Current Issues in Education
Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Volume 16, Number 2 August 11, 2013 ISSN 1099-839X

Criterion-Referenced Exit Examinations: An Institution's Internal Process for Psychometric Analysis

Cristian Lieneck, Eileen Morrison, and Larry R. Price
Texas State University-San Marcos

The Texas State University-San Marcos undergraduate healthcare administration program requires all bachelors of health administration (BHA) students to pass a comprehensive examination to demonstrate their knowledge of specific core competencies. This also demonstrates completion of their didactic coursework in order to enter a practical internship or residency experience. Since this examination provided important documentation of student learning, the program conducted a detailed psychometric analysis of its three most recent undergraduate comprehensive exit examinations. In order to determine the value of this examination psychometrically, an evaluation of item validity evidence, between-exam reliability, and related assessment of descriptive statistics with regard to overall exam results and individual healthcare administration competency outcomes was necessary. Using Classical Test Theory (CTT) as a methodological framework, the psychometric analysis involved calculating item-level indices that assessed descriptive, validity, difficulty, and discrimination characteristics. This allowed the program's faculty to better interpret student exam outcomes at the overall exam and within-exam competency levels. Additionally, this analysis provided an evaluation of the score reliability of the three alternate exam forms, as well as within-exam healthcare administration competency items, furthering the program's comprehensive exit exam test development process. The outcomes of the analysis included an increased awareness of potential non-equivalent test forms for the total exam and within each exam (competency) level, increased level of interpretation the descriptive results for each exam, and the establishment of a more robust test development process to guide future comprehensive examination efforts.

Keywords: healthcare administration, alternate form reliability, comprehensive exit examination, test psychometric analysis.

The impetus for this article resulted from the collaboration of university faculty interested in conducting a criterion-referenced exit exam for the BHA undergraduate program of study, located in the School of Health Administration at Texas State University. These faculty members initially did not have sufficient experience in conducting the sophisticated psychometric analyses necessary to evaluate exam outcomes. Often, exam outcomes of this type are sent to a third party statistical analysis firm for ensuring that an exam exhibits adequate evidence of score reliability and validity. However, because of increased budget constraints and the requirement for university programs to further understand their exam quality, the School of Health Administration conducted an internal psychometric evaluation of their exam examinations. The findings of this study provided guidance for the School of Health Administration specific to future steps to ensure successful comprehensive exit

examinations in their undergraduate healthcare administration program. The psychometric process was conducted using Classical Test Theory (CTT); a psychometric test theory that can be utilized for any field of study. The study's limitations, while thoroughly addressed in the Results section, primarily highlight the impact of small sample size on the integrity of psychometric analyses. To this end, exam descriptive statistics are analyzed, while further psychometric analyses serve as a demonstration of the process of item level analyses conducted on the program's comprehensive exit exam.

Literature Review

Healthcare Administration Competency

The field of healthcare administration possesses several types of competencies required of healthcare administrators (Calhoun, Davidson, Sinioris, Vincent, & Griffith, 2002; Calhoun et al., 2004; Healthcare Leadership Alliance, 2005; Maurer & Grazier, 2001; National Center for Healthcare Leadership, 2010; Robbins, Bradley, & Spicer, 2001; Stefl, 2008). Additionally, the wide range of careers available for healthcare administrators, as well as extensive range of competency domains in healthcare administration led to numerous certifying bodies and related assessments. These evaluations methods assess the competency of their respective membership and often grant fellowship, or other credentialed status upon completion of a specialized examination or similar assessment instrument (American College of Healthcare Executives, 2011; American College of Medical Practice Executives, 2011; Healthcare Financial Management Association, 2011). These certifying bodies often describe the successful completion of their competency assessment process as "board certification," thus allowing potential healthcare administrators who have demonstrated a specific level of competency the privilege of calling themselves, "board certified" in their respective association's area of specialization.

The History of "Board Certification" in Health Care

The concept of board certification has primarily surrounded the field of medicine and allied health professions. Medical doctors, doctors of osteopathic medicine, specialty physicians, pharmacists, nursing, as well as many other clinical professions offer board certification examinations in order to discriminate among those candidates who have met a specific level of competency related to their field, and those who have not. Furthermore, the board certification examination certifies that an individual has successfully completed a course of study and possesses the required knowledge and skills for that specialty (American Board of Medical Specialties Public Education Program, 2011). The method to establish one's ability to master a specific level of competency is usually applied through the use of a certifying examination, which may include a written and/or practical examination, depending on the course of study. For all board certification assessments, including those previously mentioned for the field of healthcare administration, a similar concept exists: the instruments are utilized to effectively discriminate among applicants who do possess a minimum level of competency in the field, against those who do not.

The Criterion-Referenced Exam and Confirmation of Minimal Competency Levels

Cohen and Swerdlik (1999) and Ebel and Frisbie (1991) define a criterion-referenced examination as one that effectively describes the behavior expected of an individual, or their relationship to a specific subject matter. Additionally, the passing of a criterion-referenced examination is often necessary prior to furthering the student learning experience through a practical learning context (Ebel & Frisbie, 1991). Because of this, the course of study subject matter is used to determine evaluation criteria. In addition, those individuals who become "board certified" must achieve a score that is representative of a particular level of mastery on a criterion-referenced examination (i.e., a "cutoff score") (Ebel & Frisbie, 1991, pp. 37-38), thus demonstrating their knowledge of the specified material.

