes in their opinions regarding item-writing guidelines and review.

METHODS

A pair of consecutive courses, Self Care I and II, was chosen to test the hypothesis that training faculty members in item writing and implementing collaborative examination item review based on peer-reviewed guidelines would improve the performance statistics of examination items used in those courses and improve faculty members' self-rated confidence, success, and opinions regarding item-writing guidelines and review. The courses are offered in the P2 spring and P3 fall semesters and cover topics regarding nonprescription medicine and basic physical assessment techniques. Ninety percent of the course grades are based on multiple-choice examinations, with the remaining 10% based on practical activity

performance. Between the 2 courses, 12 clinical faculty members serve as instructors. Of the 12 faculty members invited to participate, 10 completed all components of the study. Item statistic comparisons were made between the pairs of courses taught in the control sequence – spring and fall of 2012 (Self Care 1 and 2, with no faculty training and no item review), and the intervention sequence – spring and fall of 2013 (Self Care 1 and 2, with faculty training and item review). Demographic characteristics of students enrolled in each course pair were collected and compared. Between 2012 and 2013, no content was changed in either course, but one instructor retired and was replaced in the 2013 sequence. An overview of the study design and summary of interventions can be found in Figure 1.

Before each semester began, the instructors were asked to complete a presurvey that gathered the subjective information stated above (Table 1). After completing the survey, investigators led a one-hour training program for instructors that covered an overview of the item-writing guidelines authored by Haladyna et al,¹ followed by a group discussion of key points to develop consensus

Figure 1. Schematic of Study Design and Implementation.

Table 1. Changes in Faculty Opinions and Confidence in Item-Writing^{a-d}

Item	Pre-survey (mean)	Post-survey (mean)	<i>p</i> value
How would you rate your confidence in writing effective multiple-choice test items? (0= not at all confident, 10=very confident)	6.0	8.1	0.002
How would you rate your success at writing multiple-choice test items? (0=very unsuccessful, 10=very successful)	6.4	7.9	<0.001
In your opinion, to what degree will peer review of examination items affect item quality? (0=very negatively, 10=very positively)	7.9	8.5	0.14
Do you plan to modify future multiple-choice items based on item-writing guidelines? (1=Yes, 2=No)	9 Y	10 Y	
In your opinion, to what degree will voluntary application of item-writing guidelines affect item quality? (0=very negatively, 10=very positively)	1 N	0 N	1.00
To what degree do you feel confident that you can properly evaluate your and your colleagues' test questions? (0=very unconfident, 1=very confident)	7.9	8.4	0.24
To what degree do you feel confident that you could implement a formal examination item evaluation process as a coordinator of another course? (0=very unconfident, 1=very confident)	6.7	8.4	0.005
	5.5	7.1	0.008

^aIf the same faculty members participated in both Self-Care 1 and 2, their responses from Self Care 2 were eliminated before analysis

^bIf faculty members responded to one questionnaire but not the other, their data were eliminated

^cIf faculty members skipped an item on one survey but not the other, their data for that item were pulled from the analysis of that item

^dScaled responses were compared with paired *t* tests; categorical responses were compared with McNemar test

among the participating faculty members and uniformity among the group's examination items. Examples of points discussed included routine use of items with stem negation; formats such as, "all of the above," "none of the above," "select all that apply," or type K questions (ie, questions with multiple-combination answer choices); and the ideal number of distractors for each item. The group decided to avoid all of these formats in the intervention year because of the documented potential negative effects on item statistics, with the exception of using "select all that apply" where appropriate in place of type K questions.¹ Items with 4 options (correct answer and 3 distractors) were chosen to be the standard, with instructor discretion to use 3-option items to avoid writing poor distractors to simply meet that standard.

