RESEARCH

Development of a Summative Examination with Subject Matter Expert Validation

Ashley N. Castleberry, PharmD, MAEd,[a] Eric F. Schneider, PharmD,[b] Martha H. Carle MEd, MPH,[c] Cindy D. Stowe, PharmD[d]

[a] University of Arkansas for Medical Sciences College of Pharmacy, Little Rock, Arkansas
[b] Wingate University School of Pharmacy, Wingate, North Carolina
[c] University of Arkansas for Medical Sciences Office of Educational Development, Little Rock, Arkansas
[d] Sullivan University College of Pharmacy, Louisville, Kentucky

Submitted January 12, 2015; accepted March 17, 2015; published March 25, 2016.

Objective. To describe the development, implementation and impact of a summative examination on student learning and programmatic curricular outcomes.

Methods. The summative examination was developed using a systematic approach. Item reliability was evaluated using standard psychometric analyses. Content validity was assessed using necessity scoring as determined by subject matter experts.

Results. Almost 700 items written by 37 faculty members were evaluated. Passing standards increased annually (45% in 2009 to 67% in 2014) as the result of targeting item difficulty and necessity scores. The percentage of items exhibiting discrimination above 0.1 increased to 100% over the four years. Necessity scores above 2.75 out of 4 increased from 65% to 100% of items over six years of examination administration.

Conclusion. This examination successfully assessed student and curricular outcomes. Faculty member engagement observed in this process supports a culture of assessment. This type of examination could be beneficial to other programs.

Keywords: summative examination, examination development, assessment, subject matter experts

INTRODUCTION

The use of summative assessment in professional pharmacy programs has multiple benefits, including providing a measure of student learning around a specific set of curricular outcomes. This measurement can inform curricular improvements and plans for individual student improvement. The development and implementation of reliable and valid summative assessments can also advance the skill set of faculty members around measurement of student learning and improvement of all assessment activities.

In professional curriculum, assessment exists along the "formative – summative" continuum to enhance student learning and improve curricular outcomes. Assessments used for learning are termed "formative," while assessments of learning are termed "summative."[1,2] Summative assessment exists on the spectrum of low-to-high stakes testing. This level of impact is based on the outcome the assessment has on an individual or program. Ultimately for professional pharmacy programs, the North American Pharmacy Licensure Exam (NAPLEX) serves as the highest stake summative assessment for pharmacy graduates; this singular examination stands between doctor of pharmacy (PharmD) graduates and active participation in the practice of pharmacy. Previous publications describe the use of summative assessments to gauge overall knowledge across the pharmacy curriculum.[3-5]

In the Accreditation Council for Pharmacy Education's (ACPE) Standards 2016, a greater focus is placed on assessment and in particular the use of summative assessment to measure student learning and curricular success.[6] In addition to ACPE's focus on assessment, the 2013 revised Educational Outcomes from the Center for the Advancement of Pharmacy Education (CAPE) provides targets for professional degree programs to aim for.[7] The four domains and 15 subdomains of the CAPE Outcomes inspire both summative and formative assessment.

In 2006, the University of Arkansas for Medical Sciences (UAMS) College of Pharmacy (COP) wanted to

Corresponding Author: Ashley Castleberry, 4301 West Markham St., Slot #522-4, Little Rock, AR 72205. Tel: 501-686-8383. Fax: 501-296-1168. E-mail: ancastleberry@uams.edu.

create a summative assessment of the foundational knowledge of the professional program that could be used as a high-stakes progression examination. The college has a traditional pharmacy curriculum with stand-alone foundational science courses, biomedical and pharmaceutical-based, as well as clinical, social, administrative, and behavioral science courses. Delivery of the curriculum is primarily lecture-based with hands-on laboratories, problem-based learning, and team-based learning integrated into the didactic curriculum. Course-level assessments of learning occur primarily using computer-based electronic testing with multiple-choice and fill in the blank testing format. Additionally, the objective structured clinical examination (OSCE) technique is used strategically in some courses for both formative and summative assessment. Introductory pharmacy practice experiences (IPPEs) in community and institutional practice are offered in the first three years of the curriculum to complement learning.

