Methodology for the Assessment of Competence and the Definition of Deficiencies of Students in All Levels of the Curriculum

Robert B. Supernaw and Reza Mehvar

School of Pharmacy, Texas Tech University Health Sciences Center, 1300 South Coulter Street, Amarillo TX 79106

Methods for determination of minimum competence (*i.e.,* cut-scores or passing grades) have been described in the higher education literature. However, the use of these methods in pharmacy education has been rare, if at all. This article describes the two major methods, namely Nedelsky and Angoff, which have been used for determination of cut-scores or passing grades in multiple-choice assessment instruments. Generally, both methods involve: (*i*) definition of a minimally competent student (*ii*) development of an assessment tool which tests specific knowledge, skills, and/or abilities; and (*iii*) determination of cut-scores by a subject matter expert panel. The major difference between the two methods lies only in the way the cut-scores are determined. Whereas the Nedelsky method establishes cut-scores by determining the likelihood of eliminating obviously incorrect options (answers), the Angoff method treats the question and its options as a whole and determines the likelihood of a minimally-competent student answering the question correctly. In both methods, the cut-score for the whole assessment is determined by the average of cut-scores for individual questions. A brief description of the methods and a modification of the Angoff method used at Texas Tech School of Pharmacy in an annual assessment are described. These methods may be useful for other schools of pharmacy that plan to assess their student learning using competency-based assessment tools.

INTRODUCTION

Perhaps spurred by the accreditation standards(1) of the American Council on Pharmaceutical Education (ACPE), many schools of pharmacy have developed or are beginning to develop assessment strategies. While many assessment models have been considered(2), with the exception of licensing and certification examinations, few have focused on achieving prescribed scores on knowledge, ability, and skill assessment instruments. Within schools of pharmacy, the major challenge facing the faculty is the problem of the methodology for determining passing scores (cut-scores) in a valid fashion. The faculty of Texas Tech University Health Sciences Center School of Pharmacy has developed a process for such determinations. Ongoing faculty development and the development of a method for cut-score determination have given rise to an annual assessment program that measures student achievement of specific academic expectations at each level of the curriculum, while defining deficiencies of individual students.

In the Year 2000, the faculty of the Texas Tech School of Pharmacy envisioned a program of academic assessment that would discern between mastery and lack of mastery of prescribed knowledge, skills, and abilities of its students. To realize this vision, a goal to develop such an assessment was set – an assessment that would serve two principal functions. First, the assessment would focus the attention of the students on their specific area or areas of mastery and of weakness. Second, the assessment would indirectly assess the ability of the Texas Tech to provide the curriculum and the academic support necessary for the students' mastery of these competencies.

Several schools of pharmacy have some form of annual testing of students(3). However, most report student scores on a relative basis. That is, results are reported as raw scores and percentages, and some are reported as deviations from the class means or as stanine scores. The faculty of Texas Tech desired to develop a system of testing that could be used to calculate the score on any test and subtest that would indicate minimum competency. In short, a student scoring at or above the calculated cut-score would be deemed as having met the standard (minimal competence), while a student scoring below the calculated cut-score would be deemed as not having met the standard.

Although description and application of competency determination methods are available in the higher education literature(4), the use of these methods in pharmacy education has not been reported. Therefore, the purpose of this article is to describe major methods available for determination of cut scores or minimum competencies and their application to annual assessments at Texas Tech School of Pharmacy.

