Assessing Pharmacy Student Knowledge on Multiple-Choice Examinations Using Partial-Credit Scoring of Combined-Response Multiple-Choice Items

Supakit Wongwiwatthananukit and Nicholas G. Popovich[1]

School of Pharmacy and Pharmacal Sciences, Purdue University, Heine Pharmacy Building, West Lafayette, IN47907-1330

Deborah E. Bennett

School of Education, Purdue University, West Lafayette, IN 47907-1440

This study developed, implemented, and evaluated the use of partial-credit scoring for the combined-response multiple-choice (CRMC) examination format utilized in an elective nonprescription drugs course. Behavioral objectives and CRMC questions were developed and classified according to the appropriate cognitive domain of *Bloom's Taxonomy of Educational Objectives*. External and internal content panels validated these objectives and reviewed developed CRMC questions prior to their use on three course examinations. Item response data for the partial credit scoring and the traditional dichotomous scoring method were then compared in terms of item difficulty, item discrimination, reliability, efficiency, cognitive domain being measured, and criterion-referenced grading assignments. The results demonstrated that the partial credit scoring method increased item difficulty and resulted in student outcomes which were more reflective of student knowledge compared to dichotomous scoring. Further, the partial credit scoring method increased the reliability and efficiency when compared to the dichotomous scoring method. The item difficulty difference attained by using the two methods of scoring was statistically different ($P < 0.05$) at different levels of the cognitive domain.

INTRODUCTION

In pharmaceutical and other health professional education, the most commonly used assessment instrument is the test consisting of multiple-choice (MC) items, particularly the single-answer (SA), four or five foil multiple-choice item and the combined-response multiple-choice (CRMC) item which is one type of complex multiple choice (Type K) item. CRMC items are combinations of three primary responses, and the combinations of the primary responses are referred to as secondary choices. The secondary choices which are specific for the choice of the three primary responses serve as options. A typical CRMC format is illustrated by the following example:

Which of the following laxative products is(are) appropriate to recommend to a man in otherwise good health who suffers from painful hemorrhoids?

1) Dulcolax Tablets
2) Fiberall Powder — Answer: D
3) Metamucil Packets

Response Format: A.1 B. 3 C. 1,2 D. 2,3 E. 1,2,3

This format, compared to the traditional single-answer, four or five option multiple-choice item is used to test student achievement within the pharmacy curriculum courses and is also utilized in the North America Pharmacist Licensure Examination (NAPLEX) to evaluate the minimal competency of the pharmacy graduate to enter the profession of pharmacy.

An advantage of the CRMC format is that it accommodates those occasions when more than one choice/response is correct. Indeed, within pharmacy practice, there are many occasions when more than one response is correct and/or appropriate. Oftentimes, too, in practice, there are also degrees of correctness (*i.e.*, most correct, somewhat correct) when responding to certain questions/inquiries. Some researchers recommend against using the complex multiple choice (Type K) format for several reasons(1-3). For example, complex multiple choice items are more difficult than single best answer multiple choice. This format requires more time to read and could affect the number of items used in a test (*i.e.*, sampling of content). This format may produce items with lower discrimination which could affect test score reliability. This format is difficult to construct and edit. Although there are several disadvantages of this format, it is important to note that the complex multiple choice format that these researchers criticized is not the one used in this study.

Historically, the scoring of multiple-choice examinations has been problematical. Dichotomous scoring consists of assigning a certain point value for a correct answer and zero points for an incorrect answer. It causes inefficiencies in the interpretation of test scores because the response to a MC item may come from different information levels (*i.e.*, complete information, partial information, misinformation, and total ignorance)(4-6). Complete information is described as the knowledge that allows that student to select the most correct response and most incorrect answer from among presented

[1]Corresponding author.

Received July 7, 1999, accepted November 4, 1999.

options in a MC item(5). Partial information is described as the information that students have which allows them to recognize some of the incorrect response options as clearly wrong while being unable to recognize all of the correct options(5-7). Partial information, in itself, is not incorrect information. However, it is incomplete knowledge, differing from misinformation in being only partially correct, rather than completely incorrect. If a student believes the correct response is wrong without guessing then the student has misinformation. Incorrect options may also be eliminated due to misinformation(5).

In test scoring of the students with partial information or misinformation, the ability to discriminate degrees of correctness or incorrectness has usually been ignored. Students are given no credit for rejecting the worst responses or choosing somewhat correct responses. This can be observed from the example CRMC question above. The correct answer is "D." Based upon conventional dichotomous scoring, students who select "D" would get full credit and all other responses would be valued at zero credit. But, students who selected "B" are also partially correct, even though they did not select the most correct response. This demonstrates how dichotomous scoring cannot reward students for less than full knowledge or mastery of the content area.

Guessing strategies and willingness to guess or omit an answer when a student does not know the answer also contribute to an item response(4). According to the dichotomous scoring system, "lucky and test-wise" students can produce an upward bias in scores as a result of guessing. For these reasons, it is important to treat misinformation, partial information, and guessing properly to provide test scores that give a better estimation of the student's ability level.

The basis for giving partial credit is to reward intermediate levels of performance between a correct answer and a wrong answer based on the degree of completeness of accuracy of that answer(5,8,9). Conceptually, it is a desirable feature for measurement efficiency because it represents the partial knowledge or skill proficiency of students. Past research efforts have been mostly devoted to considering how to award partial credit for MC examinations and the impact of partial credit scoring on test scores(2-7,9-19). Currently, many methods are used to correct the limitations of dichotomous scoring to provide test scores that give a better estimation of the student's ability level. Some scoring methods for awarding partial credit for partial knowledge (*e.g.*, multiple correct options, item response theories [IRTs], finite state theory) on multiple-choice tests seem to be a promising trend in the area of measurement study. Most of the scoring methods singly or jointly, however, provide no panacea or even consistent potential for worthwhile improvement of the psychometric properties of test scores. Although some methods have enhanced internal-consistency reliability, this result is devalued by one or more of the following such as reduced validity, increased time required for testing, increased resource usage, scoring complexity, difficulty of explaining the scoring to students and other users, difficulty of explaining the response mode and training students in its use, among others.

In multiple-answer (MA) multiple-choice items, which resemble CRMC, one study utilized multiple correct options scoring(18). Hsu, Moss, and Khampalikit used MA multiple-choice items and studied six different scoring formulas for both SA and MA items. These formulas varied in the assignment of partial credit and the correction for guessing. The authors found the method which accounted for partial knowledge yielded more reliable and valid scores. Items of the MA type were substantially more discriminating when partial credit was given. This implied that scoring with the MA format option was potentially a better choice.

Nevertheless, research on the effects of giving partial credit for CRMC items when compared to conventional dichotomous scoring with the same examination instrument in terms of item difficulty, item discrimination, cognitive domain being measured, reliability, and efficiency has not been established. In addition, the use of CRMC items on achievement tests designed for health professions is still continuing, although research in this student evaluation area in pharmaceutical education is non-existent. Thus, this study applied the concept of awarding partial credit and developed a partial credit scoring mechanism for CRMC items (*i.e.*, without penalty to an incorrect choice and guessing).