The intent of this study was to measure student learning and retention of core curricular comprehension as well as to provide a measure of metacognition.[8] The purpose of this paper is to describe the development and implementation of a summative examination along with the measures used to show its validity and reliability and, secondarily, to examine the impact of this type of assessment on the culture of the college and on resource utilization.

METHODS

Summative assessment is one method used by the college to assess student learning across the curriculum. The summative examination described here is designed to reflect learning in the first two years of the curriculum and to help identify students at risk of poor performance as they progress through the program. An overview of the process is shown in Figure 1. To begin developing the examination, a blueprint based on the curriculum map was developed by the assessment committee and approved by the faculty members. Course content was grouped into three domains to match the four domains of Appendix 1 in Standards 2016: biomedical sciences, pharmaceutical sciences, and a combination of clinical with social, administrative, and behavioral sciences.[3] The summative examination is administered online at the end of the second professional year. It consists of 75 multiple-choice questions (MCQ) based on program-wide competencies and course weighting in years one and two and 35 short-answer and matching questions relating to the top 200 drugs (trade/generic names and therapeutic class). Faculty members teaching in the first two years of the curriculum were asked to provide questions

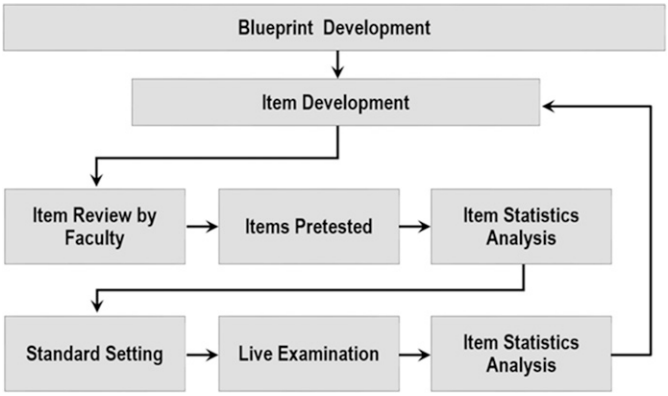

Figure 1. Development and Administration of the Summative Examination.

that were concept-based, linked to course objectives, and avoided trivial detail. The intent behind the summative examination is to test high-level concepts as retained knowledge rather than the detailed information tested within the course term. Faculty members were asked to provide questions that would not be used for assessment within the course. Next, the assessment committee, in sessions open to all faculty members, reviewed and edited all questions to follow a basic single best answer, multiple-choice format with 3-4 reasonable distractors. Information about writing quality MCQs and a checklist for assessing item construction, quality of the question stem, the answer options and the correct answer were provided to reviewers (Table 1). Items from the initial item reviews were pretested on second-year students prior to high-stakes administration the subsequent year. The pretest consisted of 125-item examinations (2 versions) using 214 unique items. Roughly twice the number of questions required by the blueprint was used to allow for loss of items as a result of the item analysis review process.

Each subsequent year, the high-stakes progression examination was formulated with scored items selected based on their performance statistics, standard setting analytics, and blueprint criteria. Classical psychometric statistics (difficulty and discrimination factors) were used to assess individual item performance and the Kuder-Richardson 20 (KR-20) statistic was used to determine performance reliability. Item difficulty (p+) was defined as the percentage of students answering the item correctly. For item discrimination, the point biserial (pbi), was used. The pbi is the Pearson correlation between item score and the overall examination score. New pretest items were submitted, processed, and selected for inclusion each year based on review of the item pool. Students were not made aware of which items were being pretested and which were actually being scored. The need for

Table 1. Item Review Checklist (Adapted from Haladyna et al[13])

Overall Question
Is the item a complex MCQ format (K-type, All of the Above, None of the Above)?
If yes, can it be rewritten?
Is the item written with good grammar, punctuation, and spelling?
Is there extraneous or unnecessary information included?
Does the item test ONE competency?
Is the item general enough (ie, NOT minutia) and defensible?
Stem
Is the stem written in question or completion form?
Is the wording clear and focused?
Is there teaching in the stem?
Is there NEGATIVE phrasing? If yes, can it be rewritten?
Answer Options
Are the options in logical or numerical order?
Is the content homogenous?
Is the length of the options fairly consistent? Does any one option stand out?
Are all the options plausible?
Are any options overly specific or overly general?
Are there any specific determiners (eg, "never" or "always")?
Correct Answer
Is there only one correct answer?

additional pretest items were identified each year based on postexamination review of item statistics, recent curricular revision, and general item-bank vitality.

