American Journal of Pharmaceutical Education **2010; 74 (5) Article 80.**

INSTRUCTIONAL DESIGN AND ASSESSMENT

Development of a Reliable, Valid Annual Skills Mastery Assessment Examination

Gregory L. Alston, PharmD, and Bryan L. Love, PharmD

Wingate University School of Pharmacy

Submitted July 30, 2009; accepted December 19, 2009; published June 15, 2010.

Objective. To develop a methodology for a reliable, valid annual skills mastery assessment examination to provide formative student feedback, inform curricular review, and comply with the Accreditation Council for Pharmacy Education (ACPE) Standards 2007.

Design. A sample of program-level ability-based outcomes skills were chosen for the examination. Test items were written, underwent quality control, and were scored for level of difficulty. Versions of the examination for first-, second-, third-, and fourth-year pharmacy students were developed and administered, the results were analyzed, reliability and validity were evaluated, and reports were generated. Item-writing guidelines, quality control procedures, and examination production steps were codified to create a criterion-referenced examination. Students and faculty advisors received detailed score reports and results were used to guide student performance and stimulate a review of curricular outcomes.

Assessment. Content, criterion, and construct validity were analyzed as defined in the literature for the intended use of this assessment tool. Data suggest the Annual Skills Mastery Assessment (ASMA) examination is both reliable and valid. Students and faculty members were surveyed regarding the usefulness of the examination. Results indicate general satisfaction with the assessment program.

Conclusion. A reasonably reliable, reasonably valid multiple-choice annual skills mastery assessment for selected outcomes statements providing formative feedback and informed curricular review was developed.

Keywords: progress examination, examination, skill mastery, ability-based outcomes, assessment

INTRODUCTION

Progress examinations are considered viable tools for pharmacy program assessment. ACPE 2007 Standards Guideline 15.1 states that PharmD programs should "incorporate periodic, psychometrically sound, comprehensive, knowledge-based, and performance-based formative and summative assessments including nationally standardized assessments."[1] Although a locally constructed examination may not be as psychometrically robust as a commercially prepared standardized examination, it has 3 potentially significant advantages: it can be tailored to assess the specific terminal ability-based outcomes (TABOs) of the PharmD program; the college or school has complete access to and control of the data; and the assessment program can provide useful formative student feedback.

Key considerations for developing a valid progress examination include how well the examination: includes important content; aligns content with the curriculum; reflects what should be learned; measures what it purports to measure; reflects individual student scores in a meaningful way; allows cost-effective production; and delivers results in a timely fashion.[2] Additionally, the examination should be reliable and reasonably valid as shown by content, thinking skills, internal, external, and consequential evidence.

The cognitive domain consists of all intellectual behaviors, including the 2 major categories of achievement and ability. Achievement refers to behavior that is easy to change, and includes 2 subcategories of knowledge and skills. Knowledge includes the facts, concepts, principles, and procedures that provide the core content of any curriculum. Skills are higher-order acts that require knowledge and involve performance in context. Ability refers to cognitive behavior that is more difficult to change. Ability is the long-term learning of a more complex behavior such as critical thinking or problem solving.[3] Ability develops from a foundation of knowledge and skills (Table 1).

Knowledge, skills, and abilities therefore exist on a continuum of increasing complexity. Performance of an

Corresponding Author: Gregory L. Alston, PharmD, Wingate University School of Pharmacy, Campus Box 3087, Wingate, NC 28174. Tel: 704-233-8329. Fax: 704-233-8332. E-mail: galston@wingate.edu

Table 1. Description of the Difference Between Knowledge, Skill, and Ability to Provide Patient Care[a]

Achievement	Ability
Knowledge: Know about adverse reactions, side effects, SOAP Notes, HIPAA, counseling techniques. **Skills:** detecting an adverse reaction, detect a side effect, write a SOAP note, maintain HIPAA compliance, counsel a patient	The ability to adequately provide pharmaceutical care to a specific patient.

Abbreviations: SOAP = subjective, objective, assessment plan; HIPAA = Health Insurance Portability and Accountability Act
[a] Format adapted from Haladyna[18]

ability requires the application of knowledge and skill, both of which can be learned and assessed within a short timeframe. The corresponding complex ability develops unpredictably and may not emerge until years later, potentially confounding the attempt of schools to measure the achievement of ability within the timeframe available to educators.