For healthcare administration education, undergraduate programs may require a comprehensive exit examination at the end of students' didactic study to objectively confirm students' retention of a minimal level of knowledge prior to entry into their internship/residency fieldwork experience, or prior to graduation. As a result, a key objective for this examination is to evaluate students' overall competency in the field of healthcare administration. In addition, criterion-referenced examinations also evaluate specific individual healthcare administration competencies. To this end, the examinations are also expected to reflect evidence of content, construct, and rational validity (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999).

In the context of the program evaluation presented here, exam items were developed in consideration of establishing evidence of intrinsic rational validity (Ebel & Frisbie, 1991) based on healthcare content domains identified by faculty and course objectives designed to measure the content for each course. The underlying concept of the exam was to affirm that students retained the basic concepts of each course prior to the beginning of their external residency experience. Therefore, this program's exit exam closely resembles the characteristics of the previously described board exam.

Interpretation of the results for a criterion-referenced examination differs from the routine, in-class examination. For a criterion-referenced comprehensive exam in healthcare administration, it is expected that the

students have already mastered the content on the exam if they have successfully passed these required courses leading up to the exam itself. This posits the assumption that the students are able to successfully demonstrate their knowledge of each course (or overall competency) by performing well on all sections on the exam that correspond with their previous classes. The verification of overall knowledge in each criterion-referenced section of the exam is necessary to ensure students are prepared to exit the classroom setting and enter their practical fieldwork experiences with adequate competency in each domain of healthcare administration.

Since all exam items should link to the subject matter and specific course objectives, scores on the instrument should technically produce low levels of discrimination among higher and lower scoring students, as they are expected to have retained a minimum level of knowledge for each competency present on the exam. In other words, students are expected to do well on the comprehensive exit examination if they retained a minimum level of knowledge from their course of study, yet still have the potential to score poorly on the exam if they did not retain a minimal level of knowledge from their undergraduate experience. This objective measure ensures that students are not entering their practical fieldwork experiences without first demonstrating a minimal knowledge level of healthcare administration competencies set forth by the School of Health Administration.

Other Reasons for an Undergraduate Comprehensive Exit Exam

While there is a strong requirement for healthcare administration undergraduate students to be highly knowledgeable in the prescribed healthcare administration core competencies upon completion of their studies, previous research suggests that a minimum level of competency is not completely established prior to fieldwork and/or practical experience (Hartman & Crow, 2002; Helfand, Cherlin, & Bradley, 2005; Hudak, Brooke, & Finstuen, 2000; Mecklenburg, 2001; Pointer, Luke, & Brown, 1986; Schneller, 1997; White & Begun, 2006). Additionally, educational accreditation bodies require empirical evidence demonstrating not only that the undergraduate curriculum encompass specific healthcare administration core competencies, but also that it demonstrates the mastery of competencies at the individual student level (Association for University Programs in Health Administration, 2011). While the course of study in a healthcare administration program may provide learning in the necessary core competencies, the use of a comprehensive exit exam empirically demonstrates overall competency in healthcare administration by evaluating each individual's competency in core areas, well as his or her overall competency in healthcare administration prior to program completion.

Assessment of an Undergraduate Comprehensive Exit Exam at Texas State

History and Development of Alternate Exam Forms

The use of a comprehensive exit examination at the Texas State School of Health Administration's undergraduate program has been a form of competency assessment for many years. Successful completion of this examination is required before the students begin their residency program. The intent of this requirement relates to quality assurance. The School wants to assure preceptors and future employers that the students who graduate from the program exhibit evidence of cognitive or knowledge-based competencies related to their degree. In addition, faculty members can use the student results on their section of the examination to affirm that their course content is correctly specified according to course objectives and test specifications.

During a recent review of the exam process, a decision was made to increase faculty involvement in question development and to link the examination closer to terminal objectives for each course (course objectives). A process was developed to facilitate this decision. The first part of this process was to establish a cutoff score that represented a pass or fail status for the exam. The Angoff Method, which is often utilized for certification and licensing exams, was implemented by subject matter experts (program faculty) conducting a review of each test item and evaluated the likelihood that a minimally competent student would respond to each item correctly (Zieky, 2008). Implementing this method, as well as the standard criteria used to delineate passing versus failing courses in the healthcare administration major, resulted with an acceptable cutoff score to be set at 70%. Therefore, students who met this overall exam result requirement successfully passed the examination. Additionally, there was no ranking of students by exam score for this examination because its goal is not to discriminate between the highest performing students and the lowest performing students. Similar to professional board exams, this examination's purpose is to measure or verify basic competency in the associated field of study, therefore allowing for mutually exclusive, dichotomous groups to be established upon completion of the exam: pass and fail.

Next, the length and scope of the exam was assessed. After a careful review of the time allotted to take the exam, as well as student performance variables and logistics, the length of the exam was established at 200 questions (items). The scope of the exam involved (a) consideration of the 16 courses included in the core curriculum of the undergraduate program and (b) the need for all courses to be equally represented. Therefore, each course contributed 12 questions to the overall examination, (16 healthcare administration courses X 12 questions per course = 192 questions) with the remaining eight questions covering areas of common knowledge and

professionalism. These final eight questions also serve as potential test (experimental) items for future exams, as well as questions directly related to previous semester events applicable to all students such as guest lectures, student healthcare administration organizational events, and/or other important current events.