Beginning in 2010, students were asked to complete an optional test-taking questionnaire during the examination aimed at assessing metacognitive skills focusing on students' awareness of their own comprehension.[8,9] The first section of the questionnaire asked students to identify up to 10 items they felt certain they had answered incorrectly. Students were asked to choose if the question was difficult to understand or too detailed. Space was provided for additional comments on why they felt they answered incorrectly. Data collected from the metacognitive questionnaire were used in the SE item pool review process.[8] The summative assessment project was deemed exempt from review by the UAMS Institutional Review Board.

The passing standard was set using a modified two-round iterative, modified Angoff procedure.[10] This process used faculty and nonfaculty pharmacist practitioners who served on the assessment committee as the subject matter experts (SME) and rated each item. Collectively, the SMEs who contributed to the standard setting comprised a diverse group ranging from generalist clinicians to basic scientists, all of whom were familiar with the expectations of student performance in the program. These item ratings ultimately determined the passing score for the examination. Prior to conducting the modified Angoff procedure, a training session was held to discuss the standard setting process. Pivotal to the modified Angoff procedure is an expectation among all SMEs of the ability level of the student who would be deemed just barely acceptable (ie, "minimally competent").[10] To generate a consensus, the group discussed the expected level of ability for a minimally competent rising third-year (P3) student. The group was instructed and frequently reminded during the process to use this level of ability when making ratings during the modified Angoff process. It is important to note that minimally acceptable did not equate with an average student.

Participants rated items on necessity, difficulty, and their estimate of the percentage of minimally acceptable rising P3 students (as defined in the consensus discussion) who would correctly answer the question. For the purpose of the examination, necessity was defined as the importance of the knowledge and skill to the safe and effective practice of pharmacy. To rate necessity, a 4-point scale was developed (4=essential/critical, 3=important, 2=useful, 1=not necessary). During the training for the modified Angoff procedure, the meaning of the necessity scores was discussed among SMEs to allow a more standardized use of the scale. Difficulty was rated on a 5-point scale (5=considerably harder, 4=slightly harder, 3=appropriate, 2=slightly easier, 1=considerably easier). Finally, each item was rated based on the percent of minimally acceptable students beginning the third year expected to answer the item correctly. For each item, the SMEs provided an initial rating, discussed the rationale for the rating among the group, and provided a second rating after the feedback from the group. Inclusion of the second round was designed to reduce group variability and promote a consensus.[10] The passing standard range was computed from the second round of SME ratings. The passing score could be determined using the SME assessment of the likelihood that a group of minimally acceptable students would answer the item alone or by weighting the score based on necessity and/or difficulty ratings (Figure 2). This resulted in three possible passing standards. Additionally, the assessment committee adopted standards for item performance (to assure item quality) and average necessity score (to assure content validity) as a means of assuring this goal.

The final step in the summative examination process was providing feedback to the students. For purposes of item security, students were not allowed to review specific examination questions. However, after administering the examination, all students were given a customized report card outlining their performance overall, on the course-based and Top 200 questions, and on Standards 2016 Appendix 1 domains (Figure 3). The director of

Step 1: Collate the round two SME Necessity (N), Difficulty (D), Probability of Success (PS), PS*N, and PS*N*D ratings by test question.

Step 2: For each test question, calculate the average of the N (aN), D (aD), PS (aPS), PS*N (aPS*aN), and PS*N*D (aPS*aN*aD) ratings

Step 3: Perform the Pass Point Calculations

$$PS\ only = \frac{\Sigma\ aPS}{number\ of\ items}$$

$$Weighted\ for\ Necessity = \frac{\Sigma\ (aPS * aN)/Number\ of\ items}{\Sigma\ aN/Number\ of\ items}$$

$$Weighted\ for\ Necessity\ and\ Difficulty = \frac{\Sigma\ (aPS * aN * aD)/Number\ of\ items}{\Sigma\ (aN * aD)/number\ of\ items}$$

Figure 2. Passing Standard Calculations.

assessment and/or associate dean for administrative and academic affairs conducted individual student meetings with students with unsatisfactory scores to develop a plan for improvement. Failure on the first attempt warranted a retake examination after remediation and a 30-day wait period. Failure on the second attempt warranted additional remediation and retake of the examination after 30 days with referral to the scholastic standing committee for consideration of progression if the third attempt is failed. During examination development, faculty authors received reports on the performance of their items. Subsequently, individual item performance was reviewed with faculty members for consistency over time.