The American Educational Research Association (AERA) states that every high-stakes educational examination program should meet several conditions including: examinees should be protected against high-stakes decisions based on a single test; examinations should be validated for each use; likely negative consequences should be explained prior to examination administration; curriculum and test content should be in alignment; validity of passing scores should be verified; opportunities for remediation should be provided; sufficient reliability for each intended use should be measured; and an ongoing evaluation of consequences of the examination should be conducted.[4]

Because examinations are comprised of test items, proper item development is critical to ensure validity. Downing and Haladyna[5] described a quality assurance procedure to provide evidence of test validity through proper test item development. An ideal process documents how items are developed, how responses to the items are studied to ensure the test items are sound, and provides qualitative and quantitative forms of evidence. Table 2 summarizes the types of evidence required to make a reasonable claim of validity for an examination. The Outcomes Assessment Committee, comprised of the assistant dean for assessment (chair) and 4 faculty members, followed this guide to design the Annual Skill Mastery Assessment (ASMA) examination. The examination was developed, printed, and scored using LXR Test Software (Logic eXtension Resources, Georgetown, SC).

The committee set the following goals for the examination: (1) minimum reliability of (0.60) as calculated by Cronbach's alpha scale, (2) face validity and content validity as evidenced by adherence to quality test design and qualitative and quantitative evidence, (3) criterion-related validity as evidenced by correlation to concurrent measures of performance, cumulative professional program grade point average (GPA), and class rank, (4) construct validity as evidenced by proper design and usefulness of this instrument for the intended purpose as measured by faculty members' and students' survey responses.

This article will describe the methodology used to improve the reliability and validity of the examination.

DESIGN

Wingate University School of Pharmacy (WUSOP) developed a multiple-choice, 4-option, single best-answer examination, with no guessing penalty, to compose the annual skills mastery assessment for the PharmD program. Motivating factors for choosing this format over other options included low-cost, limited faculty resources, and the limited availability of a trained pool of standardized patients needed to perform valid clinical simulations. Four versions were assembled to address curriculum-specific, grade-level skills for each class year. The 4 examinations were administered in March 2008. The examination included 62 items (6 TABO) for first-year (P1) students, 94 items (11 TABO) for second-year (P2) students, 128 items (16 TABO) for third-year (P3) students, and 170 items (21 TABO) for fourth-year (P4) students. Reliability was computed using Cronbach's alpha scale, and evidence of validity was gathered. An annual examination, administered at WUSOP each year since 2004, is organized around program-level outcomes (skills) rather than course-level knowledge. Properly executed, this strategy should create a criterion-referenced examination capable of providing a direct link to the assessment of program ability-based objectives to meet accreditation guidelines. The process was updated in 2008 to take advantage of the database utility of LXR Test software.

WUSOP has developed a constellation of assessments that are administered in an attempt to triangulate assessment data and guide institutional decisions. In isolation, each of these unique data sets describe only a snapshot of a single moment in time. The information from multiple assessments adds context to all other assessments. Only when multiple assessments are considered

Table 2. Model of Qualitative Item Validity Evidence Adapted from Downing and Haladyna[5] and Used in Assessment Design Methodology

Type of Evidence	Activity	Evidence Needed	WUSOP Evidence Documented
Content definition	Practice analysis, job analysis, CAPE outcomes, ACPE standards	Document the method used to select content	Content Selected from CAPE, ACPE Guidelines, And WUSOP TABO.
Test specifications	Table of specifications or test blueprint created	Document the linkage between content and blueprint	Mastery Specifications report by WUSOP TABO, Test Blueprint created.
Item writer training	Develop training materials; train item writers	Document methods, principles, strategy	Faculty development and group writing sessions held, Item writing guidelines developed.
Adherence to item-writing principles	Standard item-writing rules used by all item writers	Document compliance with rules and process used to review	Item writing guidelines used as basis for item review, TABO dropped from examination due to non-compliance with guidelines.
Cognitive behavior	Cognitive classification system used	Documentation and literature support of system used and rationale	References provided to target items around higher order skills. TABO are based on demonstration of ability.
Item content verification	Content experts review and judge items	Document review sessions and credentials of reviewers	All reviewers were active teaching faculty with a doctorate degree and followed the item writing process.
Item editing	Review items and professionally edit	Document credentials, experience of editors, and guidelines	All reviewers were active teaching faculty with a doctorate degree and followed the item review process.
Bias-sensitivity review	Bias policies and procedures developed	Documentation of review and rationale	Accommodations made for students with disabilities, all students held to same standards, further analysis required.
Item tryout and pretesting	Field test items; item performance data, examinee interviews	Document item test and procedures to include or drop an item	Items reviewed by faculty panel during Angoff sessions and poor items dropped from test after item analysis and before scoring, some TABO not tested because of low quality items.
Key validation and verification	Correctness of keyed answer verified by experts	Document policies and procedures for verification	All reviewers were active teaching faculty and followed the item review process.
Test security plan	Test security policy developed and implemented	Document policies and procedures, post-event analysis	Full security employed and described, all test booklets and score sheets tracked, examination versions created.

Abbreviations: CAPE = Center for the Advancement of Pharmaceutical Education; ACPE = Accreditation Council for Pharmacy Education; TABO = Terminal Ability-based Outcomes

in context can they provide value to inform good decision making.