METHODOLOGY

Several techniques have been developed for the determination of the test score necessary to demonstrate minimal competence on a criterion-referenced examination(4). In professional knowledge, skill, and ability assessment and certification and licensing testing, the most widely used techniques for this determination are the Nedelsky(5) and the Angoff(6) methods. The overall aim of both methods is to discriminate between individuals who are competent and those who are not. For the

Am. J. Pharm. Educ., **66,** 1–4(2002); received 9/25/01, accepted 1/21/02.

purposes of competency assessment, knowledge, skill, and ability demonstration is defined by the criterion-referenced test score. With both methods, a cut-score (minimum score required to demonstrate competence) is defined through a prescribed process. Both methods entail deliberate processes that lead to the development of specific ability sets to be tested(7). Expert judges are selected and empanelled; and definitions of the "minimally competent" candidate developed. The expert panel is trained in the process, and judgments are made. The cut-scores (minimal competency scores) are then calculated from the collected judgments. The main difference in the two methods is in the item leveling processes - the processes employed to judge individual test items as to the likelihood of a minimally competent-candidate answering the question correctly.

Two methods that are not as commonly employed are the Ebel method and the Bookmark method. The Ebel method(8) employs the expert panel to classify questions into categories of importance and degree of difficulty. Based upon the panel's categorization of a question, the degree of difficulty for a minimally competent candidate is set. The composite categorization of a set of questions then constitutes the cut-score. In the Bookmark method(9), test items are ordered sequentially, from the easiest to the most difficult. Panelists then evaluate the items, beginning with the easiest, moving towards the more difficult items, until they reach the point at which the panelists believe the minimally competent candidate can no longer answer the item correctly. This point is "bookmarked" as the cut-point. This method can also be employed to bookmark cut-points for various levels of competence.

Additionally, combined methods can be employed. That is, more than one method can be used, and the results can be averaged, or a principal method can be used and the results adjusted to constitute an adjusted cut-score. The NAPLEX passing score is calculated in this manner. While a modified Angoff method is the principal method used for the NAPLEX, the results are adjusted secondary to consideration of post-test results.

Nedelsky Method

The Nedelsky method(5) calls for questions to correspond to specific knowledge, skill, or ability statements. After a question has been reviewed for content, accuracy, and format, a panel of experts calculates its degree of difficulty for the minimally-competent examinee. The main premise of the Nedelsky method is that the test takers who do not know the correct answer to a question will eliminate as many answers as possible before making their final selection or guess. To that end, each answer option is evaluated and judged as to its likelihood of being eliminated by the minimally-competent examinee. For a four-option question (A-D), the "expected score" is derived from this evaluation; it is the reciprocal of the number of options NOT eliminated. Therefore, for the four-option multiple-choice question, expected scores can be 1.00, 0.50, 0.33, or 0.25. That is, if it is judged that the minimally-competent test taker should be able to eliminate all three incorrect responses, then the number of options NOT eliminated is 1, the reciprocal of which is 1.00. If just two of the responses are likely to be eliminated, then the number of options NOT eliminated is 2, the reciprocal of which is 0.50. Similarly, if just one of the responses is likely to be eliminated, then the number of options NOT eliminated is 3, the reciprocal of which is 0.33. And, finally, if the minimally-competent examinee is not expected to be able to eliminate any of the responses, then the number of options NOT eliminated is 4, the reciprocal of which is 0.25. As can be easily seen in these illustrations, for an assessment using questions that are clearly above the competence level of a candidate, in a given set of four-option questions, a minimally-competent candidate will still be expected to achieve a score of 0.25 or 25 percent by merely guessing. The passing or cut-score calculated using the Nedelsky method is the average of the degree of difficulty (reciprocal of the options NOT eliminated) of all questions asked. This cut-score is the expected score of the minimally-competent examinee. A test taker scoring at or above the cut-score is deemed to be competent.

An advantage of the Nedelsky method is its simplicity. It is relatively easy for a panel of experts to project which distractors a minimally competent candidate should be able to eliminate. However, a disadvantage of this system is that its primary focus is on the negative – those answer options that are not correct. There is only indirect attention on the correct answer. In pharmacy education, most faculty would agree that it is preferable for a student to know the correct answer rather than arrive at the correct answer by process of elimination.