To allow time for the group to review items before each of the courses' 3 examinations, course coordinators requested that items be submitted 2 weeks prior to the examination. Items were formatted into a draft examination and sent to the group 2-3 days before the pre-examination meeting. Item authors were not identified, but this information was available to participants by noting which instructors covered specific content as outlined in the course syllabi. All participating

instructors reviewed the items, attended the roughly one-hour meetings, and offered critiques and suggestions. Changes were made based on group discussion and majority opinion and were recorded and tabulated for later review as a way for all instructors to see the most common mistakes made by the group.

After the 2013 sequence of Self Care 1 and 2 was complete, instructors were asked to complete a postsurvey. The surveys were administered in both Self Care 1 and 2, so instructors of both courses only completed the survey instruments in Self Care 1 as an attempt to isolate the subjective effects of guideline training and item review in relation to these faculty members' more accurate baseline. This project was approved by the University of Louisiana at Monroe Institutional Review Board.

For the surveys, continuous data were compared using 2-tailed *t* tests, and categorical data were compared using the McNemar test. Student performance data was collected and analyzed by comparing mean examination scores on all examinations in the control sequence vs the intervention sequence. Mean item difficulties (*p*=percent of students answering correctly) and discrimination values (*rpb*=point biserial correlations) were calculated for each course by the examination administration software, ExamSoft (ExamSoft Worldwide, Inc., Dallas,

TX), and compared to corresponding data from the 2012 offerings of Self Care 1 and 2. Because a substantial amount of literature suggesting that application of item-writing guidelines increases item discrimination, one-tailed independent *t* tests were used to compare these data.

Kuder-Richardson 20 (KR20) values calculated by ExamSoft were compared between the control and intervention sequences with independent *t* tests. All items from both sequences also were assigned classifications based on difficulty and discrimination.¹¹ See Figure 2 for a graphical representation of these classifications. Level 1 items (p of 45-75% and rpb>0.2) had the best statistics because of their reasonable difficulty and relatively high discriminatory ability, and we recommended using items in this range if possible. Level 2 items (p of 76-90% and rpb>0.15) were considered relatively easy because of their difficulty range while still providing adequate discrimination, and we recommended using them sparingly. Level 3 items (p of 25-44% and rpb>0.1) were difficult with lower discrimination, and we recommended they should be reviewed and rewritten, if possible. Level 4 items (p of <25% or >90% and any rpb) were considered too difficult or too easy, and we recommended they be avoided unless associated content was essential. Differences in level classification were compared with 2-tailed independent *t* tests. *P* values of <0.05 were defined as significant for all tests and analyses were performed with Statistix 9.0 (Analytical Software, Tallahassee, FL).

Figure 2. Classification Guide of Examination Items by Difficulty and Discrimination.¹¹

RESULTS

Demographic characteristics of students enrolled in each course sequence (control vs intervention) are provided in Table 2. Ten of 12 instructors (83%) completed both surveys. Survey questions and responses are summarized in Table 1. Six participating faculty members reported they had been teaching in a professional pharmacy curriculum for ≤5 years, and 4, for 6 to 10 years. Seventy percent reported previous training in item writing, with faculty training programs (n=8) and credentialing board training (n=3) reported as the most common experiences. When asked which of the following factors affected their sense of success in item writing, faculty members responded as follows: item statistics (n=9), previous training in item writing (n=4), and student challenges to examination items (n=3). At baseline, only 5 faculty members (50%) had ever participated in item peer-review, 4 reported using peer-review “half of the time,” and one, “a minority of the time.” Similarly, only 5 faculty members at baseline had ever modified examination items based on item-writing guidelines.

Mean student scores (standard deviation, SD) on examinations in the control sequence vs intervention sequence were 88.3 (4.5) and 85.6 (6.0), *p*<0.001, respectively. There was no significant difference between the control and intervention items with regard to mean difficulty (SD) [86.3 (16.6%) and 84.4 (16.6%), *p*=0.097] or discrimination [0.23 (0.153) and 0.25 (0.149), *p*=0.05], respectively. Mean KR20s (SD) between the control and intervention sequences were 0.11 (0.49) and 0.30 (0.49), respectively, *p*=0.5. However, the distribution of item classifications did improve (Table 3). Scatter plots of item distributions into levels can be found in Figures 3 and 4. Higher numbers of questions fell into the more desirable categories (levels 1 and 2), and fewer into the less desirable (levels 3 and 4). Comparison of the mean number of items per test in each level between the control and intervention sequences is summarized in Table 3. Of the reviewed items, 110 (41.7%) were