Assessment data from student focus groups, course evaluations, teacher evaluations, faculty surveys, student surveys, student reflections, faculty reflections, exit interviews, preceptor surveys, and employer feedback, combined with our annual assessment examination, provide sufficient contextual support for guiding decisions.

The Outcomes Assessment Committee developed the WUSOP examination process by leaning on the methodology and experience at Texas Tech University as described by Supernaw and Mehvar.[6,7] The committee developed a test blueprint, item-writing guidelines, item development procedures, examination security protocols, and examination scoring procedures based on a review of studies by Case and Swanson,[8] Downing,[9] and Kehoe.[10,11] To minimize item preparation time, the number of options per item was set at 4. Evidence does not support the need for more than 4 options and suggests that fewer are appropriate.[12,13]

Progress examinations can be either *norm-referenced*, which compares the scores of the current test taker to a previous group of test takers, or *criterion-referenced*, which compares the score of the current test taker to a set of standards.[14] Typically, the content for a norm-referenced test would be selected based on how well it distinguishes 1 student from another, while the content for a criterion-referenced test would be selected based on how well it matches the learning outcomes deemed most important to the curriculum.[14] Given the goal to create valuable formative student feedback, a criterion-referenced examination was selected.

Examination items were designed to measure the acquisition of program-level skills, not course-level knowledge. The Outcomes Assessment Committee selected a representative sample of skills to be tested on each examination. This sample was approved by faculty ballot to be representative of the major outcome statement goals for each year of the curriculum. P1 examinees were tested on P1 abilities; P2 examinees were tested on P1 and P2 abilities; P3 examinees were tested on P1, P2, and P3 abilities; and P4 examinees were tested on P1, P2, P3, and P4 abilities. All faculty members were invited and participated in item writing, editing, verification, cut score development, examination proctoring, and score distribution. A test blueprint, grading table, and mastery specifications were developed (Table 3).

All faculty members were instructed about the methodology and techniques for writing effective test items. Materials were drawn from a variety of sources, most notably the Case and Swanson[8] and Haldyna studies.[9] Each faculty member was assigned to a group headed by a member of the Outcomes Assessment Committee with sessions held to elucidate item-writing procedures for this examination. Specific item-writing guidelines were developed and are available upon request from the authors.

Each item writer developed 8 to 12 items, and forwarded their completed items to the group leader to edit for adherence to the item writing standards, as well as grammar, spelling, and formatting. The completed items were organized by TABO for inclusion in the question bank. Faculty development included training in how to write higher-order test items using the cognitive levels of Bloom's taxonomy.[15]

The Outcomes Assessment Committee approved the test items, associating each with a TABO, and using a modified Angoff system, calculated a minimal competency cut score for each test item. Two separate panels of 10 faculty members reviewed each test item and estimated the percentage of 100 minimally competent students who would answer the question correctly. For example, P1 test items were scored with the panel envisioning students who had successfully completed the P1 year. The high and low estimates were dropped, and the average of the remaining scores was determined to be the mastery cut score for that test item.[16]

The average cut score for the actual items included on the examination was used as the mastery score percentage for the corresponding group of items. The weighted average percentage for all questions was used as the mastery score for the entire examination. The percentage Angoff score was multiplied by the point value of the examination to create the cut score in points. Each version of the examination had a unique cut score based on the actual test items used to assemble the examination. A student raw score above the Angoff cut score was defined as *mastery*. A student raw score that was less than 1 correct answer below the Angoff cut score was defined as *partial mastery*. In addition, a student raw score that was more than 1 correct answer below the Angoff cut score was labeled *non-mastery*. All items were assigned a point value of 1. Between 8 and 14 items were included for each TABO. A unique examination was created for each grade year to address curriculum specific, grade-level skills for each class year. Each examination was offered in 2 versions created by scrambling answer choices. Care was given to diversify the items across a broad range of content and disease states. Test items were selected to approximate 50% to 60% difficulty in aggregate. A difficulty level of 57% to 67% appears to maximize the reliability of the examination.[17]

To ensure the reliability and validity of the examination over time and consistency in examination administration and test security, the following procedures were followed: all files were properly secured throughout the examination production cycle; multiple versions of each examination were created by scrambling questions to random order; a limited number of test booklets were printed and each was assigned a unique serial number; and a control sheet was created for each test room to identify the students expected to take the examination. The students signed their examination booklets, signed a cover page

Table 3. WUSOP 2008 Annual Skills Mastery Assessment (ASMA) Examination Test Blueprint