Angoff Method

The Angoff method(6) is similar to the Nedelsky method(5) in that its use leads to the determination of a cut-score which defines the performance level of the minimally competent candidate. It also calls for questions that test the mastery of specific knowledge, skill, or ability statements. However, the expert panel in the Angoff method is asked to judge the likelihood of a minimally-competent candidate answering a specific question correctly, rather than judging which distractors can be eliminated by such a candidate. In contrast to the Nedelsky method, judges using the Angoff method are not limited to assessments of success of just 25, 33, 50, and 100 percent for a four-option question. Instead, a judge may opine that the likelihood of success in answering a question is any number between zero and one hundred, representing any percent likelihood of success. However, judges will never rate a question lower than 25 percent, as this figure represents the likelihood of a pure guess being correct. In short, in the Angoff system, the judge is asked to project the likelihood of a candidate knowing the correct response rather than the candidate's ability to rule-out distractors. Herein lies its major advantage. Additionally, it allows the panelists to argue their cases on the basis of the correct answer, rather than having to argue their cases related to all three of the distractors and the correct option.

Application of the Angoff Method at Texas Tech

At Texas Tech, the members of the Outcomes Assessment Committee evaluated each method and opted to use the modified Angoff method. Instead of establishing minimum competencies at a single point (*i.e.*, at the conclusion of the pharmacy program), the faculty decided to use the method to define the expected scores of students at four points in the curriculum. In this manner, for a given question, an expected score was calculated for the P1, P2, P3, and P4 students at the mid-point (January) in each year. Questions designed for the P1 students were administered to all students; however, there were different cut-score calculations for each level of the curriculum. For example, a typical expectation for the P2 students is as follows.

Knowledge, Skill, or Ability Statement: The P2 student should be able to perform pharmacokinetic calculations.

A question developed to test the mastery of this ability statement might be as follows.

P2 Question. Cefonicid has a renal clearance of 20 ml/min in healthy subjects. The free fraction of the drug in plasma is 0.02. Assuming a glomerular filtration rate of 120 ml/min, what is/are the mechanism(s) ABSOLUTELY involved in the renal excretion of the drug?

A. glomerular filtration only
B. glomerular filtration and tubular reabsorption
C. glomerular filtration and active secretion
D. glomerular filtration, active secretion, and tubular reabsorption

Question Leveling. After each question is quality controlled by the expert panel (in this case, the expert panel on *Use of Basic Sciences in the Practice of Pharmacy*) for accuracy, correspondence with the stated knowledge, skill, or ability statement, and format, the question is forwarded to the Outcomes Assessment Committee for leveling.

Before the leveling process begins, the seven to ten-member panel is asked by a facilitator to envision a class of one hundred students, all of whom are minimally competent. It is important to remind the panelists that the term minimally competent is not a term of disparagement. Clearly, a minimally-competent student is one who meets the expectations and passes the course. At Texas Tech, to pass a course, a student must receive at least a 70 percent. Therefore, the panel is asked to visualize a class of students who always score 70 percent on every assignment and test – no higher, and no lower.

The question leveling process is electronically facilitated, and each panelist is provided with a networked laptop. To level the question, it is first shown on a screen and carefully evaluated by each member of the seven to ten-member panel. After some discussion of the question, additional modifications are made, if indicated. When the panel agrees that the question is appropriate, each member expresses an opinion as to what percentage of the visualized class of minimally-competent students will be able to answer the question correctly. After each has entered his/her scores (estimated percentage of the class who will answer correctly), all scores are shown on the screen. The panelists who project the highest and lowest scores are then asked to state their rationale for the benefit of the entire panel. After these discussions, the panel is asked to re-score the question. The computer program then discards the high and the low re-scores and calculates a mean from the remaining scores. This mean then becomes the cut-score for P1 students for the question. This process is termed question leveling. In essence, the question leveling process is a determination of the expected degree of difficulty (pi) of a question administered to a minimally-competent student.