The test development coordinator (internal to the department) requested multiple-choice format examination questions from each primary faculty member for all of the 16 courses in the BHA curriculum. The faculty member designed his or her questions to measure content specified on the course syllabi through specified terminal course objectives. Multiple-choice questions were the only item format to be constructed by the individual faculty members; Haladyna (2004) item creation and form guidelines served as a standard reference for all faculty. Faculty submitted their questions to the test development coordinator who assembled the questions into a test format, which included attention to stem and foil integrity, specific wording, answer sequence, and duplication (Haladyna, 2004). After specific item-level adjustments to address overall exam flow and formatting, the test coordinator asked each faculty member to review his or her questions for both accuracy and design. The test coordinator used this process to (a) link the content of the exam with the concepts and details provided in the curriculum and (b) to provide the opportunity for establishing content validity evidence by expert review (AERA/APA/NCMA, 1999).

The School of Health Administration conducted its undergraduate exit exams at the end of the fall and spring semesters. Due to the frequency of these examinations as well as test security purposes, there was a need to develop a minimum of three exit examinations. The process for developing these examinations was the same for each examination. This process provided a stable methodology for test construction since each test involved the same faculty members, course objectives, and review process.

Methodology

The Test Development Process and Psychometric Analysis

As previously noted by Calhoun et al. (2002) solid preparation and establishment of psychometric protocols in art of measurement and the science of testing is important to quality test development. Initially, the School of Health Administration lacked these sophisticated psychometric protocols. Therefore, program faculty members decided to conduct an internal review of the psychometric properties of the tests after administration of the three comprehensive exit examinations. The rational for waiting until there were three tests was to verify that the test development process yielded multiple forms of a test that produced scores that meet the requirements of equivalence.

This was a complex task. Specifically, the process involved ensuring that tests were targeted to produce equivalent scores, were constructed to the same explicit content and statistical specifications, and were administered under identical conditions (AERA/APA/NCME, 1999). For example, content experts must (a) work diligently to ensure that test items reflect the overall universe or domain of all possible items that a test could be composed of; and, (b) that items are constructed in light of the goals and objectives for the courses comprising the curriculum. Once the issues of content and statistical specification are delineated, scores on parallel forms of a test can be psychometrically evaluated for their equivalence in terms of (a) item-level and total score descriptive statistics, (b) score reliability for the total test, and (c) score validity at the level of the item and the total test. The aforementioned steps are conducted with regard to individual student healthcare administration competency levels. Since the School of Health Administration recently established three new alternate forms of the exit examination, a test development process similar to the guidelines offered above was conducted. Once test scores were obtained from the pilot sample of examinees, psychometric analyses were performed on each comprehensive exit exam to assess score equivalence for the three forms of the test.

The Test Development Process

Item-level analyses are essential to understanding how scores obtained from examinees on test items are functioning in relation to the goals of the test. As previously mentioned, the case study presented herein is based on classical test theory (CTT) (Crocker & Algina, 1986). A shortcoming of classical test theory is that test and item-level statistics are sample dependent. Because CTT-based statistics are sample dependent, aligning scores on different test forms is difficult, if not impossible. Thus, the sample-dependency issue was identified as a major problem for the program. Cohen and Swerdlik (1999) and Schmeiser and Welch (2006) suggest the following steps in the test development process:

- Test conceptualization;
- Test construction;
- Test tryout;
- Test analysis;
- Test revision.

Focusing on principles of quality assurance and improvement, this process contains a feedback loop which allows for test revision after analysis of data, including descriptive and psychometric exam statistics, followed by subsequent test tryouts over future testing periods. After development of three semesters of exit examinations (Fall 2009, Spring 2010, and Fall 2010 semesters), and considering the process used to develop alternate forms of the examination, it can reasonably be concluded that the

department has successfully engaged in the first three steps of the test development process. In order to understand each exam's characteristics and level of competency measurement in healthcare administration, a more complete review of the examination, including descriptive statistics and specific psychometric analyses was necessary. This analysis can direct the future test revision processes within the department.

Psychometric Analysis

The data collection and/or tryout sample acquisition process involved gaining access to each semester's exit exam raw data. The test development coordinator provided the University's testing center with the key, student examinations, and any other required information. All three comprehensive exit exams items consisted of a multiple choice format and were administered using Scantron forms. All of the raw data was made available in electronic format (MS Excel) by the University's testing center. The test development coordinator used this information to determine which students passed the exit examination and informed students individually of their results. Confidentiality was maintained throughout the score reporting process.

The testing center provided a number of assessments for each examination, including some descriptive, item, and test-level psychometric statistics, as well as returning each exam's raw data file to the test coordinator. It was from these initial raw data files that additional item-level psychometric indices were calculated using MS Excel to establish item-level difficulty and discrimination characteristics. A brief summary of each psychometric statistic calculated for each exam form is listed below.

1. Kuder-Richardson Formula 20 (KR-20 statistic) – an evaluation of internal reliability for measurements with dichotomous results (correct answer, incorrect answer on item-level responses). A KR-20 coefficient > 0.90 indicates an internally consistent test structure (Kuder & Richardson, 1937; Crocker & Algina, 1986). Statistical Package for the Social Sciences (SPSS) was used to calculate the statistic.

2. Item-difficulty index (p) – an item-level analysis that describes the proportion of test takers that scored an individual item correctly (Ebel & Frisbie, 1991; Nunnally & Bernstein, 1994).

Directionality assessment: High values (0.5 to 1.0) indicate an item that most of the students scored correctly. Low values (0.0 to 0.49) indicate an item that most of the students scored incorrectly.

Formula:

p = (# of subjects scoring the item correctly) / (# of subjects taking the exam)

3. Item-discrimination index (d) – an item-level analysis that compares the performance on an item with the upper and lower regions of the continuous overall exam scores (Ebel & Frisbie, 1991; Cohen & Swerdlik, 1999).

Directionality assessment: Negative discrimination indices indicate an item with a stem or foil problem and require immediate revision or omission altogether.