The purpose of this study was to utilize combined-response multiple-choice (CRMC) to accommodate and reward students with partial credit when responding to multiple-choice questions that had responses with varying degrees of completeness or correctness (*i.e.*, most correct, somewhat correct). It was intended to optimize the measurement of student knowledge without penalty for incorrect responses and guessing. Specific objectives of this study included:

1. Determining if a difference existed in item difficulty between dichotomous scoring and partial credit scoring methods for each of the three CRMC examinations and in the content areas of analgesics and laxatives.
2. Determining if a statistical difference existed in item discrimination between dichotomous scoring and partial credit scoring methods for each of the three CRMC examinations and in the content areas of analgesics and laxatives.
3. Determining the differences in the reliability coefficients for dichotomous scoring and partial credit scoring methods for the three CRMC examinations and in the content areas of analgesics and laxatives.
4. Comparing the efficiency of dichotomous scoring and partial credit scoring methods for the three CRMC examinations and in the content areas of analgesics and laxatives.
5. Determining if statistical differences existed in item difficulty and item discrimination for written objectives at selected levels within the cognitive domain of Bloom's Taxonomy of Educational Objectives (*i.e.*, knowledge, comprehension, application, analysis, evaluation) between dichotomous scoring and partial credit scoring methods in the content areas of analgesics and laxatives.
6. Determining if a difference existed in the effect of criterion-referenced grade assignment between dichotomous scoring and partial credit scoring methods for the three CRMC examinations.

METHODS

Definitions of Terms

For purposes of this study several terms are operationally defined as follows:

Item difficulty was defined as the mean item score (M_I) divided by the weight assigned to the item (W), Item difficulty = $(M_I)/(W)$, where M_I = Σtotal scores on the item being analyzed/ number of students(18,20). In this study, the assigned weight was two points because the two semester examinations

had 50 questions and one final examination had 75 questions. Item difficulty can be interpreted as the proportion of students who answer the item correctly.

Item discrimination was defined as the item-total correlation (Pearson product-moment correlation coefficients)(20-22). Its interpretation depends on the types of examination (*i.e.*, norm- or criterion-referenced tests). Normally, item discrimination determines whether an item discriminates between the students who do well on the test and students who do not.

Reliability was defined as the Cronbach's coefficient alpha(20,22). It compares performance on each item to performance on all other items. Coefficient alpha was chosen because it is the only reliability estimate for the test whose items are weight scoring (*i.e.*, polytomous scoring) and single-test administration.

Efficiency was defined as the coefficient of effective length ($CEL_{KK/LL}$) which is comparative reliabilities, $CEL_{KK/LL} = {}^{r}_{KK}(1-r_{LL})/r_{LL}(1-r_{KK})$, $CEL_{KK/LL}$ = the relative efficiency of scoring formula KK to scoring formula LL, r_{KK} and r_{LL} = the Cronbach's coefficient alpha of scoring methods KK and LL(5,18). It is a measure of how much of a test administered in one scoring method would have to be lengthened to produce the same reliability as that same test administered by another scoring method.

Methods of CRMC test scoring were defined as dichotomous scoring and partial credit scoring.

- Dichotomous scoring (S1).
  S1 = W(2) if all choices correct. S1= 0, if omitted or any choice incorrect.
  In this study, the full point value or assigned weight (W) for each item is two points.
- Partial credit scoring (S2) is defined as the first best, the second best, and the third best answers to a question. The assigned weight for the first best answer is two points, the assigned weight for the second best is one point, and the assigned weight for the third best is 0.5 point. In the fol lowing examples, the number in parentheses for each scoring method indicates the point value assigned for that particular response. Within this format, when:

the correct answer was A.

Select	A. 1	B. 3	C. 1, 2	D. 2, 3	E. 1,2, 3
S1	W(2)	0	0	0	0
S2	W(2)	0	0	0	0

the correct answer was B.

Select	A. 1	B. 3	C. 1,2	D. 2, 3	E. 1,2,3
S1	0	W(2)	0	0	0
S2	0	W(2)	0	0	0

the correct answer was C.

Select	A. 1	B. 3	C. 1,2	D. 2, 3	E. 1,2,3
S1	0	0	W(2)	0	0
S2	W/2(1)	0	W(2)	0	0

the correct answer was D.

Select	A. 1	B. 3	C. 1,2	D. 2, 3	E. 1,2,3
S1	0	0	0	W(2)	0
S2	0	W/2(1)	0	W(2)	0

the correct answer was E.

Select	A. 1	B. 3	C. 1,2	D. 2, 3	E. 1,2,3
S1	0	0	0	0	W(2)
S2	W/4(0.5)	W/4(0.5)	W/2(1)	W/2(1)	W(2)

Content areas taught and tested in a Nonprescription Drugs course (PHPR 470), Spring semester, 1998, were the major OTC product categories. There were two hourly semester examinations scheduled and one final examination. The first examination (50 items, 100 points) tested three topics (*i.e.*, nonprescription internal analgesics, home diagnostic products, diabetic accessories/products). The second examination (50 items, 100 points) tested five topics (*i.e.*, smoking cessation products, foot care products, corn, callus, and wart products, pediculicide/scabicide products, topical antiinfective products). The final examination (75 items, 150 points) tested four content areas (*i.e.*, laxative products, hemorrhoidal products, antidiarrheal products, antacids/OTC H2-antagonist products).

Within the course, there were also five written case scenario homework assignments for the students to complete. The content areas of these were cold and cough products, nonprescription internal analgesics, ophthalmic products, OTC contraceptive products, and laxative products. The point value for each homework assignment was 60 points. Of those 60 points, 15 points were devoted to writing attributes of the completed assignment. These were as follows: three points each for focus (clearly stated main idea; clear introduction), organization (clear structure; effective transitions between paragraphs; each paragraph containing a single major idea; beginning and end of paper relate to the whole), readability (sentences within the paragraphs relate to each other and to the paragraph topic; writing flows), development and quality of thinking (supported ideas with relevant and specific facts, reasons, and evidence; evidences in-depth thought), and mechanics (fulfills all assignment requirements and follows all directions, no mechanical errors [subject/verb agreement, pronoun reference agreement, spelling, punctuation]). The remaining 45 points were allocated on the basis of how well the student solved the scenario.

Levels of cognitive domain were defined as the six levels of Bloom's taxonomy of the cognitive domain: knowledge, comprehension, application, analysis, synthesis, evaluation(21,23-25).

Course Enrollment

One hundred fifty pharmacy students enrolled in the three credit hour elective Nonprescription Drugs (PHPR 470) course during the second professional year (spring 1998) in the baccalaureate program or in the doctor of pharmacy program at the School of Pharmacy and Pharmacal Sciences, Purdue University, and were the subjects in this study. The entire class attended three 50 minute traditional presentation/lectures per week over the course of the semester. Prior to the start of the course, each student purchased a course manual which contained lecture outlines and other course aids (*i.e.*, professional articles that complemented the instruction) utilized by some instructors during his/her class periods. The students also were encouraged to read these materials prior to each lecture.