	Ability Set, Item Count			
Terminal Ability-based Outcome Statements	**P1 Examination,**	**P2 Examination,**	**P3 Examination,**	**P4 Examination,**
P1 Year				
Perform a selected pharmaceutical calculation	12	10	8	8
Cellular process essential to life	8	0	0	0
Structured and function of human anatomy	8	8	8	8
How systems function to maintain homeostasis	10	10	8	8
Explain the mechanism of action of a selected drug class	14	10	8	8
Use appropriate literature/resources to solve a problem	10	8	8	8
P2 Year				
Elements that influence drug absorption, metabolism, and excretion	0	8	8	8
Make an appropriate dosing adjustment	0	8	8	8
Create a patient care plan	0	0	8	8
Interpret a clinical lab	0	8	8	8
Evaluate a patient therapy	0	8	8	8
Recognize an adverse effect	0	8	8	8
Determine the correct application of law	0	8	8	8
P3 Year				
Communicate in Spanish	0	0	8	8
Create a problem list	0	0	8	8
Adjust the patient care plan	0	0	8	8
Evaluate therapeutic outcomes	0	0	8	8
P4 Year				
Resolve a patient problem	0	0	0	8
Develop an alternate course of therapy	0	0	0	10
Detecting adverse drug reaction	0	0	0	8
Recommended dosage adjustment	0	0	0	8
Differentiate disease states by symptoms	0	0	0	8
Total Item Count per Examination	62	94	128	170
Items Scored	61	93	127	169
Calculated Cut Score, %	56	55	51	52

describing the security policy, and signed their score sheet to verify their identity. They were allowed to write on their test booklets and used only school-assigned calculators.

The purpose of the examination was explained to all examinees, but they were not told which specific TABOs would be tested on the examination to eliminate the effect of focused test preparation on scores. Faculty members were advised not to discuss examination items with students before or after the examination to avoid the temptation to teach to the test. Test items were generated to cover the entire curriculum, not just coursework. As a general design principle, it was considered appropriate to include test items that were never covered by a lecture, to expect students to learn during their experiential education, to include novel stretch items in limited quantities, and to include items with a range of difficulty. All students were required to take the examination. (Samples of the test booklet instructions are available upon request from the authors.)

The score sheets were mechanically scored by LXR Test software, and both test and item statistics were reviewed. Poor test items were removed from the examination, and the examination was rescored. Items were considered poor if less than 25% of examinees answered the item correctly, the point-biserial correlation was negative, or the item was identified as containing an error. The point-biserial is the correlation between an item score and the total score on a test. Positive values indicate that the item differentiates between high-ability and low-ability examinees. The final Angoff cut scores were recalculated based on the updated point count for the test, and the grading table was adjusted to reflect these changes.

Student score reports were designed to create criterion-referenced formative value, eliminate the reporting of student class rank scores, and provide detailed subscores by TABO to inform curricular development. The individual mastery reports were printed and made available to students through their faculty advisor within 3 days of the examination. Each student had the opportunity to discuss their strengths and weaknesses with their advisors, and remediation or corrective plans for students who failed to demonstrate mastery were developed when appropriate. Although considered "high stakes," the annual assessment has not been used as a standalone barrier to student progress. Students failing to demonstrate mastery have been counseled by their advisor and the dean for student affairs regarding appropriate remediation.

Student performance data was analyzed by comparing the student raw score to the cumulative pharmacy program GPA achieved by the end of the spring 2008 semester, the class rank, the Pharmacy College Admission Test (PCAT) score, and prepharmacy GPA using EZanalyze 3.0 software (Timothy Poynton, Boston, MA) to plug into the spreadsheet software. Examination results were discussed by curriculum committee and outcomes assessment committee members. Student focus group and faculty member feedback was compared with actual examination results.

EVALUATION AND ASSESSMENT

For the spring 2008 administration of the examination, 235 out of 239 examinees demonstrated overall mastery, 2 demonstrated partial mastery, and 2 failed to demonstrate mastery of the overall composite skill sets. All P3 and P4 students demonstrated mastery, and only 1 P2 failed to demonstrate mastery.

The reliabilities as calculated by Cronbach's alpha were P1, $\alpha = 0.65$; P2, $\alpha = 0.81$; P3, $\alpha = 0.82$; and P4, $\alpha = 0.80$, which were above the minimum goal of 0.60. The examination statistics for 2008 are presented in Table 4. One item was removed from scoring on each version of the examination due to poor item statistics.