Table 2. Demographics of Student Samples Each Year

Characteristic ^a	2012 (Control)	2013 (Intervention)	<i>p</i> value
Gender ^b , n			
Male	28	35	0.87
Female	44	55	
Math/science GPA	3.40	3.30	0.11
Prepharmacy GPA	3.52	3.44	0.10

^aCalculated with 2-tailed independent *t* tests, unless noted otherwise

^bCalculated with chi-square test

GPA=grade point average

Table 3. Classification Assignments of Examination Items by Group^{a,b}

Item Class	Control		Intervention		p value
	n (%)	Avg/Examination (SD)	n (%)	Avg/Examination (SD)	
Level 1	31 (11.4)	5.17 (2.5)	52 (19.7)	8.7 (5.3)	0.17
Level 2	70 (25.7)	11.67 (4.1)	76 (28.8)	12.7 (3.9)	0.60
Level 3	7 (2.6)	1.17 (0.8)	3 (1.1)	0.5 (0.8)	0.17
Level 4	142 (52.2)	23.67 (5.6)	122 (46.2)	20.33 (6.9)	0.37
Uncategorizable	22 (8.1)	3.50 (2.1)	11 (4.2)	1.83 (1.7)	0.17
Total	272		264		

^an (%) represents the number (percent) of items in the 6 examinations in each sequence that were categorized into each level

^bComparisons were made between the average number of items per examination in each course sequence that were categorized into each level; independent *t* tests were used

modified based on one or more of the item-writing guidelines. The most common change was to include the central idea of the question in the stem rather than the choices. Other modifications are in Table 4. Comparisons of frequency of undesirable item formats in the final versions of the tests between the control and intervention sequences are in Table 5.

DISCUSSION

A primary goal of this project was to improve the performance statistics of examination items used in the self-care modules. Although no significant differences were found in item difficulty or in item distribution among the 4 item classes, there were trends of positive change in item classifications (Table 3). The small number of examinations (6 per course sequence) likely limited our ability to find a true difference. An increase in level

1 items and a decrease in level 4 items suggest that a positive effect was made by the intervention. Level 1 items had the best item statistics and were the most desirable questions to use on examinations, while level 4 items were either too difficult or too easy. Neither level 4 extremes should be used unless content in the question is essential to test and cannot be rewritten in a more effective way. Several items in each group (22 in control sequence and 11 in intervention) were not categorizable because they did not meet both criteria for any level of classification. This number decreased with the review process, but it is not clear whether the presence of uncategorizable questions affects student or item performance. Student performance was slightly lower in the intervention sequence as indicated by the approximate 3-point decrease in mean examination scores. This was expected because of the decrease in the number of level 4 items, specifically

Figure 3. Scatter Plot of Examination Items from Control Sequence.

Figure 4. Scatter Plot of Examination Items from Intervention Sequence.

Table 4. Item-writing Guidelines Informing Editing Decisions by Frequency¹

Guideline	n (% total changes)
Include the central idea in the stem instead of the choices.	37 (33.6)
Use correct grammar, punctuation, capitalization, and spelling.	17 (15.5)
Minimize the amount of reading in each stem.	13 (11.8)
Use the question, completion, and best-answer versions of the conventional multiple choice (MC), the alternate choice, true-false, multiple true-false, matching, and the context-dependent item and item set formats, but avoid the complex MC (Type K) format.	10 (9.1)
Keep choices independent; do not overlap choices.	7 (6.4)
Word the stem positively, avoid negatives such as NOT or EXCEPT. Use the negative words cautiously and ensure the word appears capitalized and boldface.	7 (6.4)
Avoid window dressing.	6 (5.5)
Avoid "all-of-the-above."	3 (2.7)
Place choices in logical or numerical order.	2 (1.8)
Develop as many effective choices as you can, but research suggests 3 is adequate.	2 (1.8)
Ensure directions in the stem are very clear.	2 (1.8)
Avoid giving clues to the right answer, such as grammatical inconsistencies that cue the test-taker to the correct choice.	1 (0.9)
Keep the length of the choices about equal.	1 (0.9)
Keep the choices homogenous in content and grammatical structure.	1 (0.9)