Developing Test Pool Items and Objectives

The objectives for the three examinations were developed using the above-named content areas by the investigator. Each objective was classified according to the appropriate cognitive

domain of Bloom's Taxonomy of Educational Objectives(21,23-25). The test objectives were revised and organized by the investigator after input from selected department faculty who instructed the various content areas of the course. These test objectives were used to develop the set of test pool items (*i.e.*, first, second, third set) for the three examinations. Several objectives in the knowledge level were deemed necessary to evaluate recognition or recall of important information about pathophysiology, disease state management, and appropriate treatment recommendation. Objectives for higher-order thought processes, *e.g.*, comprehension, application, analysis, and evaluation, were also necessary for testing the ability of students to use, to extend, to apply information to a new situation, and to justify the rationale for the use of a particular treatment.

Objectives requiring synthesis thought processes were not used in the test item construction because of the limitations in measurement associated with the multiple-choice item format(21,24,26). However, synthesis level was measured through the homework assignments. Synthesis requires the student to put together important elements (*e.g.*, concepts, knowledge) in an integrative fashion to solve the problem(23,25). It is a process of continual development for the student in that the student learns to develop a pattern of thinking that will allow for the solution of the problem.

Preparing Test Pool Items and Examinations

The CRMC format of multiple-choice items was chosen to prepare of the three sets of test pool items. Not all of the objectives for each content area were used in creating the test pool items that constituted the three examinations used in the study. Only those statements and cognitive tasks that were determined to be the most fundamental requisites for this course and clinical practice were selected for inclusion. All of the test pool items were developed as a result of the investigator's own pharmacy practice experience and from prior examinations used in this course. The course manual, standard pharmaceutical reference books, and current professional journals served as resources to validate the information and to develop the combinations of three primary responses for CRMC format. Each set of test pool items was forwarded to the external content review panel member to judge content validity. Based on the criteria for content validity, each set of test pool items was edited and revised to improve the test item quality.

The external content review panel was constituted by six educators and/or practitioners who have taught or currently instruct a nonprescription drugs or a self care course at a school/college of pharmacy in the U.S. Each external content review panel member was asked beforehand to indicate his/her familiarity and "comfort" level of understanding the specific area that each question set examined. Because these members might not be familiar with each of the content areas that would be examined, the course instructors responsible for the content areas (*e.g.*, internal analgesics, diabetes accessories and devices, laxatives) in the nonprescription drugs course also reviewed his/her specific content area and served as an internal control. Internal control panel faculty were asked to review examination items germane to their content area prior to the administration of the examination. They were directed, however, not to consciously teach to the questions. Typically, developed examination questions from a content area exceeded the number actually used on an examination. There were three CRMC examinations developed in this study. The test items for each CRMC examination were selected from each set of test pool items (*i.e.*, first, second, third set) according to the following factors:

- The investigator's own pharmacy practice experience, which helped determine the most fundamental information for this course.
- The item review and agreement of the instructors who taught and served as the internal control for their specific content areas in this course.
- Achieving an adequate and representative sample of both the subject matter and the particular skills in the test blueprints that should have been learned as a result of instruction.

All test pool items forwarded to the external review panel members were prepared on a content review form developed by Pray(27) and modified by the investigator (Appendix). The form required the reviewers to rate each item according to its degree of representativeness of the knowledge or skills required to achieve the corresponding the test pool items' objectives. If in the opinion of the reviewer, the item was deemed to be more appropriate for measuring an objective other than the one originally designated, the form requested that a more suitable objective be indicated by the reviewer if one existed. It also provided space for the reviewers to comment on the correctness of the keyed response, and provide the second best, third best, fourth best, and fifth best responses. This was intended to assist the investigator when assigning partial credit. It was decided prior to beginning the review that the investigator would consider the reviewers' assigned partial credit responses/suggestions and incorporate them as one method of partial credit scoring used in the study. However, the majority of the reviewers' responses (*i.e.*, more than 80 percent of reviewers) did not provide the second best, third best, forth best, and fifth best more than the keyed responses originally assigned, therefore, there were no consistent answers other than the keyed response. The reason may come from the unclear instruction that explained the procedures to assign partial credit. This study would use only the partial credit scoring method that was developed by the investigators to compare with dichotomous scoring method. Lastly, if the reviewers felt the item could be revised and reworded, the form also provided space to make specific suggestions for improving the effectiveness of individual item stems and three primary responses of CRMC format.

A nonstatistical analysis technique similar to that used by Pray was applied to the responses of the external content review panel members to the degree of representativeness and the keyed response of each item(27,28). The 80 percent decision rule was not achieved by three items on the first and second set of pool items sent to the external content review panel members. However, the majority of the panel members (*i.e.*, 5/6 of panel members) agreed on an alternative objective for each of these three items. There were no revisions or corrections to the keyed response of the first, second, and third set of pool items sent to the external content review panel members based on a simple majority rule. Minor suggestions made by the reviewers to improve the clarity of item stems and to increase the plausibility of individual three primary responses or item distractors, *i.e.*, wording changes were also incorporated. It was also decided prior to beginning the review that, if it appeared that major revision of item stems or distractors were

Table I. Means and standard deviations (SD) of item difficulty and discrimination for the performance of 150 pharmacy students on the first, second, and third CRMC examinations using dichotomous scoring (S1) and partial credit scoring (S2)

	Mean (SD)			
	Item difficulty		Item discrimination	
Examination and content area	S1	S2	S1	S2
Examination 1	0.7256(0.1918)	0.7725 (0.1847)	0.2079 (0.0989)	0.2074(0.1153)
Examination 2	0.7451 (0.1918)	0.7957 (0.1855)	0.2189(0.1031)	0.2183 (0.1234)
Examination 3	0.7215 (0.2017)	0.7748 (0.1717)	0.1954(0.1195)	0.1999 (0.1289)
Analgesics	0.8200(0.1350)	0.8720(0.1097)	0.1936(0.0693)	0.1989(0.0839)
Laxatives	0.7537(0.1408)	0.8120(0.1321)	0.1644(0.1559)	0.1698(0.1610)

necessary to make an item acceptable to panel members, that particular item would be omitted from the candidate item pool for developing examinations used in this research. However, none of 299 generated items reviewed had to be omitted based on this criterion. There were some minor suggestions made by the reviewers to improve the clarity of item stems and to increase the plausibility of individual three primary responses or item distractors/foils (*i.e.*, wording changes, grammatical corrections). All of these suggestions were incorporated by the investigator to improve the quality of the items. The students were informed of and provided the behavioral objectives for each content area that was tested on each examination. These objectives were the same as objectives used for the first, second, and third set of pool items sent to the external content review panel members.