The Outcomes Assessment Committee documented compliance with examination development criteria (Table 2). The Pearson's correlation between examinee ASMA raw score and cumulative GPA ranged from $r = 0.533$ to $r = 0.659$, class rank ranged from $r = 0.513$ to $r = 0.653$, PCAT score ranged from $r = 0.198$ to $r = 0.403$, prepharmacy GPA ranged from $r = 0.137$ to $= 0.307$. The class rank for ASMA examination scores correlated to the class rank by GPA ranged from $r = 0.513$ to $r = 0.653$. In addition, all 2008 P4 students (51) passed their NAPLEX examination on the first attempt. Correlation results, including statistical significance for all examinations are reported in Table 4. Online faculty survey results (N = 16

Table 4. WUSOP Test Statistics for Spring 2008 Administration of the Annual Skills Mastery Assessment (ASMA)[a]

	Test Names and Dates[a]			
Variables	**P1 2008**	**P2 2008**	**P3 2008**	**P4 2008**
No. of Examinees	69	65	53	51
No. of Items	62	94	128	170
Maximum Points	61	93	127	169
High Score	53	83	114	159
Low score	29	38	78	116
Median	44	69	97	132
Mean	43.6	68.5	96	134
Standard Deviation	5.2	8.4	9.8	10.2
Test Reliability (Chronbach's α)	0.7	0.8	0.8	0.8
Standard Error of Measurement	3.1	3.7	4.2	4.6
Standard Error Divided by Max Points	0.05	0.04	0.03	0.03
Cut Score Points by Points Earned (CSP)	35	51	64	87
ASMA Raw Score to Cumulative GPA (p value)	0.533 (< 0.01)	0.659 (< 0.01)	0.563 (< 0.01)	0.598 (< 0.01)
ASMA Raw Score to Class Rank (p value)	0.513 (< 0.01)	0.653 (< 0.01)	0.542 (< 0.01)	0.592 (< 0.01)
ASMA Raw Score to PCAT (p value)	0.396 (< 0.01)	.255 (< 0.04)	0.403 (< 0.01)	0.198 (< 0.16)
ASMA Raw Score to PrePharmacy GPA (p value)	0.265 (< 0.03)	0.305 (< 0.01)	0.137 (< 0.33)	0.307 (< 0.03)
ASMA Score Rank to Class Rank by GPA (p value)	0.555 (< 0.01)	0.701 (< 0.01)	0.568 (< 0.01)	0.582 (< 0.01)

Abbreviations: ASMA = Annual Skills Mastery Assessment Exam; GPA = grade point average; PCAT = Pharmacy College Admissions Test
[a] All tests administered on March 26, 2008.

Table 5. Results of Anonymous Online Faculty Survey Regarding the Annual Assessment Examination, (N = 16)[a]

Question	Mean (SD)[b]
The annual assessment is a useful tool for students to identify areas of strength and weakness	4.6 (0.5)
The annual assessment is a useful tool for the school to identify problems with the curriculum	3.7 (0.7)
I was clear about my role in the Spring 2008 assessment process (writing questions, Angoff scoring, explaining results to students)	4.4 (0.6)
The outcomes assessment committee provided adequate help for faculty throughout the Spring 2008 process	4.3 (0.6)
The results of the annual assessment were available to faculty and students in a timely manner	4.6 (0.5)
I agree with the policy of distributing the results to students through their academic advisor	4.6 (0.5)
In the future, the results of the annual assessment examination should be used to make decisions on academic progression to the next year	3.3 (1)
Overall this Spring 2008 assessment process was an improvement over previous years	4.0 (0.7)
The students score report used in spring 2008 was an improvement over older versions of the score report	4.2 (0.8)
Students are more likely to understand their strengths and weaknesses by reviewing the Spring 2008 students score report than they were reviewing previously used score reports	3.8 (0.8)

[a] Twenty-three faculty members were sent surveys and 16 responses were received.
[b] Items were rated using the following scale: 1 = strongly disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree

out of 23) indicated strong faculty support for the assessment examination. Over 87% of faculty members responded agree or strongly agree to 9 different quality measures surveyed (Table 5). All students received a comprehensive student score report and had the opportunity to meet with their faculty advisors to review results. (A sample student score report is available from the author.) Approximately 90% of P1 - P3 students met with their advisor. One hundred fifty-nine (out of 201) students supported the program and expressed a strong intent to use the formative data generated to improve future performance (Table 6). Separate reports provided a snapshot of the performance of each set of examinees in a mastery summary report. This data was presented to a curricular summit of all faculty members in May 2008 and used to revise program goals and course-level objectives. Areas identified as needing review were medical Spanish, biomedical informatics, pharmacokinetics, and patient problem-solving skills.