Formula:

$$d = p_u - p_l,$$

- where p_u = proportion of the upper 25% who answered the item correctly;
- p_l = proportion of the lower 25% who answered the item incorrectly.

4. Point-biserial correlation coefficient (r_{pbi}) – an additional method for evaluating discrimination among items and how they separate better performing students on each section of the exam with the lower performing students in that same section of the exam (Crocker & Algina, 1986; Ebel & Frisbie, 1991).

Directionality assessment: The higher the r_{pbi} correlation, the better the item is at discriminating among examinees.

Formula:

$$r_{pbi} = M_p - M_q / S_t (\sqrt{pq})$$

- Where M_p = the whole-section mean (12 questions) for students answering the item correctly, M_q = the whole-section mean (same 12 questions) for students answering the item incorrectly;
- S_t = standard deviation for the same 12 question section;
- p = proportion of students answering correctly;
- q = proportion of students answering incorrectly.

Results

Limitations

There are several limitations relevant to current study. These limitations are highlighted to provide guidance for future work in the area of examination development and validation for health care administration programs. The first limitation is the composition and size of the sample used for tryout analyses. As with any quantitative study, sample size is critical to ensure accurate estimation of score reliability of the instrument. The sample used herein served as a working example in order to establish future department-level psychometric protocols. To this end, improving the psychometric processes employed in the School of Health Administration addresses the important goal of furthering the program's quality assurance initiative and evaluation

of student healthcare administration competencies before the fieldwork experience (internship/residency in healthcare administration). With the progression of each future spring and fall semester, additional subjects and their related scores will be continuously evaluated and thus increase overall sample size for the established process. For reasons previously mentioned, an in-house process to evaluate the exam was necessary and the faculty employed principles based on Classical Test Theory (CTT), understanding that sample size will grow with future student matriculation, yet also utilizing caution when interpreting and inferring item-level psychometric indices upon such an initial, small sample.

The second issue concerns the suggestion or recommendation for using items interchangeably across test forms when the item statistics are clearly sample dependent. In this study each sample item-response pool (semester) provided less than 30 subjects; a number representing the number of alternate forms evaluated between semesters. Therefore any psychometric statistics were calculated and based (interpreted) on the sample size from which they originated. With limited and varying sample pools between semesters, it is important to note the limitation and interpret these indices with caution also, until future semesters progress through the exam process and further increase overall sample size. Once an adequate sample size is available, the program should employ item response theory (IRT) (DeAyala, 2009) to calibrate all items in the pool onto the same metric. In this way, parallel forms of the test will be easily assembled for future use.

A third concern relates to the validity (Kane, 2006) of the examinations used by the School of Health Administration. Due to sample size constraints, evaluation of construct validity using confirmatory factor analysis (CFA) (Brown, 2006) was not possible. Programs should plan on conducting CFA to provide evidence of construct validity of their exams once an appropriate sample size is achieved.

Descriptive Analysis

Initial analysis of each comprehensive exit exam began with the descriptive results. These values, calculated using the raw data from each exam, provide an initial, broad review of the exam results. The descriptive statistics shown in Table 1 were used for initial interpretation of exam results.

Table 1
Healthcare Administration Exit Exam Descriptive Statistics by Semester

Semester	Mean	Standard Deviation	Min	Max	Median
Fall 2009 n=22	75.3	6.51	64.5	91	74.3
Spring 2010 n=14	69.1	7.30	58.0	81	69.3
Fall 2010 n=19	76.2	4.24	68.5	83.5	76

Notes: total n=55.

Table 2
Item-Level Difficulty and Discrimination Indices for Each Healthcare Administration Exit Exam (Single Course)

Course Name	Item Number	Difficulty Index (*p*)			Discrimination Index (*d*)		
		Fall 2009 n=22	Spring 2010 n=14	Fall 2010 n=19	Fall 2009 n=22	Spring 2010 n=14	Fall 2010 n=19
Healthcare Organization and Delivery							
	1	0.09	1.00	0.89	0.16	0.00	0.40
	2	0.95	0.92	0.94	0.00	-0.25	0.00
	3	0.09	0.50	0.89	0.16	0.25	0.20
	4	0.50	0.92	0.05	0.50	-0.25	0.00
	5	0.72	0.42	1.00	0.83	0.75	0.00
	6	0.86	0.28	1.00	0.16	0.50	0.00
	7	0.54	0.35	0.52	0.66	-0.75	0.80
	8	0.95	1.00	0.73	0.00	0.00	0.40
	9	0.77	0.78	0.89	0.33	0.50	0.20
	10	0.50	0.28	0.94	0.50	1.00	0.20
	11	0.41	0.78	0.36	0.33	0.25	0.80
	12	0.86	0.92	0.94	0.00	0.25	0.00

Demonstration of a single course's difficulty and discrimination indices at the item-level to allow for both horizontal and vertical analyses.

Central tendency was evaluated by comparing the total mean score for each semester. Two semesters' total exam means were very close in average score (the Fall 2009 and Fall 2010 semesters) while the Spring 2010 semester's total mean fell below the "C" letter grade level (i.e., a score of 70). While variability was quite similar among exams, with the Fall 2010 semester exam possessing the lowest variance (4.24) and the Spring 2010 semester exam having the largest variance (7.30), further descriptive analysis demonstrates overall variability among exam forms. For example, while the range of total scores within each semester showed low variance, the Spring 2010 semester's minimum score fell within the "F" letter grade range and the other two semesters' minimum scores fell within the "D" letter grade range. Similar variability exists with regard to maximum and median scores for each semester.