extremely easy items. Although comparisons of test reliability with the KR20 did not demonstrate a significant difference, the mean values between sequences appeared to improve, and the small sample size of 6 examinations per sequence likely limited the power of this comparison.

Another goal of this project was to improve the involved faculty members' self-rated confidence, success, and opinions regarding item-writing guidelines and review. Comparison of presurvey and postsurvey responses demonstrated that faculty members felt more confident and successful at writing effective multiple-choice test items after the guideline review, discussions, and application of principles in formal item-review meetings. The analysis of the control vs intervention items did not demonstrate a significant difference in difficulty. Given that new items were written for each year rather than simply editing previously used items to conform to guidelines, there were differences between questions in each course sequence other than those that were purely guideline-related. The fact that the number of items with more

desirable statistics appeared to increase, combined with the postsurvey indication of increased perceived item-writing success, suggests that the intervention may have contributed to quantitative and qualitative improvements in the item-writing process.

Another objective of this project was to encourage the use of peer-review of examination items in other courses in the curriculum. Responses to the last 2 items on the surveys demonstrated that faculty members felt more confident in evaluating examination questions after the intervention. They also felt more certain of their ability to implement a similar process in another course. The participants reported that implementation of this process in other courses would be a major benefit to the program. The program consists of an integrated, modular curriculum in which the majority of modules is team-taught and involves several faculty members from the clinical and basic sciences departments.

Coordinating efforts in writing effective examination items may benefit the quality of course assessments. By attempting to limit variability in the examination items (ie, variability introduced by deviating from guidelines), examinations may more accurately measure students' true knowledge. This process would provide students with better examination-derived formative feedback and allow the program to more accurately assess curricular effectiveness. The process also would create a venue in which a collaborative group could discuss and adjust the content validity of examination items and ensure item alignment with course objectives, though no such adjustments were identified as necessary throughout this project.

Table 5. Comparison of Undesirable Item Format Frequency

Item Type	2012 (Control)	2013 (Intervention)
All of the above	15	1
None of the above	2	0
Select all that apply	0	9
Stem negation	10	12
Type K*	9	3

*Questions with multiple-combination answer choices

Table 4 lists the common errors identified during the peer-review process. Errors that authors initially anticipated occurring at the highest frequency (eg, stem negation, using “all of the above” as an answer, and the use of K-type questions) were not frequent offenders. This suggests that the initial review of item-writing guidelines and the discussion that followed were effective at changing some participants’ typical patterns of writing. Table 5 summarizes what may account for the poorer than predicted performance in the intervention sequence. In several cases, more difficult formats were used more frequently in 2013. A decrease in the number of “all of the above” options, which often make examination items easier, and an increase in the “select all that apply” format (which, while considered difficult, was chosen as a replacement for type K items because of its better discriminatory capability), may have contributed to this poorer performance.

Although most remaining modifications were made to remove unnecessary difficulty, these changes may have added some degree of difficulty. Use of stem negation did increase by 2 items in the intervention sequence, but this format was used more intentionally in 2013 (eg, to identify pertinent negatives) and was consistently highlighted with bolded and capitalized text, whereas this highlighting was not consistently used in 2012.

One of the intangible benefits of this process is the opportunity for junior faculty members to receive instant feedback on their examination items when, under ordinary circumstances, they may not seek help and advice. This process also may correct errors in items that would normally have triggered a formal challenge by students after the examination. The review method also provided a venue for communication between all faculty members involved in the courses, giving everyone a better understanding of what each instructor was teaching and allowing a view of questions from others’ perspectives.