RESULTS

The development of this summative examination took approximately 18 months, from 2007-2008, and included blueprint development, item writing, peer item review, item mapping, item pretesting, item analysis, standard setting, and formulation of the first high-stakes summative examination. The pilot phase was conducted in spring 2008 with high stakes administration for rising P3 students in 2009. The implementation of the examination resulted in a database of items with student performance data and modified Angoff metrics determined by SMEs. Keeping the image of the "minimally competent" student and the subsequent estimation of the percent expected to answer the item correctly proved challenging for the SMEs.

Allowing participants to explain their rankings, particularly when they were outliers, and the ensuing group discussion provided perspective and more group consensus. As of this writing, almost 700 items written by 37 faculty members were tested for this examination. In addition to producing items for inclusion on future examinations, the process of implementing such an examination contributed to changes in the culture of assessment at the college. Tangible results were observed such as increased attention to examination construction and maintenance across the curriculum without administrative directives, new assessment measurements throughout the curriculum, and improved faculty understanding of item performance data.

The passing standard of the summative examination calculated based on the SME rankings from the modified Angoff process increased over time (Table 2). A conservative approach was adopted using the lowest of the calculated passing standards depicted in Figure 2. The passing standard of the first scored implementation of the examination (2009) was 45%. This number gradually increased to 67% for administration in 2014. The reliability of the summative examination was calculated each year using the KR-20. This number consistently ranged from 0.71 to 0.78. Table 2 shows the passing standard and KR-20 of the examination each year.

Item statistics were collected on each item to determine inclusion on future examinations. Statistics included discrimination (pbi) and difficulty (p+). Necessity scores generated from SMEs were evaluated as was student input on metacognitive forms. This information was compiled to assess overall item quality.

As items were pretested, one of the criteria for inclusion on future examinations was discrimination above 0.10. While discrimination values above 0.10 would be more ideal, the assessment committee chose the pbi threshold to strike a balance between maintaining as robust an item pool as possible while still having reasonably discriminating items.[11] As the number and quality of items in the database increased over the years, the percentage of items on the examination exhibiting this characteristic increased to 100% on each examination given 2011-2014. Table 3 shows the mean discrimination score, range, distribution of scores at various cut points, and the percentage of items on the examination having a discrimination score above 0.1 for each year.

For the summative examination, our desired range for difficulty (p+) was greater than 0.5. Since a summative examination should measure general concepts required for progression to the next academic year, we believe that more than 50% of the class should get each item correct. Table 3 shows the mean difficulty score,

UAMS COLLEGE OF PHARMACY
UNIVERSITY OF ARKANSAS FOR MEDICAL SCIENCES

Summative Examination Score
Spring 2011

STUDENT NAME

	# of Questions	Your Exam	Total Group (n=116) Average	Range
Overall Score	110	75%	83.1%	(67.0% - 97.0%)
Passing Score		64%		
Top 200 Drugs	35	86%	91.2%	(54.0% - 100.0%)
Provide Generic Name	11	82%	87.6%	(36.4% - 100.0%)
Provide Brand Name	11	73%	91.6%	(36.4% - 100.0%)
Identify Drug Class	13	100%	94.2%	(69.2% - 100.0%)

On the Top 200 Portion of the exam you met the minimum acceptable standard (for this portion of the exam 80%). It is important that you maintain your knowledge of the Top 200 drugs (brand and generic names along with their drug class). The expectation over the third year is that you will know all of this information along with indications, dosage forms, and available strengths. The Class of 2012 are being required to take a Top 200 test prior to beginning APPEs, and we anticipate the same will occur next year.