Initial observation of the overall results by semester may suggest that either the Spring 2010 students were not as competent in the healthcare administration criterion-referenced material, or the exam that semester was more difficult. But, this conclusion cannot immediately be made based upon the overall descriptive statistics for each semester in Table 1, especially because each semester's comprehensive exit exam consisted of different test items and therefore were mutually exclusive. Further item-level analysis was required to thoroughly investigate potential differences in student overall score and individual competency levels on each of the three exit exams.

Psychometric Results

The Kuder-Richardson formula for internal consistency was computed for each 200-item exam using SPSS, as it is analogous to Cronbach's alpha in evaluating internal consistency, except that Cronbach's alpha is not sensitive to continuous variables (Cortina, 1993). The dichotomous data (correct versus incorrect student responses for each item) for each exam resulted in KR-20 reliability statistics of 0.83 (Fall 2009), 0.84 (Spring 2010), and 0.65 (Fall 2010). While it is expected for students with a high level of knowledge in the healthcare administration content domain to answer a high number of items correctly on the exam, and students with a low level of knowledge in the content domain to answer a low number of items incorrectly, the statistic did not reach the preferred 0.90 coefficient to demonstrate homogeneous exams (Crocker & Algina, 1986). However, further evaluation of the KR-20 reliability statistic is recommended for future study. Sensitive to item difficulty, range of exam scores, and length of the exam,

Table 3
Point-Biserial Correlation Coefficients by Semester (Single Course)

Course Item Number	Point-Biserial Coefficient (r_{pbi})		
	Fall 2009 n=22	**Spring 2010 n=14**	**Fall 2010 n=19**
Healthcare Organization and Delivery			
1	0.273	a	**0.491***
2	0.042	-0.326	0.042
3	0.273	0.141	**0.347**
4	**0.384**	-0.143	-0.042
5	0.693**	**0.543***	a
6	0.170	**0.535***	a
7	**0.432***	0.679**	0.608**
8	0.042	a	**0.305**
9	**0.410**	**0.532**	0.204
10	**0.320**	0.743**	**0.436**
11	0.296	0.188	0.687**
12	-0.017	**0.404**	0.042

Demonstration of a single course's point-biserial coefficient index results at the item-level to allow for both horizontal and vertical analyses.
a = Correlation cannot be computed because one of the variables is a constant.
**. Correlation is significant at the 0.01 level.
*. Correlation is significant at the 0.05 level.
Bold items represent an effect size range from 0.30 to 0.60.

increased sample size may posit more interpretable KR-20 coefficient results (Horst, 1953).

Additional comprehensive results for the study included both descriptive (Table 1) and psychometric analyses at both the semester and individual student level (Lieneck, 2011). To facilitate interpretation of the item-level analyses, comparative data tables were created in order to allow both vertical analyses of each exam's psychometric statistics, as well as horizontal analysis across semesters. This allowed for evaluation of individual competencies within exams, as well as comparison of competency analysis results across the three alternate forms (semesters). Table 2 provides an example of an abbreviated results table. It demonstrates how to format the results of both the difficulty and discrimination indices for a single course (competency) in the healthcare administration program (Lieneck, 2011).

Table 3 demonstrates the format utilized to allow for within exam, as well as across exam investigation into the point-biserial correlation coefficient results (Lieneck, 2011).

Analysis of Psychometric Results

Calculation of psychometric statistics at the item-level creates an abundance of data relevant to effective interpretation of results. Therefore, data summary tables may be created to easily interpret results at a comparative level. One approach to creating a smaller, summary table of psychometric results is to display frequencies of items not meeting specific criteria by healthcare administration competency. This step allows for immediate identification of exam items that require additional review. Several psychometric index characteristics (provided in the next section) provide metrics for further interpretation of exam items, student scores specific to each item, thereby informing the test development process regarding potential item revisions and/or omissions.

Negative Discrimination Indices

Schmeiser and Welch (2006) and Cohen and Swerdlik (1999) define an effectively discriminating item as one that most of the high scorers answer correctly and the low scorers answer incorrectly. In theory, this allows

for effective discrimination at the item level between more competent students and less competent students. Therefore, a negative d index value identifies an extremely poor exam item, resulting in the lower scoring group answering the item correctly and the better performing students answering the item incorrectly. This instance, termed a "nightmare" by Cohen and Swerdlik (1999, p. 207) could be a result of several issues, including the possibility of a confusing or poorly worded item stem and/or item distracters (also known as foils). Immediate identification, revision, and/or item omission is required. Table 2 demonstrates three negative discrimination indices, all falling within the Spring 2010 exam. Individual, item-level analysis of these three items is required by the test coordinator. Furthermore, as sample size continues to increase, a more detailed examination of student responses for items with negative discrimination indices may provide further insight into their occurrence.

As a criterion-referenced comprehensive mastery exam in healthcare administration, it is expected that the students have already mastered the content on the exam and are therefore able to successfully demonstrate their knowledge of each course (overall competency) by performing well on all sections on the exam, for both upper and lower performing groups. This verification of overall knowledge in each criterion-referenced section of the exam (course level) is necessary to ensure the students are prepared to exit the academia setting and enter their practical fieldwork experiences with a strong level of competency in healthcare administration. Therefore, low discrimination indices are expected to occur on this type of criterion-referenced examination, with the lower scoring group answering several items correctly, as did the upper scoring group.