DISCUSSION

By verifying that students have developed skills, PharmD programs may logically infer that graduates will

Table 6. Results of an Anonymous Online Survey of Pharmacy Students Regarding the Annual Assessment Examination, N = 159[a]

Question	Mean (SD)[b]
The annual assessment is a useful tool for students to identify areas of strength and weakness	3.9 (1)
The annual assessment is a useful tool for the school to identify problems with the curriculum	3.9 (1)
The results of this Spring 2008 assessment were available to faculty and students in a timely manner	4.2 (0.8)
I agree with the policy of distributing the results to students through their academic advisor	4.1 (0.9)
In the future, the results of the annual assessment examination should be used for making a decision on academic progression to the next year	2.4 (1.3)
Overall this Spring 2008 assessment process was an improvement over previous years	3.5 (0.9)
I was well-informed by the school about the process for obtaining my assessment results	4.2 (0.9)
My advisor or another faculty member acting on his or her behalf took the time to explain the results to be adequately	4.2 (0.9)
I understand the results of my annual assessment examination taken Spring 2008	4.4 (0.7)
The students score report used in spring 2008 was helpful in identifying my areas of strength and weakness	3.9 (1)
Before leaving campus for summer break in May 2008 I have enough information about my strengths and weaknesses to create an action plan for self-improvement before beginning the fall 2008 semester	3.6 (0.9)
Based on the detailed information presented in the students score report I plan to make changes to improve my areas of weakness	3.6 (1)

[a] Due to delays in review board, approval to survey was not sent until October, at which time 50 P4 students had graduated and were no longer enrolled or responding to survey e-mails. The total N in the survey software was 254. The number of accurate e-mail addresses delivered was 201 and the number of respondents was 159.
[b] Items were rated using the following scale: 1 = strongly disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree

develop ability and become competent. Because the majority of pharmacy program graduates (> 90%) nationwide are certified minimally competent by the NAPLEX licensing examination, this may be true.[18]

WUSOP's terminal ability-based outcomes statements mirror the knowledge, skills, and abilities defined by ACPE Standards 2007. The curriculum was designed to develop these skills into abilities that will translate into competent graduates. The multiple-choice testing format is appropriate for measuring mental skills or abilities when there is clearly a right answer to the problem posed. Skills or abilities requiring creative answers would be better assessed by a constructed response test format. Physical skills or abilities are not appropriate for multiple-choice testing.[19]

Reliability can be defined as "the extent to which measurements resulting from the test are the result of characteristics of those being measured."[20] Reliability combines the characteristics of both the examination and a pool of examinees. The test itself, the characteristics of the students taking the test, and the nature of the scoring can all introduce error. The internal consistency measurement, Cronbach's alpha, was chosen due to the inherent time constraints of conducting a test-retest analysis, and the potential errors introduced by split-half, or alternate form reliability studies. Because every assessment score contains an error of measurement, it is impossible to say with certainty that any individual's observed score accurately mirrors the true score. Our goal, therefore, is to estimate the standard error of measurement and then use that value to estimate the probability that an observed score is reasonably close to the true score[21] (Table 4). The observed calculations for the ASMA examinations are reasonably reliable (> 0.65), and the standard error of measurement as a percentage of examination points ranges from 2.7% to 5%, which suggests that at least 95% of the observed score is not due to random error.

The literature suggests that an alpha of 0.60 is reasonable for a course examination but may need to be as high as 0.95 for a high-stakes credentialing examination.[20] Neither reliability nor validity is considered a property of a test or a test score; rather both are properties of the use and interpretation of test scores.[22] Depending upon the use of the examination results, evidence of validity should differ. Because this examination was designed on a criterion-referenced model to provide useful individual student feedback as a referendum on curricular quality, the validity evidence required was less strict than if we had been trying to establish a norm-referenced standard among different groups of students. In this instance, it was sufficient to show that students were mastering properly defined educational outcomes skills, demonstrating the precursors of professional competence, and were receiving valuable, formative feedback to improve individual performance prior to graduation.

WUSOP sought to optimize examination validity by developing examination content under strict controls: sampling content carefully for inclusion on the examination; setting appropriate standards for the pool of examinees; adhering to proper item writing guidelines; not using alternate test forms; using the same test form for all students in the same grade year; requiring all students to take the examination; creating examinations tailored for each grade year under tight process controls; mechanically scoring the multiple-choice examination; and conducting stringent item analysis to remove bad items from the question bank. Additionally, validity was enhanced by not allowing construct-irrelevant variance to creep into the process by introducing scoring errors, student cheating, guessing penalties, unethical test preparation, inconsistent test administration, or rater unreliability. Student anxiety was minimized by deemphasizing test preparation and student-student comparison reports. Test fatigue was minimized by administering the test during a week with no other scheduled examinations and cancelling all classes on testing day. Motivation to take the test was enhanced by the lack of punitive consequences attached to the examination and faculty support of the formative nature of the score reporting.