Core P1 and P2 Courses	75	71%	79.3%	(64.0% - 97.0%)
Content Breakdown*				
Biomedical Sciences (52% of Section)	39	69%	78.4%	(59.0% - 100.0%)
Pharmaceutical Sciences (19% of Section)	14	71%	79.7%	(42.9% - 100.0%)
Clinical/Behavioral Sciences (29% of Section)	22	73%	80.5%	(54.6% - 100.0%)

On the Core P1 and P2 courses: Biomedical Sciences portion of the exam you met the minimum acceptable standard (for this portion of the exam 54%). It is important that you maintain your knowledge of the Biomedical Sciences.

On the Core P1 and P2 courses: Pharmaceutical Sciences portion of the exam you met the minimum acceptable standard (for this portion of the exam 51%). It is important that you maintain your knowledge of the Pharmaceutical Sciences.

On the Core P1 and P2 courses: Clinical and Behavioral Sciences portion of the exam you met the minimum acceptable standard (for this portion of the exam 55%). It is important that you maintain your knowledge of the Clinical and Behavioral Sciences.

*Content Breakdown:
Biomedical Sciences: Anatomy/Physiology/Pathology, Biological & Cellular Chemistry, Principles of Drug Actions, Pharmacology, Medicinal Chemistry, and Molecular Biology/Biotechnology
Pharmaceutical Sciences: Pharmaceutics, Pharmacokinetics, and Clinical Pharmacokinetics
Clinical/Behavioral Sciences: US Healthcare, Career Orientation and Communications, Drug Information, Dispensing, Therapeutics I, Nonprescription Drugs, and Pharmaceutical Calculations

Figure 3. Sample Student Report Card.

range, distribution of scores at various cut points, and the percentage of items on the examination having a difficulty score above 0.5 for each year.

Necessity scores, as determined by SMEs, were used a marker of validity for the examination. These scores serve as an indicator of content validity, showing that the examination was assessing what it was meant to assess. Items with scores above 2.75 are ideal for inclusion on a summative examination that assesses minimal competency to progress to the next curricular year. Much like the discrimination cutoff, the threshold necessity score was chosen by the assessment committee to allow the highest possible level of necessity while still maintaining an adequate pool of examination questions. Figure 4 displays the percentage of items on the examination having a necessity score above 2.75 and how this increased each year (65%, 71%, 90.7%, 100%, 98.7%, and 100%, respectively).

Table 2. Overall Examination Performance

	Passing Standard %	KR-20
2009	45	0.76
2010	55	0.75
2011	64	0.78
2012	66	0.71
2013	67	0.73
2014	67	0.77

Item quality was assessed using item statistics, as well as student responses on metacognitive forms starting in the second year of the high stakes examination administration. When items did not perform as anticipated, they were reviewed to determine if a writing error had occurred that would account for the poor performance. Student metacognitive data were used to help determine the reason for self-identification of questions they believed were not answered correctly. The identification of missed items may have been based on their belief that the question was confusing or difficult to understand, too detailed, or simply because the student was not prepared for the item. In addition to checking one of these three options as reasons for missing the item, space for additional explanation of perceived incorrect questions was provided for students to write in their rationale. Triangulation of all available data (ie, student identification, item performance, and item necessity) determined further use of examination items and ultimately curricular revisions. This process also empowered students to contribute to the quality of the examination. Additional analysis of metacognitive results can be found in a previously published study.[8]

As a part of the metacognitive assessment by students in year 2 of the summative examination, 25 items were identified as being missed by at least 10 students. The most common reasons cited by students for their certainty that they missed an item were confusing item construction and content that was too detailed/specific. Items that students selected as having missed were reviewed. Twenty of the items identified by students as frequently missed were removed from the examination operational pool based on the psychometric criteria set up by the assessment committee.

Student input was important as new items were created. The metacognitive form continued to be used despite our experience with the high-quality items. Student comments predominantly addressed the pretest items for which there was no past performance data and items where curricular revision or change in faculty members resulted in content gaps. Inconsistencies were further investigated with faculty members to determine potential causes for change in student performance. This feedback was important to item development and commonly reported items were reviewed in more detail and discussed with authors before inclusion as a high-stakes test item for subsequent summative examination administrations.