Communication among this group was open and free from personal defensiveness, which may have been partly because of the voluntary nature of the process and because of the explanation of the review method and purpose before the project began. The instructors in the course also were in the same department, which potentially contributed to the group’s cohesiveness.

One limitation of this project is related to the homogeneity of the 2 course sequences involved (ie, there were likely differences resulting from the change of instructor that occurred between 2012 and 2013). Although the 2012 instructor’s lecture materials were made available to the instructor who replaced him, the 44 items covering the related content were all rewritten and may have

differed. Survey questions focused on self-reported rather than measurable knowledge, resulting in an inability to detect true pre/post differences in this domain. Consequently, the participants’ self-perceptions of confidence and success cannot be correlated to a demonstrated knowledge increase. Improvement in item classification may have been a result of the peer-review process itself, participant knowledge increase, or a combination of the two, but was not directly measured or assessed in this study.

Another limitation may be related to the ease with which instructors were recruited. In similar projects, there may be a concern about instructor willingness to participate because of the potential for perceived examination criticism during the review process. In our case, examination items were not identified by author. But, because all members of the group had access to the syllabi and course schedules, authors could be identified according to the material they taught in the course. Regardless, our review sessions were open, direct, and candid, yet free from any displays of ego or posturing. The success of the review method in terms of the cooperation of faculty members may have been related to the number of instructors involved, the fact that all were members of the same department, and the straightforward nature of the course content. These factors also may limit the generalizability of our results to other courses, especially those taught collaboratively by members of different departments. Finally, analyses of objective data, such as KR20s and item classification distributions, were limited by the small sample size of examinations involved in the courses.

We plan to implement similar interventions in other courses with multidisciplinary teaching teams. More research should be completed to determine whether peer-based item review initiatives could positively affect examinations in team-taught courses.

CONCLUSION

A systematic, collaborative examination review and improvement process can be successfully completed in a team-taught course and may improve overall examination quality. Although the results of this project did not demonstrate significantly improved objective criteria such as item statistics and student performance, the data presented a compelling case for further research on the implementation of such review processes in other courses at other schools. The interventions of training with peer review and editing significantly improved faculty member confidence and self-perceived success in item writing, while also increasing positive perceptions of item-writing

guidelines, the peer-review process, and the number of instructors planning to implement this type of review in other courses.

REFERENCES

1. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ.* 2002;15(3):309-334.
2. Case SM, Swanson DB. 2003. Constructing written test questions for the basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners. *Radiographics.* 2006;26(2):543-551.
3. Downing SM, Haladyna TM, eds. Handbook of Test Development. Mahwah, NJ: Erlbaum; 2006.
4. Collins J. Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics.* 2006;26(2):543-551.
5. Burton SJ, Sudweeks RR, Merrill PF, Wood B. How to prepare better multiple-choice test items: guidelines for university faculty. Brigham Young University Testing Services and the Department of Instructional Science. 1991.
6. Accreditation Council for Pharmacy Education. Accreditation standards and guidelines for the professional program in pharmacy leading to the doctor of pharmacy degree. <https://www.acpe-accredit.org/standards/default.asp>. Accessed August 2014.
7. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. *Acad Med.* 2002;77(2):156-161.
8. Caldwell DJ, Pate AN. Effects of question formats on student and item performance. *Am J Pharm Ed.* 2013; 77(4):Article 71.
9. Pate A, Caldwell DJ. Effects of multiple-choice item-writing guideline utilization on item and student performance. *Curr Pharm Teach Learn.* 2014; 6(1):130-134.
10. Naeem N, Vleuten C, Alfari EA. Faculty development on item writing substantially improves item quality. *Adv in Health Sci Educ.* 2011;17(3):369-376.
11. Haladyna TM. *Developing and Validating Multiple-choice Test Items.* (3rd ed.) Mahwah, NJ: Lawrence Erlbaum Associates, 1999.