Design and Procedure

This study was a descriptive research project which involved collecting data to test hypotheses or to answer questions or objectives(29). Students were administered the three CRMC examinations as two regular semester (*i.e.*, two, hourly mid-term) examinations and one final examination (*i.e.*, two-hour examination) in preselected content areas of the nonprescription drugs course. Transoptic score sheets were used by students to record their answers. These answer sheets were graded at the Center for Instructional Services (CIS), Purdue University, and utilized a program for scoring created by the investigators. Data on students' test scores for three examinations and in the content areas of analgesics and laxatives were collected and coded according to the two methods of scoring (S1, S2). The two content areas (*i.e.*, analgesics, laxatives) were arbitrarily selected by the investigators to examine if differences existed in item difficulty and item discrimination when measuring the written objectives at different levels of the cognitive domain between dichotomous scoring and partial credit scoring methods.

Data Analysis

Data were managed and analyzed using Microsoft Excel(30) and Statistical Analysis System (SAS) software systems(31). Means and standard deviations of item difficulty and item discrimination for each of three CRMC examinations were performed for objective one and two, respectively. It is important to mention that it was unnecessary to test statistical inference *(i.e.,* one-tail (-test) for the difference in item difficulty between dichotomous scoring and partial credit scoring methods because the partial credit scoring mechanism developed for CRMC examination in this study always gave higher item difficulty values compared to that of dichotomous scoring. The paired-samples t test (two-tailed test) was used to evaluate the statistical difference of methods of scoring (*i.e.*, S1, S2) on item discrimination. Cronbach's Coefficient Alpha was used to compute reliabilities for objective three. Coefficient of Effective Lengths (CEL) was computed for objective four. Multivariate Analysis of Variance (MANOVA) was used to evaluate objective five. A criterion-referenced (CR) grading assignment where A = 90-100 percent, B = 80-89 percent, C = 70-79 percent, D = 65-69 percent, and F = 64 percent and below, was used to assign grades based upon students' scores on each of the three CRMC examinations. The percentage of students assigned to CR grade categories by each scoring method (S1, S2) was calculated for each examination. An a priori level of significance of 0.05 was used for all statistical tests.

RESULTS AND DISCUSSION

Objective One. Table I shows the results of means and standard deviation of item difficulty for the performance of the students on the first, second, third CRMC examination and in the content areas of analgesics and laxatives using dichotomous scoring and partial credit scoring. Assigning partial credit clearly made the item difficulty values higher or more reflective of student knowledge than dichotomous scoring because the partial credit scoring method developed in this study was structured to give some credit for the students with partial information or reward students for less than full knowledge or mastery of the content area. Clear evidence for the existence of partial information mediating responses to CRMC items was demonstrated.

Objective Two. Table I shows the results of means and standard deviation of item discrimination for the performance of the students on the first, second, third CRMC examination and in the content areas of analgesics and laxatives using dichotomous scoring and partial credit scoring. Results from paired-samples *t*-tests revealed no significant difference between mean item discrimination using dichotomous scoring and partial credit scoring on the first [t(49) = 0.0858; P = 0.9319], second [t(49) = 0.1041; P = 0.9175], third [t(74) = 1.1570; P = 0.2510] examinations; analgesics [t(10) = -0.9215; P = 0.3785]; and laxatives [t(19) = -0.5489; P = 0.5894]. These results indicated that using dichotomous scoring and partial credit scoring provided the same information usable in distinguishing between low- and high-ability students in examinations and in the content areas of analgesics and laxatives.

Objective Three. Figure 1 illustrates the column chart of Cronbach alpha reliability coefficient for the first, second, third

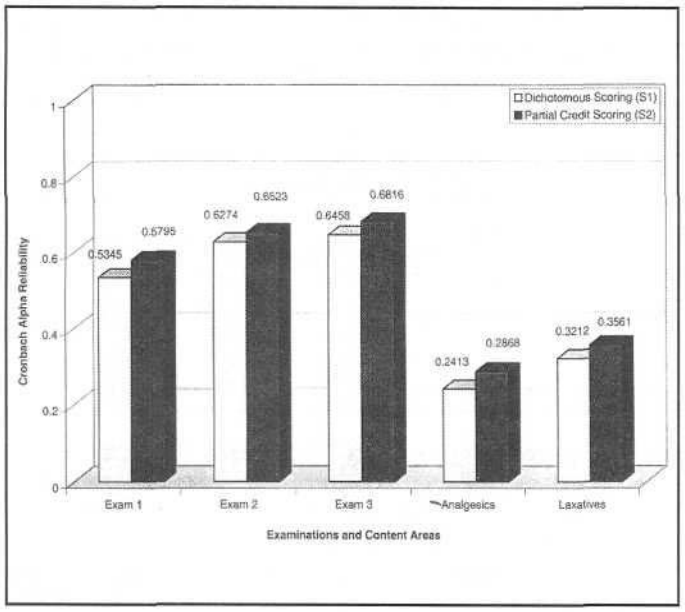

Fig. 1. Cronbach alpha reliability coefficient for the first, second, third CRMC examinations and in the content areas of analgesics and laxatives using dichotomous scoring (SI) and partial credit scoring (S2).

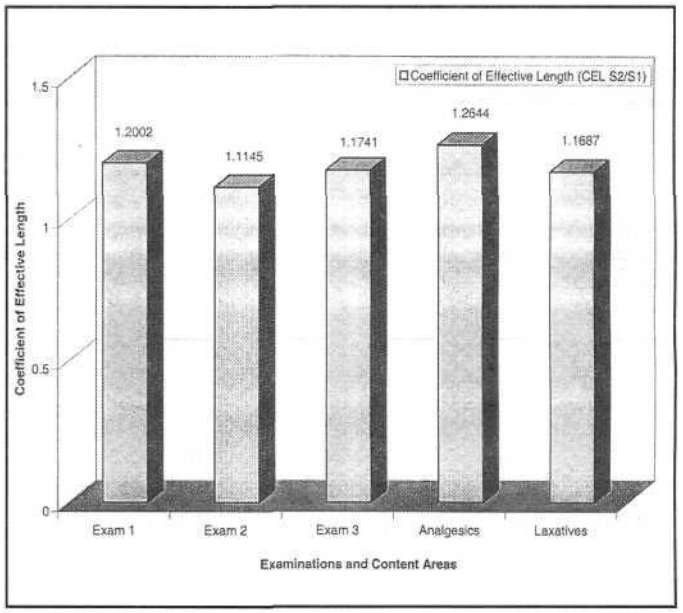

Fig. 2. Coefficient of effective length (CEL S2/S1) between partial credit scoring (S2) and dichotomous scoring (SI) for the first, second, third CRMC examinations and in the content areas of analgesics and laxatives.