Low Difficulty Indices

As a criterion-referenced comprehensive exit exam, the overall goal is not to perfectly discriminate the student population, or otherwise maximize the variance of scores (Crocker & Algina, 1986). If this were the case, then the exam would successfully result in an estimated number of students failing the exam on a regular basis, based on the prescribed level of overall item difficulty for each exam. Instead, this program's comprehensive exit exam was established to verify the retention of healthcare administration competency knowledge prior to the field experience. Therefore, it is an assumption that if each student has established a specific level of competency in each competency assessed, low difficulty indices should not occur. If low difficulty indices result, several variables exist that can be evaluated to address the items with poor difficulty indices (Crocker & Algina, 1986):

- Potential instructional effectiveness on this healthcare administration competency item may be inadequate at the instructor and/or program level.
- Item specification may be inadequate at either the stem and/or foil level.
- Other potential confounders may exist that have yet to be determined, based upon specific circumstances and individual analysis of each item with a low difficulty index (other unknown reasons not mentioned above).

At the School of Health Administration, difficulty index results were also provided from the Texas State University testing center and form the basis for a test review and question revision. There was a review of questions missed by 50% or more of the students each semester (d values < 0.50) and these questions were changed or replaced by the faculty members who initially wrote them. Table 2 demonstrates nine difficulty indices falling below the 0.50 difficulty cutoff score (Fall 2009 exam=3; Spring 2010 exam=4; Fall 2010 exam=2). As each exam was completed, there was a repetition of this process. The intent of this repetition was to improve reliability and validity of the examinations. Continued matriculation of students will further enable the study to approach an adequate sample size, allowing for further investigation into suspect items with difficulty indices < 0.50 by analyzing the percentage of students who chose each response option.

The test development coordinator also gave faculty members feedback on the results of their individual section on the examinations. This provided them with information to review their course objectives, teaching styles, classroom testing, and other areas that affect retention of information. Using this technique, an understanding of the results of individual course sections on the exit exam can improve teaching and content retention in the curriculum.

Lessons Learned

Implications with Non-Parallel, Alternate Forms and Sample Characteristics

When non-parallel alternate forms are found to exist, it is inappropriate to assume equivalency of descriptive statistics (example: means, medians, and standard deviations) across exam versions. As a result, an individual's overall score of 70% on the program's comprehensive examination cannot be accurately compared to another individual's overall score of 70%, when both subjects took different versions of the exam with varying psychometric analysis results. Additionally, sample size and homogeneity were limiting factors specific to the psychometric quality of the scores acquired on the test forms. Both item difficulty and item discrimination values must be taken into consideration before concluding that one student's score (or a group of students' mean score) describes a specific level of healthcare administrative competency when compared to another student's (or group of students') overall score when alternate forms exist. Based on these evaluations, methods to verify score equivalence using the psychometric results of the scores acquired from examinees taking the exams be used by following

AERA/APA/NCME (1999) and Cohen and Swerdlick (1999) test development process.

Therefore, by offering a comprehensive exit examination to assess competency in healthcare administration prior to students entering their fieldwork experience, the undergraduate program possesses a duty to understand the descriptive and psychometric statistical properties of each form and assure equivalence.

Development of an Item Pool for Constructing Parallel Test Forms

The final step of Cohen's test development process is to revise suspect items on the exam after psychometric analysis and enter a continuous feedback loop to reconstruct, tryout, and then analyze results after administration of each exam. Items may also be interchanged by using an item pool that contains questions with similar psychometric characteristics (Schmeiser & Welch, 2006). Therefore, item specification characteristics are calculated (as described by the psychometric indices) and each exam should possess items that can be used interchangeably, based upon their specific psychometric characteristics. Ultimately, this process will continue to improve and refine the creation of parallel (i.e. alternate) exam forms, specifically with regard to item difficulty and discrimination (Crocker & Algina, 1986).

Using the test development process outlined by Cohen and Swerdlik (1999) and Crocker and Algina (1986), as well as quantitative item analysis tools to investigate equivalency of alternate forms as discussed in Anastasi and Urbina (1997), the program's exam forms may be revised or substituted (at the item level) to possess similar difficulty and discrimination results, therefore establishing optimal reliability among forms. Since most students from the healthcare administration program for the current study usually only take the exam once, the single test administration method of assessing reliability among several versions of exams is necessary, with an overall goal of establishing internal consistency across each alternate form (Crocker & Algina, 1986). This is done by calculating difficulty index composite scores for each section (or competency tested) within each exam, therefore assessing the average level of difficulty for each section (competency) across multiple forms (Cohen & Swerdlik, 1999). Similar reliability assessments and composite scores may be completed for discrimination indices (Cohen & Swerdlik, 1999).

Based on the psychometric analysis results for each exam, those items not meeting specific criteria can then be either revised or omitted from the exam altogether (Cohen & Swerdlik, 1999). Revision may entail rewording a single distracter of a specific item, to rewording or replacement of the entire item stem (Cohen & Swerdlik, 1999). Item-level revision allows for those items not meeting specific psychometric properties to be addressed. Once completed, the newly revised forms can then be entered back into the test development process for new students taking the comprehensive exit exam. Furthermore, this continuous process (test tryout, analysis, revision, back to test tryout, etc.) will eventually allow for cross-validation to occur, establishing reliability from a new sample of test subjects with which new psychometric findings for each sample can be compared (Cohen & Swerdlik, 1999). When alternate forms of the healthcare administration comprehensive exit exam exist with equivalent reliability, only then can descriptive statistics be examined across multiple semesters. Based on the results of this study, the School of Health Administration has already begun the continuous process of test tryout, analysis, revision, and test tryout to improve reliability.