Validity evidence has traditionally been grouped into content, criterion, and construct-related evidence. There are no rigorous distinctions between these categories, and they are not distinct types of validity.[23] Criterion-related validity evidence refers to the hypothesis that test scores are systematically related to one or more outcome criteria.[23] The strong correlations between student raw scores and their cumulative GPA and class rank in pharmacy school, suggests that the Annual Skills Mastery Assessment examination has criterion-related validity (Table 4). Content-related validity evidence supports the hypothesis that test items represent the skills in the specified subject area,[23] and the WUSOP examination's rigid adherence to proper test design suggests that the examination has content validity. Construct-related validity evidence supports the hypothesis that the test measures the right psychological constructs.[23] Evidence can take several forms. One approach is to demonstrate that the items within the examination are interrelated and therefore measure a single construct. Because the WUSOP TABOs were developed directly from CAPE outcomes statements developed after expert panel analysis of the constructs that generate competence in pharmacy, a rudimentary compliance with construct validity is suggested. Another element of construct validity evidence would be a measurement that

demonstrates the test behaves as expected. Because all P4 students who took the ASMA examination in 2008 demonstrated mastery of the skills tested, and passed their NAPLEX board examinations which certify competence, additional evidence of construct validity can be inferred. In addition, the students who failed to demonstrate mastery (N = 4) had been identified by our traditional academic review process as students at risk.

The raw score a student achieves reflects not only ability, but effort within the group of examinees. No statistical manipulation of data can alter the fact that any assessment can provide only probabilistic inference about causation or future performance. A well-crafted examination can be rendered invalid if the scorers are misinterpreted and consequences misapplied. Szilagyi reported significant challenges in getting students to perform well on their initial attempts at the Milemarker examination.[24] The design of a student-centered examination gives WUSOP students a vested personal stake in honestly performing to their capabilities. The students understand that the score report provides a snapshot of their performance in the curriculum, and the score will be meaningful only if they actively attempt to perform well.

The advantages of developing this ASMA tool were low-cost, rapid, and flexible reporting; absence of inter-rater reliability problems; and the likelihood of success, given our institution's culture and resources. A locally developed examination can increase curricular structure, develop more consistency, improve linkage between courses, improve question writing and examination construction, and create more opportunities for students to apply knowledge.[25]

The chief concerns with this examination format are the item sampling error, the dependence upon quality test-item writing procedures, and the potential for distortion of the intended curriculum if the examination results were misinterpreted. Creating a high-stakes examination that rewards rote memory could unintentionally subvert the true educational mission of the program.[26,27] Roediger suggests lower-order knowledge-based examinations may actually cause false knowledge to be implanted in students.[28] Lack of faculty members, insufficient student engagement, and inappropriate use of the results have been identified as barriers to success with this type of examination.[29]

There is an inherent difficulty in categorizing test items by Bloom's Taxonomy level. The cognitive processes required to answer a question are as dependent on the background of the test taker as they are on the question content. Students with more practice experience or advanced prepharmacy education may simply recall an answer with little or no conscious thought, whereas students testing this material for the first time may need to reason the answer from basic principles.[8] In other words, a P4 student may use a different cognitive skill to answer a question than a P1. Potential improvements include: rate each test item with an Angoff cut score for each class level; ensure the spread of case examples across all major drug classes and disease states; improve the test items based on statistical analysis; and develop better evidence around construct and predictive validity.

The cost for the software package is approximately $3600. The estimated total work hours devoted to the preparation, administration, and review of the ASMA examination was 540 faculty hours, or roughly 20 hours per faculty member, including the OAC committee members' work. In addition, the assistant dean for assessment devoted approximately 240 hours to developing and managing the process. Both estimates are spread over the school year, with a disproportionate share of hours invested between January and March. A one-time cost for the assistant dean of assessment to attend 3 days of advanced training on the software was $2500.

It is critical to the success of an annual assessment program that faculty members embrace the examination process. To do so they must feel confident that the program has value and creates useful data. Therefore, it is important to create a detailed faculty score report and thoroughly explain the results to the faculty members.

SUMMARY

Analysis of the WUSOP Annual Skills Mastery Assessment suggests it is a reasonably reliable and reasonably valid tool to provide formative student feedback, inform curricular improvement, and comply with ACPE accreditation guidelines to continuously improve the curriculum.

ACKNOWLEDGMENTS

The authors wish to acknowledge the faculty and staff of Wingate University School of Pharmacy for the gracious use of their time and energy in support of our assessment program, especially the members of the Outcome Assessment Committee, both past and present. The assessment tool was awarded the AACP Award for Excellence in Assessment in July 2009. In addition, the authors thank Dean Robert Supernaw of Wingate University and Katherine Kelley of The Ohio State University for providing valuable review and guidance on this article.