CRMC examinations and in the content areas of analgesics and laxatives using dichotomous scoring and partial credit scoring. The partial credit scoring method increased the reliability in all three examinations when compared to dichotomous scoring method. Thus, partial credit scoring reflected the proportion of the differences among students in performance on the examination (test score variance) which is attributable to differences among students in stable knowledge or characteristics (true score variance) more so than that of dichotomous scoring. However, as is observed from the calculated value of the Cronbach's coefficient alpha for the three examinations, all were low (*i.e.*, ranging from 0.5345 to 0.6816). This may have been due to the difficulty of the three tests (*i.e.*, item difficulty values ranged between 0.7215 and 0.8720 for the three examinations). These values suggested that the tests were relatively easy(25-27). Reliability tends to decrease as tests become too easy or too difficult(25-27). As tests become very easy (nearly everyone answers all of the items correctly) or very difficult (nearly everyone answers all of the items incorrectly), score distributions become homogeneous. When distributions are homogeneous, significant shifting of ranks and a lowering of the correlation coefficient will occur.

It is important to note that the CRMC examinations used in this study were intended to be criterion-referenced (CR) tests. However, the Cronbach alpha reliability coefficient used in calculating the reliability in this study was developed for use in estimating the consistency of scores (*i.e.*, dichotomous scores, partial credit scores) obtained from tests designed to be used in making norm-referenced (NR) interpretations(32-35). Thus, the reliability coefficients should be cautiously viewed when using CR tests. The goal of CR tests used to assess mastery learning is to differentiate between mastery and nonmastery, whereas NR tests are intended to maximize differences among students. Under conditions where most students obtain mastery, variability among students is restricted (*i.e.*, the scores are not as spread out) and tends to make the usual estimates of reliability low or to underestimate the tests' true reliability. For this reason, some researchers recommend using alternate methods (*i.e.*, reliability of classification decisions, reliability of CR test scores, reliability of domain score estimates) to assess the reliability of CR tests(32-35). However, this study did not employ alternate methods because they are still rarely reported in test evaluation and controversial in selection whether which methods provide the best reliability estimates(32,33). In addition, most of the methods require two parallel test administration(32). Some methods yield biased estimates for short tests, need larger sample sizes, and need complex computation beyond the capability of available computer program(32). Alternate methods also require selecting only one cut-off score for mastery/nonmastery and this study employed five cut-off scores based on a CR grading assignment for each of the three examination regardless of whether the student did or did not master the content areas in the course.

Objective Four. Coefficient of effective length (CEL) between partial credit scoring and dichotomous scoring for the first, second, and third CRMC examinations and in the content areas of analgesics and laxatives are presented in Figure 2. The CELs were all larger than unity and ranged from 1.1145 to 1.2644. The CEL could be interpreted as the length of the examination under the partial credit scoring method in units of the length of the examination under the dichotomous scoring method. A CEL = 1.00 indicates that the examination under the partial credit scoring method had effectively the same length as when under the dichotomous scoring method. A CEL > 1.00 indicates that the partial credit scoring method was more efficient at determining student knowledge compared to the dichotomous scoring method. Another way of looking at the index is that the CEL is a measure of how much of an examination scored under the dichotomous scoring method would have to be lengthened in order to produce the same reliability by the partial credit scoring method. For example, in the first examination (50 questions), it would require of an examination of 60 questions (*i.e.*, [1.2002][50]) using dichotomous scoring to have comparable reliability as an examination of 50 questions using partial credit scoring.

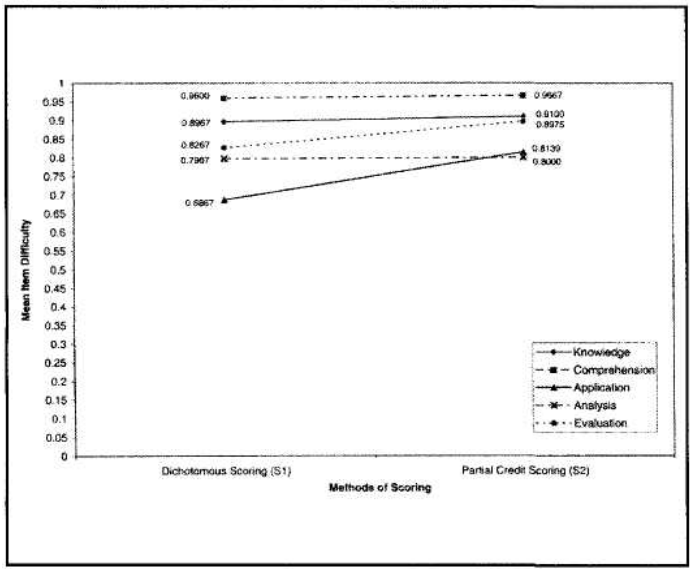

Fig. 3. The interaction effect of scoring methods (S1, S2) and cognitive levels on item difficulty in the content area of analgesics on the first CRMC examination.

Objective Five. Analgesic Content Area. In the first examination, eleven out of fifty test items were used to test students' knowledge in the analgesic content area (*i.e.*, two items in the knowledge level, two items in the comprehension level, three items in the application level, two items in the analysis level, two items in the evaluation level). For item difficulty, a significant interaction was found between methods of scoring (S1, S2) and cognitive levels on the mean item difficulty, Wilks' lambda = 0.1438, F(4, 6) = 8.9310; $P = 0.0106$. The graph in Figure 3 depicts this interaction. When changing the methods of scoring from dichotomous scoring method to the partial credit scoring method, the mean item difficulties of application and evaluation levels were slightly increased in the same direction, but for those with comprehension, knowledge, and analysis levels, there was much less of an increase or no change. As a follow-up to analyses that produced a significant interaction effect, Student-Newman-Keuls post hoc tests (type I experimentwise error rate = 0.05) were made to determine if differences existed in the mean item difficulty difference between dichotomous scoring and partial credit scoring at each level of the cognitive domain. Results indicated that the difference between dichotomous scoring and partial credit scoring was not statistically significant at application and evaluation levels. However, the difference between dichotomous scoring and partial credit scoring was found to be statistically significant in knowledge, comprehension, and analysis levels. The difference between dichotomous scoring and partial credit scoring of evaluation, knowledge, comprehension, and analysis levels were not statistically different among themselves. Therefore, in the analgesic area, the mean item difficulty was increased significantly for the application level when changing from dichotomous scoring to partial credit scoring. The dichotomous scoring and partial credit scoring were not statistically significant for the other cognitive domains. This indicated that partial credit scoring may capture some of partial information (*i.e.*, reflected in terms of cognitive levels being measured) about students who have less than full knowledge or mastery of the content area.