Exam Construct Validity for Continued Item and Competency Refinement

In addition to furthering the premise of establishing equivalent forms, construct validation is then required in order to assess each exam form and the extent to which it measures specific theoretical constructs, or competencies, in healthcare administration. Cohen and Swerdlik (1999, p. 197) further defined a construct as, "…unobservable, presupposed traits that a test developer may invoke to describe test behavior or criterion performance." In other words, each form should be evaluated to determine if it is measuring what it is intended to measure (Cohen & Swerdlik, 1999), in this case specific healthcare administration competencies.

If the test serves as a valid method of evaluating the constructs (competencies) related to healthcare administration, those individuals who pass the exam will have the knowledge level as predicted. Therefore, one can assume that they possess the specific level of knowledge in healthcare administration necessary for them to matriculate from the classroom learning environment to the practical fieldwork experience. The intended goal for the School of Health Administration was to establish multiple equivalent forms that assess the same competencies in healthcare administration, as well as similar levels of assessment for each individual competency evaluated. Due to the current study's sample size limitation, the validity estimates are not possible because the low statistical power disallows inferential analyses (Crocker & Algina, 1986). In the meantime, the program will continue to establish validity of the examinations by ensuring that all items are directly related to each course's learning objectives, as well as expert review by faculty members.

In order to further the rigor of psychometric analysis and obtain both convergent and discriminant evidence, an identification of constructs present within each comprehensive exit examination is required. Once sufficient data is collected for each alternate form, factor analysis may be utilized to further assess validity. Factor analysis is an appropriate method of identifying these

latent constructs, therefore evaluating the interrelationships of behavioral data (Anastasi & Urbina, 1997). The factors, or common traits identified within each exam can then be assessed among exams to establish content validity and increased equivalency (Anastasi & Urbina, 1997). Each factor is identified and the weight or loading of each factor assesses the contribution of that underlying concept towards the overall test. Cohen and Swerdlik (1999) describe the factor analysis procedure as highly complex and recommend a computer program to assist with construct identification. As an ongoing task, the assessment and refinement of each comprehensive exit exam will improve upon criterion-related validity as sample size increases. To this end, it will allow a healthcare administration program to effectively judge the utility of the exam regarding the assessment of each student's ability on the criterion-referenced measures (Cohen & Swerdlik, 1999), in this case competency in healthcare administration.

Conclusion

The requirement for undergraduate healthcare administration students to successfully complete a criterion-referenced exit examination is one part of a process that does not end after creation of the initial exam. The feedback loop of Cohen and Swerdlik's (1999) test development process illustrates the critical steps necessary to create alternate exam forms that exhibit adequate validity and reliability evidence for each examination so descriptive exam results may be interpreted more accurately. The healthcare administration department offering such a comprehensive, detailed criterion-referenced exam possesses a responsibility and duty to both its students and accreditation bodies to ensure the method of measurement and interpretation of examination results remains sound and fair for all stakeholders involved.

References

American Board of Medical Specialties. (2011). *What board certification means*. Retrieved from the American Board of Medical Specialties website: http://www.abms.org/about_board_certification_means.aspx

American College of Healthcare Executives. (2011). *Board certification in healthcare management.* Retrieved from the American College of Healthcare Executives website: http://www.ache.org/mbership/credentialing/credentialing.cfm

American College of Medical Practice Executives. (2011). *Board certification in medical group practice management.* Retrieved from the Medical Group Management Association/American College of Medical Practice Executives website: http://www.mgma.com/acmpe/

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Upper Saddle River, New Jersey: Prentice Hall.

Association for University Programs in Health Administration. (2011). *Criteria for undergraduate program certification.* Retrieved from the Association for University Programs in Health Administration website: http://www.aupha.org/i4a/pages/index.cfm?pageid=3519

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Calhoun, J. G., Davidson, P. L., Sinioris, M. E., Vincent, E. T., & Griffith, J. R. (2002). Toward an understanding of competency identification and assessment in health care management. *Quality Management in Health Care, 11*(1), 14-38.

Calhoun, J. G., Vincent, E. T., Baker, G. R., Sinioris, M. E., & Chen, S. L. (2004). Competency identification and modeling in healthcare leadership. *Journal of Health Administration Education, 21*(4), 419-440.

Cohen, R. J., & Swerdlik, M. E. (1999). *Psychological testing and assessment: An introduction to tests and measurement*. New York, N.Y.: McGraw Hill.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York, N.Y.: Harcourt Brace Jovanovich.

De Ayala, R. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement, 5th ed*. Englewood Cliffs, NJ: Prentice-Hall.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, N.J.: Lawrence Erlbaum Associate Publishers.

Hartman, S. J., & Crow, S. M. (2002). Executive development in healthcare during times of turbulence: Top management perceptions and recommendations. *Journal of Management in Medicine, 16*(5), 359-370.

Healthcare Financial Management Association. (2011). Certified Healthcare Financial Professional (CHFP) and Fellow of the Healthcare Financial Management Association (HFMA). Retrieved

from the Healthcare Financial Management Association website: http://www.hfma.org/certification/

Healthcare Leadership Alliance. (2005). *HLA competency directory: User's guide*. Retrieved from: http://healthcareleadershipalliance.org/HLA_Competency_Directory_Guide.pdf

Helfand, B., Cherlin, E., & Bradley, E. (2005). Next generation leadership: A profile of self-rated competencies among administrative residents and fellows. *Journal of Health Administration Education, 22*(1), 85-105.

Horst, P. (1953). Correcting the Kuder-Richardson reliability for dispersion of item difficulties. *The Psychological Bulletin, 50*(5), 371-374.