Table 3. Summary of Item Analysis Data

	2009	2010	2011	2012	2013	2014
Discrimination (pbi) Mean (range)	0.19 (-0.07-0.42)	0.19 (-0.03-0.46)	0.17 (-0.15-0.48)	0.19 (-0.14-0.45)	0.17 (-0.07-0.39)	0.22 (-0.01-0.45)
≤0.09	18	22	27	24	27	16
0.1-0.15	16	15	17	19	16	8
0.15-0.25	39	31	31	27	31	35
≥0.25	27	32	25	31	27	41
% Examination Items[a] ≥0.1	84	79	100	100	100	100
Difficulty (p+) Mean (range)	0.67 (0.15-0.98)	0.76 (0.2-1)	0.79 (0.23-1)	0.77 (0.35-1)	0.76 (0.22-1)	0.76 (0.17-1)
≤0.59	35	22	13	11	13	20
0.6 to 0.69	15	8	11	17	19	13
0.7-0.79	14	14	20	28	23	16
0.8-0.89	19	29	28	23	21	20
≥0.9	17	27	28	21	24	31
% Examination Items[a] >0.5	77	90	97	98	93	97

[a]% Exam Items is calculated based on the number of items meeting our criteria for inclusion out of the total number of scored items on that examination

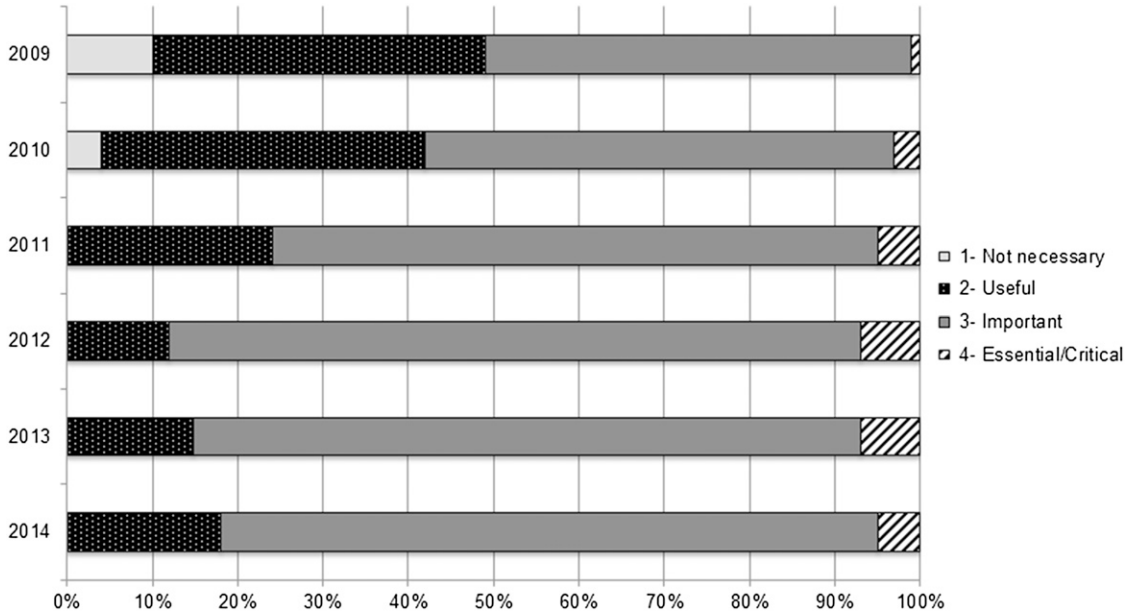

Figure 4. Necessity Scores by Year. Subject matter experts' second response to the following question for each item: To what extent is the knowledge, skill, or ability measured by this item necessary for minimally acceptable performance as a rising third-year student? (1=item is not necessary, 2=item is useful, 3=item is important, 4=item is essential or critical).