Item difficulty index was "the expected relative score on an item by a population of students," and was interpreted as the proportion of students answering the item correctly(20-22,36). In this study, when compared to dichotomous scoring, the partial credit scoring method took students' partial information or knowledge into consideration. It perhaps could be used to evaluate student performance in terms of proportion of students answering the item correctly (*i.e.*, item difficulty) for those items measuring objectives written at the different cognitive levels. Thus, when compared to dichotomous scoring, this method may allow the instructor to more readily identify students with partial information in a cognitive domain in selected content areas, and those needing remediation. For example, in the analgesic content area, the mean item difficulty was significantly increased from dichotomous scoring (*i.e.*, 0.6867) to partial credit scoring (*i.e.*, 0.8139) at a probability level of 0.05 for the application level. This could be interpreted as the proportion of students answering the item correctly (complete information) and who had less than full knowledge or mastery of the content area (partial information) when using partial credit scoring (*i.e.*, from 68.67 percent of students by using dichotomous scoring to 81.39 percent of students). Thus, 12.72 percent of students had partial information for the application level. This is helpful for the instructor to know because it demonstrates student weakness in the domain of application. Thus, in future presentation of this content area, the instructor can pay more attention to developing student application skills. The mean item difficulty did not, however, significantly increase for other cognitive levels when the partial credit scoring method was compared to the dichotomous scoring method. However, 96.67 percent of students at the comprehension level, 91 percent of students at the knowledge level, 89.75 percent of students at the evaluation level, and 80 percent of students at the analysis level answered the item correctly and who had less than full knowledge or mastery of the analgesic content area when partial credit scoring was utilized.

For item discrimination, the results indicated that there were no significant interaction effects of scoring methods (S1, S2) for each of cognitive levels within the analgesic content area, *i.e.*, Wilks' lambda = 0.5459, F(4, 6) = 1.2476; $P = 0.3843$. The analysis revealed no significant difference in the mean item discrimination for methods of scoring (S1, S2) when averaged across all cognitive levels, *i.e.*, Wilks'lambda = 0.7944, F(1, 6) = 1.5520; $P = 0.2593$. This indicated that the mean item discrimination was not statistically different between dichotomous scoring and partial credit scoring in the analgesic area. MANOVA with follow-up univariate analysis of variance on the cognitive levels in the analgesic content area also showed that there was no statistical difference in the mean item discrimination at each level of the cognitive domain when averaging all over the two methods of scoring (S1, S2), *i.e.*, F(4, 6) = 0.1400; $P = 0.9609$.

In NR evaluation, high discrimination was desirable(20,36). Low or negative discrimination index items were not included in NR examinations because they do not contribute to the ability to distinguish between low- and high-achieving students. In the context of CR evaluation, no specific attention was warranted to small positive or zero discrimination items because the desired instructional outcome as measured by the measurement instrument is high achievement(36). Negative discrimination indices were demonstrative, as in NR situation, that students who did well on the examination did not do well on an item(36). In the CR case, this could be indicative of poor item construction, ineffective instruction, and/or poor instructional materials. A large positive discrimination index indicated some deficiency in the instructional program because

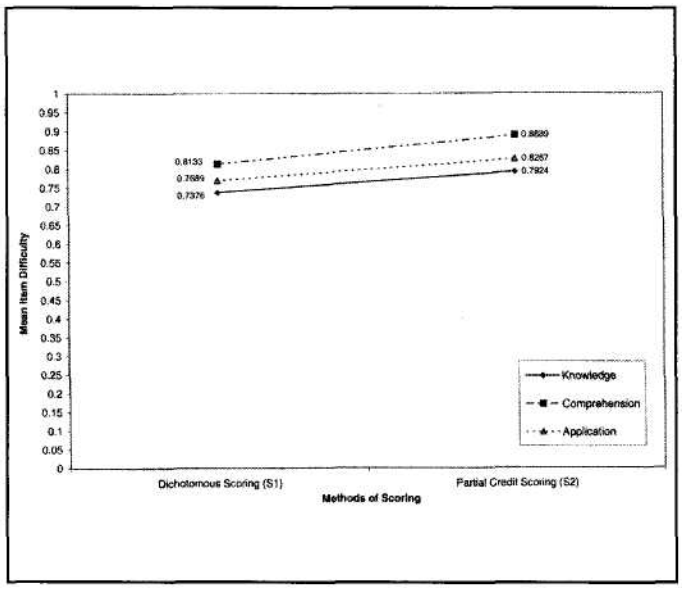

Fig. 4. The interaction effect of scoring methods (S1, S2) and cognitive levels on item difficulty in the content area of laxatives on the third CRMC Examination.

not all students demonstrated acquisition of desired content at the desired cognitive levels(36).

All of the mean item discrimination indices at each level for the cognitive domains demonstrated small positive values (*i.e.*, ranging from 0.1634 to 0.2296) in both dichotomous scoring and partial credit scoring for the content area of analgesics. This meant that no specific attention was warranted toward these small positive values. The instructional outcome (*i.e.*, measuring each of levels of cognitive domain) as measured by the examination for the analgesic content area in the nonprescription drug course was desirable. Most of the students were acquiring the desired content at each of the cognitive levels in analgesic area.

Laxative Content Area. In the third examination, twenty out of seventy-five test items were used to test students' knowledge in the laxative content area (*i.e.*, fourteen items in the knowledge level, three items in the comprehension level, three items in the application level). For item difficulty, data were analyzed using MANOVA. The interaction effect between the methods of scoring (S1, S2) and cognitive levels on the mean item difficulty was not statistically significant, *i.e.*, Wilks' lambda = 0.9835, F(2, 17) = 0.1422; $P = 0.8685$. Figure 4 illustrates this interaction effect. Thus, the mean item difficulty difference attained by using the two methods of scoring (S1, S2) was not different at the various levels of the cognitive domain. MANOVA with follow-up univariate analysis of variance on the methods of scoring in the laxative content area showed that there was a statistically significant higher value for the mean item difficulty for partial credit scoring than dichotomous scoring when averaging all over cognitive level in this content area, *i.e.*, F(1, 17) = 12.7762; $P = 0.0023$. MANOVA with follow-up univariate analysis of variance on the cognitive levels in the laxative content area revealed that there was no statistically significant difference in the mean item difficulty for each level of the cognitive domain when averaging all over the two methods of scoring (S1, S2), *i.e.*, F(2,17) = 0.5100; $P = 0.6101$.

When partial credit scoring was employed, item difficulty was interpreted as the proportion of students answering the item correctly (complete information) and those who had less than full knowledge or mastery of the content area (partial information) (*i.e.*, 88.89 percent of students at comprehension level, 82.67 percent of students at application level, and 79.24 percent of students at knowledge level). Data analysis revealed that 5.48 percent, 5.78 percent, and 7.56 percent of the student population demonstrated partial information at the knowledge, application, and comprehension levels, respectively.

For item discrimination, the results indicated that there were no significant interaction effects of the scoring methods (S1, S2) for each of the cognitive levels in the laxative area, *i.e.*, Wilks' lambda = 0.9927, F(2, 17) = 0.0620; $P = 0.9400$. The analysis revealed no significant difference in the mean item discrimination for the methods of scoring (S1, S2) when averaged across all the cognitive levels, *i.e.*, Wilks' lambda = 0.9974, F(1, 17) = 0.0438; $P = 0.8366$). This indicated that the mean item discrimination was not statistically different between dichotomous scoring and partial credit scoring in the laxative content area. MANOVA with follow-up univariate analysis of variance on the cognitive levels in the laxative area also showed that there was no statistical difference in the mean item discrimination at each level of the cognitive domain when averaging all over the two methods of scoring (S1, S2), *i.e.*, F(2, 17) = 2.18; $P = 0.1432$. All of the mean item discrimination indices at each cognitive level were near zero or demonstrated small positive values (*i.e.*, ranged from 0.0027 to 0.2197) for both dichotomous scoring and partial credit scoring. This indicated that no specific attention was warranted to these near zero or small positive values. The instructional outcome (*i.e.*, measuring each of levels of cognitive domain) as measured by the examination for the laxative content area in the nonprescription drug course was desirable. Most of the students were acquiring the desired content at each of the cognitive levels in the laxative area.