After a question developed for the P1 class is leveled, the process is repeated for the same question for each higher level of the curriculum. Therefore, each P1 question is leveled four times (for P1, P2, P3, and P4 students), whereas P2, P3, and P4 questions are leveled, respectively, three times (P2 questions for P2, P3, and P4 students), two times (P3 questions for P3 and P4 students), and one time (P4 questions for P4 students only). For example, for the P2 question in the example above, the cut score for the P2 students was 0.45; meaning 45 percent of the minimally-competent P2 students were expected to answer the question correctly. Also, this anticipated pi for the minimally competent student (cut score) indicates that a minimally-competent individual P2 student had a 45 percent likelihood of answering the question correctly. The panel also determined the cut scores for the same question for the P3 and P4 classes, respectively. In this manner, this particular P2 question was leveled for three classes (*i.e.*, the three classes at or above the question level).

When all questions were leveled, a mean was taken for all questions asked for each class. This mean then represented the specific benchmark or composite cut-score that defined overall competence for each level of the curriculum. The test was administered to all the P1 (n = 80), P2 (n = 85), P3 (n = 52), and P4 (n = 52) students.

DISCUSSION

An integral component of the assessment of student performance is measurement of long-term mastery of knowledge, skills, and abilities. In-class examinations, while indisputably valuable indicators of content mastery, arguably measure short-term memory. Annual assessments are superior indicators of the mastery of global competencies and, indirectly, the ability of the curriculum to deliver what it purports to deliver. However, a recent survey(10) indicates that the use of annual or cumulative examinations is not widespread among the U.S. schools of pharmacy. Additionally, most annual assessments report scores on a relative basis. Class standing, comparisons with means, and percentage scores are most often the substance of the results. At Texas Tech, the faculty has developed a process for the calculation of scores indicative of competence, based upon a validated method of cut-score determination. Modifications of the prescribed system have been made to determine cut-scores for each level of the curriculum. At Texas Tech, the 2001 annual assessment, consisting of a total of 216 questions (the number of questions for the P4 students) was very reliable, as indicated by a calculated reliability coefficient of 0.745. This was an indication that an alternative test of the same knowledge, skill, and ability statements would yield the same results. That is, a reliability coefficient of 0.745 indicates that 74.5 percent of the observed score variation can be attributed to true score variance(11). Also, it can be concluded that the correlation coefficient that expresses the relationship of observed score to true score (reliability index) is 0.863, indicating a "high" degree of correlation, therefore, a high degree of test reliability(12).

In terms of validity, the process for the development of the instrument followed the prescribed steps in establishing face validity and content validity(13). The items tested the mastery of the global knowledge, skills, and ability statements developed by the faculty and practitioners. However, it must be clearly understood that a written examination cannot measure actual performance in a clinical setting. Additionally, while of debatable meaning, the process described led to an assessment that yielded results which significantly correlated with: (*i*) didactic course grades (Pearson correlation 0.521, $P < 0.001$); (*ii*) fall semester clerkship grades (Pearson correlation 0.459, $P = 0.001$); and (*iii*) NAPLEX scores (Pearson correlation 0.485, $P = 0.001$).

At Texas Tech, there are over 400 knowledge, skill, and ability expectation statements. Clearly, all statements cannot be tested in an efficient manner in any one or two-day exam.

Therefore, the faculty opted to select representative expectation statements to be tested as indicators of categorical competence. Seventy-two were selected (three per curricular level in each of six categories). Each representative statement was tested with three corresponding questions, yielding a total of 216 questions for the P-4 students, 162 for the P-3 students, 108 for the P-2 students, and 54 for the P-1 students.

A consideration of the process of question leveling and cut-score calculation must include the amount of faculty time involved. At Texas Tech, the question leveling process consumed an entire weekend for the faculty who comprised the expert panel for question leveling. However, those faculty who participated in the leveling process all remarked that they felt their time was well spent and that the process was a valuable one. On average, each faculty member was required to write only four or five questions. With panel review and leveling, each faculty member, on average, expended less than three hours on the construction, quality control, and leveling processes.