DISCUSSION

The development of a summative examination is an extensive process requiring continuous evaluation and revision. Implementation of this examination at the college brought expected and unexpected changes for both faculty members and students. Expected changes included the development of a databank of examination items that measured retention of foundational knowledge. Additionally, validity was assessed using necessity scoring as determined by SMEs. This marker of validity is not described in the literature for this particular use but we believe this method to be a useful surrogate marker of validity as faculty members and practitioners are experts on what students need to know in order to progress successfully through the curriculum.[12]

Unexpected changes during and after examination development arose from the goal of making better examination items in the review process. Numerous faculty development activities coincided with examination development and spilled over into formative and summative assessment in individual courses. The examination development process encouraged faculty members to understand assessment at a deeper level and resulted in program level changes in the culture of assessment. Development of this examination may have also added value to the overall curriculum by mapping all items to competency statements. Improvements for both the examination and curriculum was symbiotic with changes in the curriculum leading to change in examination performance as well as examination performance leading to change in the curriculum.

Development and implementation of a summative examination like this one can bring about culture change for students, faculty members, and the school. These types of examinations help students and faculty members focus on the most important pieces of the curriculum while highlighting the importance of knowledge retention and lifelong learning.[1] This assessment culture plays a critical role in how a new form of assessment is valued and implemented.[1] Students are able to self-assess their knowledge of content areas and reflect on performance in critical areas. Report cards are given to students to help facilitate this process for self-assessment and improvement. When first implementing an examination of this nature, students were fearful or apprehensive of taking the examination. Yet, students began to see the value of the assessment over time. They became more curious about their performance as the examination indicated what they had retained across the curriculum.

In addition to the benefits of a summative examination offered students, the attention that was given to the summative examination process began to change the way faculty members viewed assessment. Faculty members became more aware of assessment procedures, item statistics, and how to write good examination items as a result of "just-in-time education" provided during the item writing process. Course coordinators began to hold item review sessions before examinations to get the input of

colleagues (both content and noncontent experts) on the quality of questions used in interim and final examinations. Use of item statistics after administration of an examination increased as faculty members learned to better interpret these metrics. Some course coordinators created student report cards to provide students with a detailed review of their performance for each course examination using a similar database as was used for the summative examination. Item writing and assessment are skills that must be practiced and the quality of assessment increased during this time. All of these changes in assessment culture persisted or coincided with new initiatives that may be attributable to an advancing culture of assessment or quality improvement.

To create a quality examination, many hours were invested. While the technology used in this process was similar to that used on other examinations at the college, the human resource hours used to create and maintain the summative examination were extensive and included database building for items, examination creation using blueprint and item statistics, development of examination for online administration, student report card creation and distribution, examination result reports to faculty members and administration, and integration of the results into the programmatic assessment algorithm. This process included about 10 faculty members and practitioners who reviewed and discussed each item for approximately 5-10 minutes each. The resources invested into this project were extensive, but we believe the return on investment far outweighed the costs. Students and faculty members benefitted from this process and the change in culture that was created during this process persisted and was of value in other strategic measures at the college.

The examination was designed to measure knowledge and cognition gained during the first two years of the curriculum using multiple-choice items. It was not designed to assess noncognitive domains that may be better measured using OSCEs or cocurricular or experiential activities. Because of the breadth of hours covered in the summative examination, it is a representative sample of the important concepts from the first two years of the curriculum. The use of the necessity score as determined by SMEs helped us narrow down the most useful items for inclusion on the examination. Therefore, we feel that the items encompass concepts truly required of a rising P3 student in our program. Reliability measures (Cronbach alpha or KR-20) are important for summative examinations and should ideally be greater than 0.90. Our reliability measures have yet to meet this 0.90 data point. Our goal is for this measure to increase each year as quality and number of items increases on the examination.

Schools looking to implement a similar examination should anticipate and account for the faculty time required to create and review all items. Inclusion of a faculty member with dedicated time and expertise in the area of assessment and evaluation is critical to successful implementation and ongoing commitment to continuous quality improvement. Additionally, before starting the process, faculty and student buy-in should also be gained on the value of a progression examination to detect nonminimally competent students.