Hudak, R., Brooke, P. P., & Finstuen, K. (2000). Identifying management competencies for health care executives: Review of a series of Delphi studies. *Journal of Health Administration Education, 18*(2), 213-243.

Kane, M. T. (2006). Validation. In R. L. Linn (Ed.). *Educational Measurement* (4th ed.). Washington, DC: American Council on Education and Macmillan.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

Lieneck, C. (2011). *An examination of individual competencies among students matriculating through an undergraduate healthcare administration program*. Retrieved from Digital Collections@TxState. (txst.10877.3030)

Maurer, R. T., & Grazier, K. (2001). Development of core competencies in health care finance. *Journal of Health Administration Education* Supplement (Fall 2001), 130-145.

Mecklenburg, G. (2001). Career performance: How are we doing? *Journal of Healthcare Management, 46*(1), 8-13.

National Center for Healthcare Leadership. (2010). *National Center for Health Leadership competency model summary.* Retrieved from the National Center for Health Leadership website: http://nchl.org/Documents/NavLink/Competency_Model-summary_uid31020101024281.pdf

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory,4th ed*. New York, NY: McGraw-Hill.

Pointer, L., Luke, R. D., & Brown, G. D. (1986). Health administration education at a turning point: Revolution, alignment issues. *Journal of Health Administration, 4*(3), 423-436.

Robbins, C. J., Bradley, E. H., & Spicer, M. (2001). Developing leadership in healthcare administration: A competency assessment tool. *Journal of Healthcare Management, 46*(3), 188-202.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement, 4th ed*. (pp. 307-353).

Stefl, M. (2008). Common competencies for all healthcare managers: The healthcare leadership alliance model. *Journal of Healthcare Management, 53*(6), 360-374.

White, K. R., & Begun, J. W. (2006). Preceptor and employer evaluation of health administration student competencies. *Journal of Health Administration Education, 23*(1), 53-68.

Zieky, M. J., Perie, M., & Livingstone, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

.

Article Citation
Lieneck, C., Morrison, E., & Price, L.R. (2013). Criterion-referenced exit examinations: An institution's internal process for psychometric analysis. *Current Issues in Education*, *16*(2). Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1052

Author Notes

Cristian Lieneck, PhD, FACMPE, FACHE, FAHM
Texas State University-San Marcos
601 University Drive, Room 250A, Office #272
San Marcos, Texas 78666
clieneck@txstate.edu

Dr. Lieneck has over 10 years of experience in the field of healthcare administration with a focus in medical group practice management. Initially entering the field as a Medical Service Corps officer in the United States Army, Dr. Lieneck led a medical clinic and evacuation platoon in an armored cavalry unit. As a Captain, he served as an Executive Officer and Company Commander, managing four dental practices and one hospital-based oral surgery clinic. As a civilian, Dr. Lieneck held positions of leadership in organizations such as the Texas Medicaid-Waiver Program (Texas DADS), Austin Radiological Association (ARA), as well as serving as the medical group practice administrator of an Austin-based physiatry and pain medicine group practice, consisting of multiple locations and an outpatient procedure suite in central Texas. Dr. Lieneck is a fellow and board certified in medical group practice management by the American College of Medical Practice Executives (ACMPE), the credentialing body of the Medical Group Management Association (MGMA). He is also a fellow and board certified in healthcare management by the American College of Healthcare Executives (ACHE), and a fellow in the Academy of Healthcare Management (AHM).

Eileen Morrison, EdD, MPH, CHES, LPC
Texas State University-San Marcos
601 University Drive, Room 250A, Office #268
San Marcos, Texas 78666
emorrison@txstate.edu

Dr. Morrison has her EdD in Counseling & Administration from Vanderbilt University and her MPH from the University of Tennessee. Prior to joining the Texas State University faculty 2004, Dr. Morrison had extensive experience in curriculum design, course development, and teaching in graduate programs in health administration. She also holds several licenses that include dental hygiene, Licensed Professional Counselor, and Certified Health Education Specialist. Her health administration experience involves positions in corporate and public health. Dr Morrison has served as a consultant to the U.S. Department of Justice, the United States Public Health Service, Head Start, and the United States Army. She also writes in the field of health administration and education and including two textbooks on ethics published by Jones and Bartlett. A new book on health care ethics issues is under contract.

Larry R. Price, PhD
Texas State University-San Marcos
601 University Drive, ASBS #325
San Marcos, Texas 78666
LP11@txstate.edu

Dr. Larry Price is Professor of Psychometrics & Statistics and Director of the Interdisciplinary Initiative for Research Design and Analysis at Texas State University. Prior coming to Texas State University, he served as a Senior Psychometrician and Statistician for the Emory University School Medicine, Departments of Psychiatry & Behavioral Sciences. Dr. Price was employed at The Psychological Corporation in San Antonio as a Senior Psychometrician during 1999-2002. He is a Fellow of the American Psychological Association, Division 5 – Evaluation, Measurement & Statistics, and a member of the American Statistical Association, the Psychometric Society, the American Educational Research Association and the American Mathematical Society. He has published extensively with over 90 peer-reviewed articles, presentations, books and book chapters in journals and books such as: Structural Equation Modeling, Journal of Educational & Behavioral Statistics, Psychological Assessment, Psychological Methods, Elementary School Journal, Journal of Experimental Education, and The Journal of Clinical and Experimental Neuropsychology and the Oxford Handbook of Quantitative Methods.

Current Issues in Education

Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Manuscript received: 08/13/2012
Revisions received: 03/30/2013
Accepted: 04/26/2013

Current Issues in Education

Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Volume 16, Number 2 August 11, 2013 ISSN 1099-839X