One of the most important functions of the Outcomes Assessment Committee was to convince the faculty that the modified Angoff method would, indeed, yield appropriate results – cut-scores that would be truly indicative of knowledge, skill, and ability statement mastery. To accomplish this, the Committee randomly selected a question from a previous annual assessment – a question that had not been subjected to the leveling process. The expert panel was asked to evaluate the question and project the likelihood of the previous year's P4 class answering the question correctly. It was made clear to the panel, that what was being asked was not the likelihood of a class of minimally-competent students answering the question correctly, rather, the pi for the entire class. This exercise was undertaken at the beginning of the leveling process each of the two days devoted to the process. In each instance, the modified Angoff method yielded a pi value that was within 0.1 percent of the actual performance of the students. This exercise is recommended to strengthen the confidence the faculty have in the approach - an approach that, at first glance, may seem to some rather unscientific. A suspension of skepticism significantly enhances the resolve of the faculty to participate willfully and to advise students regarding the results of the assessment meaningfully.

In conclusion, the methods available in the literature for determination of cut-scores or competency levels are reviewed and application of one of these methods to the Annual Assessment at Texas Tech is described here. However, the Annual Assessment at Texas Tech has been an evolving multi-step process, which, in addition to the determination of cut-scores, involves methods for selection of curricular abilities to be tested, development of the assessment tool, administration of the assessment, preparation and distribution of reports, and evaluation of the assessment by both faculty and students. These issues will be reported in a subsequent paper.

Acknowledgment. The authors would like to acknowledge significant contribution of all the faculty of Texas Tech School of Pharmacy and, in particular, the members of the school's Outcomes Assessment Committee and the members of the faculty expert panel who participated in the determination of cut-scores in the 2000-2001 academic year.

References

(1) Accreditation Standards and Guidelines for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree, American Council on Pharmaceutical Education, Chicago IL (1997).
(2) Madaus, G.F., Scriven, M. and Stufflebeam, D.L., *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*, Kluwer-Nijhoff Publishing, Boston MA (1991).
(3) Bouldin, A.S. and Wilkin, N.E., "Programmatic assessment in U.S. schools and colleges of pharmacy: A snapshot," *Am. J. Pharm. Educ.*, **64,** 380-387(2000).
(4) Livingston, S.A. and Zieky, M.J., *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*, Educational Testing Service, Princeton NJ (1982).
(5) Nedelsky, L., "Absolute grading standards for objective tests." *Educational Psychol. Measurement*, **4,** 3-19(1954).
(6) Angoff, W.H., "Scales, norms, and equivalent scores," in *Educational Measurement*, (edit., Thorndike, R.L.), American Council on Education, Washington DC (1971) pp. 514-515.
(7) Fabrey, L.J., *Standard Setting: Nedelsky vs. Angoff*, Applied Measurement Professionals, Lenexa KS (1991).
(8) Ebel, R., *Essentials of Educational Measurement*, Second Edition, Printice Hall, Englewood Cliffs NJ (1972).
(9) Lewis, D.M., Mitzel, H.C., Green, D.R., "Standard setting: A bookmark approach." Presented paper, Annual CCSSO National Conference on Large Scale Assessment, Phoenix AZ (1996).
(10) Ryan, G.J. and Nykamp, D., "Use of cumulative examinations at U.S. schools of pharmacy," *Am. J. Pharm. Educ.*, **64,** 409-412 (2000).
(11) Crocker, L. and Algina, J., *Introduction to Classical and Modern Test Theory*, Harcourt Brace Jovanovich College Publishers, Fort Worth TX (1986) pp. 114-117.
(12) *Op. Cit.* (9), pp. 32-33.
(13) *Op. Cit.* (9), pp. 217-242