Accrediting agencies mandate proof of student learning and retention of learning. This examination was valid to measure curricular outcomes and used SMEs to assess the necessity of each item for inclusion on the examination. Data were managed at the college level and therefore, could be used in innovative ways for curricular improvement, alignment of curricular content with educational outcomes, and creation of individualized student education plans. The collective effort to create an examination of this caliber also spurred interest among faculty members. Faculty development in item writing and assessment techniques was an unexpected yet positive result of this process. Since faculty members were writing and reviewing items themselves and within small groups, faculty development was occurring in ways that might not happen using a prepackaged examination. Summative assessment of this nature could be used with cohort schools as a benchmarking examination to evaluate student success within different curricula. With the requirement in the ACPE Standards 2016 to use the Pharmacy Curriculum Outcomes Assessment (PCOA) for assessment of foundational knowledge, benchmarking between programs may be accomplished.[3] With strategically placed examinations, programs could compare the performance of their students long before student performance on the NAPLEX is known. For schools with distant campus sites, this type of examination could serve as a comparator between campuses.

We believe there are certain characteristics that make a good summative examination. Items should be of high necessity, discriminate well, and be easy enough to discriminate well at the minimal competence level. For the purpose of this examination, we strive for a necessity rating of at least 2.75, a pbi of at least 0.1, and item difficulty around 0.75. We have found items meeting these criteria test the big picture concepts, which were our intentions when creating this examination. This type of examination is not meant to discriminate between top performers in the class; instead, the examination can identify students not minimally competent and therefore not ready to progress to the next year of the curriculum. As the databank of acceptable examination items increases, we

anticipate increasing the number of scored items on the examination, which will increase the overall reliability of the examination. Increasing the number of items in the databank will also allow for increasing our minimum standards for item quality leading to an even better examination. Using individualized student report cards is another unique offering of this examination and provides students with detailed, longitudinal feedback often not available to them within the curriculum.

CONCLUSION

Summative examinations are valuable forms of assessment. The development of this examination resulted in positive outcomes for faculty members and students with positive changes in the assessment culture that influenced other aspects of the program. The examination provided data on curricular effectiveness while giving student feedback on knowledge, emphasizing the need for lifelong learning. This result outweighs the costs of resources used during its creation. Other schools may want to consider establishment of similar examinations. The use of SMEs to assess the item necessity as a marker of content validity could be used in other testing situations. With the increasing desire to prove student learning outcomes and retention of knowledge, this type of examination was beneficial to the college.

ACKNOWLEDGMENTS

We would like to thank faculty members who contributed questions to the examination and to the members of the UAMS COP Assessment Committee since 2007 who helped develop and advance it.

REFERENCES

1. Anderson HM, Guadalupe A, Bird E, Moore DL. A review of educational assessment. *Am J Pharm Educ*. 2005;69(1):Article 12.
2. Formative Assessment, The Glossary of Education Reform, Great Schools Partnership. http://edglossary.org/formative-assessment/. Accessed October 20, 2014.
3. Alston GL, Love BL. Development of a reliable, valid annual skills mastery assessment examination. *Am J Pharm Educ*. 2010; 74(5):Article 80.
4. Plaza CM. Progress examinations in pharmacy education. *Am J Pharm Educ*. 2007;71(4):Article 66.
5. Szilagyi JE. Curricular progress assessments: the MileMarker. *Am J Pharm Educ*. 2008;72(5):Article 101.
6. Accreditation Council for Pharmacy Education. Accreditation Standards and Key Elements for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree, 2016. https://www.acpe-accredit.org/pdf/Standards2016FINAL.pdf. Accessed March 15, 2015.
7. Medina MS, Plaza CM, Stowe CD, et al. Center for the Advancement of Pharmacy Education 2013 Educational Outcomes. *Am J Pharm Educ*. 2013;77(8):Article 162.
8. Schneider EF, Castleberry AN, Vuk J, Stowe CD. Pharmacy students' ability to think about thinking. *Am J Pharm Educ*. 2014; 78(8):Article 148.
9. Schraw G, Moshman D. Metacognitive theories. *Educ Psychol Rev.* 1995;7(4):351-371.
10. Cizek GJ. Chapter 10: Standard Setting In: Downing SM, Haladyna TM, Eds. *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2006: 225-258.
11. Varma S. Preliminary item statistics using point-biserial correlation and p-values. Morgan Hill, CA: Educational Data Systems, Inc; 2006.
12. Wilson FR, Pan W, Schumsky DA. Recalculation of the critical values for Lawshe's content validity ratio. *Meas Eval Couns Dev.* 2012;45(3):197-210.
13. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ,* 2002;15(3):309-334.