REFERENCES

1. Accreditation Council for Pharmacy Education. Accreditation Standards and Guidelines for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree. http://www.acpe-accredit.

org/pdf/ACPE_Revised_PharmD_Standards_Adopted_Jan152006.pdf. Accessed April 20, 2010.
2. Plaza CM. Progress examinations in pharmacy education. *Am J Pharm Educ.* 2007;71(4):Article 66.
3. Haladyna TM. *Writing Test Items to Evaluate Higher Order Thinking.* Needham Heights, MA: Allyn & Bacon, A Pearson Education Company; 1997.
4. American Educational Research Association. Position statement on high stakes testing in pre-K - 12 education. http://www.aera.net/?id=378. Accessed April 20, 2010.
5. Downing SM, Haladyna TM. Test item development: validity evidence from quality assurance procedures. *Appl Meas Educ.* 1997;10(1):61-82.
6. Supernaw R, Mehvar R. Method for the assessment of competence and the definition of deficiencies of students in all levels of the curriculum. *Am J Pharm Educ.* 2002;66(1):1-4.
7. Mehvar R, Supernaw R. Outcome assessment in a PharmD program: the Texas Tech experience. *Am J Pharm Educ.* 2002;66(3):243-253.
8. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. *Natl Board Med Examiners* http://www.nbme.org/publications/ Accessed April 20, 2010.
9. Haladyna T, Downing S, Rodriguez M. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ.* 2002;15(3):309-334.
10. Kehoe J. Writing multiple-choice test items. *ERIC Clearinghouse on Assessment and Evaluation, The Catholic University of America, Department of Education, O'Boyle Hall, Washington, DC* http://www.eric.ed.gov/ED398236. Accessed April 20, 2010.
11. Kehoe J. Basic item analysis for multiple-choice tests. ERIC Clearinghouse on Assessment and Evaluation, The Catholic University of America, Department of Education, O'Boyle Hall, Washington, DC http://www.eric.ed.gov/ED398237. Accessed April 20, 2010.
12. Green K, Sax G. Test reliability by ability level of examinees. Annual Conference of the National Council on Measurement in Education. April 11-17, 1981; Los Angeles, CA.
13. Taylor AK. Violating conventional wisdom in multiple choice test construction. *Coll Stud J.* 2005;39(1):141-148.
14. Bond LA. Norm- and criterion-referenced testing. ERIC Clearinghouse on Assessment and Evaluation, 210 O'Boyle Hall, The Catholic University of America, Washington, DC. http://www.eric.ed.gov. Accessed April 20, 2010.
15. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. *Taxonomy of Educational Objectives, Handbook 1: The Cognitive Domain.* New York, NY: Longman; 1956.
16. Impara JCP, Barbara S. A comparison of cut scores using multiple standard setting methods. Annual Meeting of the American Educational Research Association. April 24-28, 2000. New Orleans, LA.
17. Feldt LS. The relationship between the distribution of item difficulties and test reliability. *Appl Meas Educ.* 1993;6(1):37-48.
18. National Association of Boards of Pharmacy on February 24, 2009. Statistical Analysis of Naplex Passing Rates for First-Time Candidates Per Pharmacy School from 2004 to 2008. National Association of Boards of Pharmacy. http://www.nabp.net/. Accessed April 20, 2010.
19. Rudner LM, Schafer, William D. Reliability. ERIC Clearinghouse on Assessment and Evaluation. http://www.eric.ed.gov/ED458213. Accessed April 20, 2010.
20. Gardner E. Five Common Misuses of Tests. ERIC Clearinghouse on Assessment and Evaluation. http://www.eric.ed.gov/ED315429. Accessed April 20, 2010.
21. Haladyna T. Roles and importance of validity studies in test development. In: Downing S, Haladyna T, eds. *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates; 2006: 739-755.
22. Kane MT. Content-related validity evidence in test development. In: Downing S, Haladyna T, eds. *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates; 2006: 131-150.
23. Brualdi A. Traditional and Modern Concepts of Validity. ERIC Clearinghouse on Assessment and Evaluation, 1129 Shriver, University of Maryland, College Park, MD. http://www.eric.ed.gov/ED435714. Accessed April 20, 2010.
24. Szilagyi JE. Curricular progress assessments: the Milemarker. *Am J Pharm Educ.* 2008;72(5):101.
25. Banta TW, Schneider JA. Using locally developed comprehensive examinations for majors to assess and improve academic program quality. Paper presented at the 70th Annual Meeting of the American Educational Research Association. April 16-20, 1986; San Francisco, CA.
26. Newble DI. The effect of assessments and examinations on the learning of medical students. *Med Educ.* 1983;17(3):165-171.
27. Abate MA, Stamatakis MK, Haggett RR. Excellence in curriculum development and assessment. *Am J Pharm Educ.* 2003;67(3):Article 89.
28. Roediger HL III, Marsh EJ. The positive and negative consequences of multiple-choice testing. *J Exp Psychol.* 2005;31(5):1155-1159.
29. Banta TW. Moving assessment forward: enabling conditions and stumbling blocks. *New Dir High Educ.* 1997;25(4):79.