Table II. Percentage of 150 pharmacy students assigned to criterion-referenced grade categories using dichotomous scoring (S1) and partial credit scoring (S2) for each of three CRMC examinations

		Grade category[a]				
Examinations	**Scoring methods**	**A**	**B**	**C**	**D**	**F**
Examination 1	S1	1.30	18.70	52.00	12.00	16.00
	S2	2.70	37.30	48.00	4.00	8.00
Examination 2	S1	4.00	30.00	37.30	14.00	14.70
	S2	8.70	46.00	33.30	10.00	2.00
Examination 3	S1	0.00	17.30	48.00	18.00	16.70
	S2	1.30	36.70	50.00	8.70	3.30

[a]A = 90 - 100 percent, B = 80 - 89 percent, C = 70 - 79 percent, D = 65 - 69 percent, F = 64 percent and below.

Objective Six. Table II demonstrates the percentage of students who would be assigned grades for each of the three examinations based upon the dichotomous and partial credit methods. The partial credit scoring method assigned more students into the "A" and "B" grade categories and fewer students into the "D" and "F" categories compared to the dichotomous scoring method. Thus, the grade distributions based on the same criterion-referenced grading scale differed substantially in the proportion of students assigned to different grade categories for dichotomous and partial credit scoring methods. This illustrated how conventional dichotomous scoring discriminates against students who have partial information and how it fails to reward students for less than full knowledge or mastery of the content area in CRMC format. Motivating students to learn could be enhanced when partial credit scoring is used to reward students for his/her partially correct in answering CRMC examination. Indeed, the most familiar way to reward students is through grades.

LIMITATIONS

Several study limitations were noted in this investigation. One such limitation involved the composition of the subject population. Because the investigators used the entire study population rather than selecting a random sample of the 150 pharmacy students enrolled in the Nonprescription Drugs course (PHPR 470) as a second professional year in either the baccalaureate program or the doctor of pharmacy program, making generalizations of the results of this study to similar populations were limited to the study subjects. Data from this study may suffer from selection bias of the subjects which could threaten internal validity. A second limitation involved the restriction of time for test administration and the small numbers of examination items in the cognitive domain that could be tested. These could affect the assessment of students' actual knowledge and/or ability in the content areas investigated, thereby, eventually lowering the test scores observed in relation to true scores and reliability of the examinations. The last limitation involved the interpretation of the reliability coefficients as mentioned in the results and discussion section.

CONCLUSION

Results from this study demonstrate that the partial credit scoring method clearly made the values of item difficulty higher or more reflective of student knowledge than the dichotomous scoring method for all of the three CRMC examinations. No statistically significant differences ($P > 0.05$) in item discrimination existed between dichotomous scoring method and partial credit scoring method for all of the three CRMC examinations. The partial credit scoring method increased the reliability and had the effect of increasing the reliability of the test's effective length when compared to the dichotomous scoring method for all of the three examinations. The item difficulty difference attained by using the two methods of scoring(S1, S2) was statistically different ($P<0.05$) at different levels of the cognitive domain. This difference depended on the content areas investigated. However, there was no statistical difference ($P>0.05$) that existed in item discrimination at each of the cognitive levels being measured and two scoring methods (S1, S2) in both of the content areas investigated. The partial credit scoring method differed in the effect of CR grade assignments from the dichotomous scoring method. The partial credit scoring method assigned a higher proportion of students into the "A" and "B" grade categories and fewer students into the "D" and "F" grade categories than did the dichotomous scoring method.

RECOMMENDATIONS

It is recommended that future research with CRMC examinations involve replication within core and elective coursework within the pharmacy curricula. It should also be evaluated with different student populations. Preferably, also, these investigations should allow examination environments that facilitate student examination taking (*e.g.*, longer test time, more test items that represent the various cognitive domain levels) and alternate methods to assess the reliability coefficients for CRMC tests designed to be used in making criterion-referenced interpretation). Replication of these findings would increase generalizability

Another direction for future research would be to correlate student performance based on either dichotomous and partial credit scoring with other methods of student achievement (*i.e.*, homework assignments) used to demonstrate student learning. Specific research might focus on which method would demonstrate better ability in predicting student achievement on homework assignments. This is important to evaluate because the MC examinations cannot assess the synthesis domain which demonstrates student ability to integrate concepts and knowledge and apply them to the solution of "real world" problems. Another direction for future research is the investigation of the relationship between students' knowledge and confidence in that knowledge when responding to CRMC examination questions graded either through partial credit scoring or dichotomous scoring. Thus, each CRMC item used on course examinations should be created or designed to assess each of the cognitive levels. It would be very interesting to determine what scoring (S1, S2) method used to assess or measure student knowledge would correlate with student confidence in knowledge according to the various cognitive levels in the content areas of the course. In addition, for those students who possess only partial information or less than full knowledge or mastery of the cognitive skills in certain content areas, partial credit scoring could be used to demonstrate student learning at each cognitive level in certain content areas or courses and their relative degree of confidence. This would provide insight to the instructor when developing or improving his/her educational plan and could help students master the cognitive levels in which students deficient and need to develop/balance their confidence in knowledge at each of cognitive levels in certain content areas or courses.

Acknowledgments. The authors wish to acknowledge the assistance of the external content validity review panel members (*i.e.*, Joye Ann Billow (South Dakota State); Timothy R. Covington (Samford); Janet P. Engle (Illinois at Chicago); Kenneth W. Lem (California at San Francisco); W. Steven Pray (Southwestern Oklahoma State) and Henry A. Palmer (Connecticut) for their help in validating the test pool items. Further, the authors wish to express special thanks to Gary L. Wright, Manager of Computer-Based Services, Center for Instructional Services, Purdue University; and Mario Rodriquez, Curriculum and Instruction Department, School of Education, Purdue University for their expertise in developing a computer grading mechanism to generate student score and item analysis.

Am. J. Pharm Educ., **64**, 1-10(2000).

References

(1) Haladyna, T.M., *Developing and Validating Multiple-Choice Test Items,* Lawrence Erlbaum Associates, Hillsdale NJ (1994) pp. 43-44.
(2) Haladyna, T.M., "The effectiveness of several multiple-choice formats,"

Appl. Measure. Educ., **5**, 73-88(1992).
(3) Albanese, M.A., "Type K and other complex multiple-choice items: An analysis of research and item properties," *Educ. Measure. Issues and Practice*., **12**, 28-33(1993).
(4) Zin, T.T. and Williams, J., "Searching for better scoring of multiple-choice tests: Proper treatment of misinformation, guessing and partial knowledge," (Reports-Evaluative/Feasibility 142). East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED 339 744)
(5) Coombs, C.H., Milholland, J.E. and Womer, KB., "The assessment of partial knowledge," *Educ. Psychol. Measure*., **16**, 13-37(1956).
(6) Garcia-Perez, M.A. and Frary, R.B., "Psychometric properties of finite-state scores versus number-correct and formula scores: A simulation study," *Appl. Psychol. Measure*., **13**, 403-417(1989).
(7) Hakistan, R. and Kansap, W., "A comparison of several methods of assessing partial knowledge in multiple-choice testing: II. Testing procedure," *J Educ. Measure*., **12**, 231-239(1975).
(8) Millman, J. and Greene, J., "The specification and development of tests of achievement and ability," in *Educational Measurement*, 3rd ed., (edit., Linn, R.L.), American Council on Education and Macmillan, New York NY(1989) pp.335-366.
(9) Abu-Sayf, F.K., "The scoring of multiple-choice tests: A closer look," *Educ. Technol*., **19**, 5-15(1979).
(10) Claudy, J.D., "Biserial weights: A new approach to test item option weighting," *Appl. Psychol. Measure*., **18**, 167-172(1978).
(11) Frary, R.B., "Partial-credit scoring methods for multiple-choice tests," *Appl. Measure. Educ*., **2**, 79-96(1989).
(12) Davis, F.B. and Fifer, G., "The effect of test reliability and validity of scoring aptitude and achievement tests with weights for every choice," *Educ. Psychol. Measure*., **19**, 159-170(1959).
(13) Patnaik, D. and Traub, R.E., "Differential weighting by judged degree of correctness," *J. Educ. Measure*., **10**, 281-286(1973).
(14) Hambleton, R.K. and Swaminathan, H., *Item Response Theory: Principles and Application*, Kluwer, Boston MA (1985) pp. 1-25.
(15) Hambleton, R.K., Swaminathan, H. and Rogers, H.J. *Fundamentals of Item Response Theory*, Sage, Newbury Park CA (1991) pp. 1-31.
(16) Muraki, E., "A generalized partial credit model: Application of an EM algorithm," *Appl. Psychol. Measure*., **16**, 159-176(1992).
(17) Hambleton, R.K. and Van Der Linden, W.J., "Advances in item response theory and applications: An introduction," *ibid*., **6**, 373-378(1982).
(18) Hsu, T.C., Moss, P.A. and Khampalikit, C, "The merits of multiple-answer items as evaluated by using six scoring formulas," *J. Exp. Educ*., **52**, 152-158(1984).
(19) Jaradat, D. and Sawaged, S., "The subset selection techniques for multiple choice tests: An empirical inquiry," *J. Educ. Measure*. **23**, 369-376(1986).
(20) Mehrens, W.A. and Lehmann, I.J., *Measurement and Evaluation in Education and Psychology*, 4th ed., Holt, Rinehart and Winston, Orlando FL (1984) pp. 160-170, 248-264.
(21) Kubiszyn, T. and Borich, G., *Educational Testing and Measurement: Classroom Application and Practice*, 5th ed., HarperCollins, New York NY(1996) pp.60-70, 127-143, 309-318.
(22) Friedenberg, L., *Psychological Testing: Design, Analysis, and Use,* Allyn&Bacon, Needham Heights MA(1995) pp.178-282.
(23) Bloom, B.S., *Taxonomy of Eeducational Objectives: The Classification of Educational Goals, Handbook I Cognitive Domain*, Longmans, New York NY(1956) pp.62-197.
(24) Gronlund, N.E., *Stating Objectives for Classroom Instruction,* 3rd ed., Macmillan, New York NY(1985) pp.33-41.
(25) Anderson, L.W. and Sosniak, L.A., *Bloom's Taxonomy : A Forty-Year Retrospective*, National Society for the Study of Education, Chicago IL(1994)pp.9-40.
(26) Gronlund, N.E., *Measurement and Evaluation in Teaching*, 6th ed., Macmillan, New York NY(1985) pp.166-191.
(27) Pray, W.S., *The Development of a Standardized Competency Examination for Doctor of Pharmacy Students*, Doctoral Dissertation, Purdue University Graduate School, West Lafayette IN(1983) pp.72-121.
(28) Pray, W.S. and Popovich, N.G., "The development of a standardized competency examination for doctor of pharmacy students," *Am. J. Pharm. Educ*., **49**, 1-9(1985).
(29) Gay, L.R., *Educational Research: Competencies for Analysis and Application,* 3rd ed., Merrill, Columbus OH(1987) pp. 220-246.
(30) Microsoft Excel for Windows Version 5.0 User's Guide, Microsoft Corporation (1994).
(31) *SAS/STAT User's Guide Release 6.03 Edition*, SAS Institute, Cary NC (1988).
(32) Berk, R.A., "Selecting the index of reliability," in *A Guide to Criterion-Referenced Test Construction,* (edit., Berk, R.A.), Johns Hopkins University Press, Baltimore MD (1984) pp. 231-266.
(33) Hambleton, R.K., "Criterion-referenced measurement," in *Educational Research, Methodology, and Measurement: An International Handbook,* 2nd ed., (edit., Keeves, J.P.), Elsevier Science, Tarrytown NY(1988) pp.719-725.
(34) Worthen, B.R., Borg, W.R. and White, K.R., *Measurement and Evaluation in the Schools*, Longmans, White Plains NY (1993) pp. 164-172.
(35) Cunningham, G.K., *Educational and Psychological Measurement*, Macmillan, New York NY (1986) pp. 197-199.
(36) Manasse, H.R. and Laskar, P.A., "Some considerations regarding norm-referenced and criterion-referenced testing in pharmaceutical education," *Am. J. Pharm. Educ*., **40**, 275-278 (1976).

APPENDIX. CONTENT REVIEW FORM

Question # _____ Assigned to Objective # _____

Keyed Response _____

Which of the following patients is (are) appropriate to treat with a nonprescription internal analgesic?

1) A 17-year-old girl with sulfa allergies who suffers from dysmenorrhea.
2) A 6 month pregnant 32-year-old woman with backache pain of three weeks duration.
3) A 39-year-old man with mouth pain and fever due to abscess associated with dental caries.

Review Process
Prior to revision of the item, if any:

A. Does the question fall into the objective it has been assigned?
_____ Yes _____ No

B. If no, which objective is more appropriate? _____

C. How representative is this question of the skills or knowledge required to carry out the objective to which it was originally assigned? (circle one number)
1. Quite representative
2. Possibly representative
3. Not representative

If you assigned it to another objective, how representative is it of the skills or knowledge required to carry out this activity using the above scale?

1 2 3

D. Is another response more correct than that keyed? If so, it is _____
Second best answer _____
Third best anlswer _____
Fourth best answer _____
Fifth best answer _____
Or there is only one, best answer _____.

Any suggested revision may be written here or directly on to the question/response above.