Athletic Training Education Journal; 2009;4(2):66-69

www.nataej.org

Development of the Online Assessment of Athletic Training Education (OAATE) Instrument

W. David Carr, PhD, ATC*; Bruce B. Frey, PhD*; Elizabeth Swann, PhD, ATC†
*The University of Kansas, Lawrence, KS; †Nova Southeastern University, Ft. Lauderdale-Davie, FL

Objective: To establish the validity and reliability of an online assessment instrument's items developed to track educational outcomes over time.
Design and Setting: A descriptive study of the validation arguments and reliability testing of the assessment items. The instrument is available to graduating students enrolled in entry-level Athletic Training Education Programs (ATEPs).
Methods: Validity was established with the creation of a national advisory board of Athletic Training educators. Construct validity was established with the creation of a test blueprint to guide the development of items for the knowledge exam. Internal reliability estimates for each domain were calculated. A single scale reliability analysis was conducted using all items. An item analysis was conducted by calculating difficulty and discrimination indexes for each item.
Results: The internal reliability estimates ranged from .23 to .44 suggesting that individual domain scores for this draft of the instrument were not reliable. The single scale total score reliability however, produced an alpha = .84 suggesting a high level of reliability. Difficulty index scores ranged from .03 to .99 (mean = .74 ± .25). Discrimination index scores ranged from -.01 to .41 (mean = .21 ± .09).
Conclusions: While the individual domain reliability was low, the overall single scale score is acceptable. Difficulty and discrimination index scores allowed the removal and revision of items to increase the overall reliability of the test bank.
Key Words: Education outcomes, programmatic evaluation, assessment, accreditation review.

Outcome assessment is a necessary process for all education programs and can be time consuming and cumbersome. Administrators, accreditors, legislators, and prospective students often demand proof of a program's success and how it compares to others,[1,2] Institutions and academic programs must be responsive to internal and external pressures.[3] Assessment systems need to be developed starting with the overall institution level, and progressing down to the individual program level. Outcomes assessment allows the program to determine what curricular areas need modification with the goal of improving the effectiveness of the overall program. Improvement, therefore, is dependent upon ongoing data assessment and use in planning. Longitudinal collection of data that directly relates to program quality or success, followed by a systematic analysis, will allow programs to improve the quality of their graduates. This concept in outcome assessment is 'continuous measurement' for 'continuous improvement'.[4] Snap-shot analysis of outcomes over a short period of time, perhaps in preparation for accreditation, will not give the same depth of insight that is available with longitudinal analysis. A one-time analysis of outcomes will not allow tracking trends across time. Accreditation standards for Athletic Training programs require the routine assessment of outcomes related to clinical and didactic instruction, student learning, and overall program effectiveness.[5] As required for accreditation, numerous institutional and programmatic systems have been developed over the years, some with and most without universal methods that would allow for comparison between programs. The lack of standard criteria among different institutions is a major roadblock to comparison that would allow programs to gauge their performance against a reference or other similar programs.

A wide variety of outcomes are commonly used to measure the success of a program. An even wider variety of methods are used to assess each outcome. Collection and analysis of the data is often the most difficult aspect and largest hindrance to program improvement. Data collection alone without conscientious analysis will not provide adequate information for program improvement. An instrument that will allow program directors to measure outcomes year after year in a consistent manner with a standardized methodology, one that is easy to implement, and whose results are user-friendly is needed to determine if curricular changes are having the desired effect. Assessment systems should also work

Dr. Carr is an Assistant Professor and the ATEP Director at the University of Kansas
wdcarr@ku.edu

Dr. Frey is an Associate Professor in the Psychology & Research in Education Dept. at the University of Kansas
bfrey@ku.edu

Dr. Swann is an Assistant Professor and the ATEP Director at Nova Southeastern University
swann@nova.edu

seamlessly with, and be incorporated into, curriculum designs so that they do not require large amounts of effort and time. An assessment instrument has been developed to address these needs. As the Online Assessment of Athletic Training Education (OAATE) instrument has been developed, the above concepts, indicators of success, longitudinal analysis, and continuous improvement, have been incorporated into the design.

In a recent article [6], Dr. Raehl, of the American Association of Colleges of Pharmacy (AACP), described the creation of an assessment system in general terms:

> "AACP's Pharmacy Education Assessment Services Program (PEAS), is an umbrella of diverse services and tools that will be integral to our programmatic assessment programs…will likely include student and faculty portfolios, peer teaching evaluations tools, and tools for curricular mapping and mapping of curricular competencies to outcomes."

While this system is now in development, it is clear that the Pharmacy profession is moving towards a mechanism for programs to track and monitor various outcomes needed for accreditation.

The field of nursing has several commercially-available education assessment systems.[7-9] These systems include a variety of surveys, licensing exam preparation tests, critical thinking exams, and can be administered via paper-and-pencil or online methods. Most systems offer longitudinal analysis reports and national comparison data. These systems differ from our system on two points: 1) our system is not a commercial entity designed to generate income since our goal is to offer a valid and reliable set of outcome measures at no cost to the users; and 2) our system is designed as a research tool. The data we obtain will be used to gain a better understanding of how well students are being taught, and what factors can be controlled that influence this learning.

This article will address the procedures that have been used to establish the validity and reliability of the OAATE item bank for Athletic Training Education Programs (ATEPs).

Methods

Instrument Design

An online instrument was developed to collect content knowledge and program satisfaction and importance ratings from graduating students. The online instrument was developed using a senior level Management Information Sciences class project. During development, the instrument was hosted on a Windows server, utilizing a Microsoft Access database, and using .Net and XHTML programming language. The system has since been migrated to a Unix server, utilizing a SQL database, and using PHP programming language.

An initial test bank of questions was developed based upon the 12 domains of the 3rd Edition of the Educational Competencies and Proficiencies.[10] The 3rd Edition was used as it was current when the project began. A minimum of ten multiple choice questions with four possible answers were collected from Athletic Training educators for each of the 12 domains. Three of the domains were populated with more than ten questions. For the exam, the instrument randomly generated ten questions for each domain for a total of 120 questions (Figure 1).

After completing a brief demographic survey, the students were prompted to begin the content knowledge assessment. The students had ten minutes to complete each domain and were forced to go in order (Figure 2). Students had access to the test portion of the web site for 30 days and could go at their own pace.

Subjects

ATEPs from around the nation were solicited to participate in the study. Mass emails were generated, messages were posted on Athletic Training list serve web sites, and word-of-mouth methods were used for generating participants among the ATEPs. Institutional Review Board for Human Subjects approval was obtained from the host institution.

Results

After two years of development testing and data collection, 808 graduating students enrolled from 83 different programs. The average age of the students was 22.6 years (range 20-36 years) with 184 females and 113 males. Of those programs, 44 (53%) had every enrolled student complete the entire instrument. Overall, 297 students (36.7%) completed the entire instrument.

Validity

Validity arguments can be drawn based upon expert consensus and content validity. The instrument has been reviewed by an advisory board of educators from around the nation for cultural, gender, racial, geographic, and religious biases. For content validity we developed test items based upon the 3rd Edition of the Education Competencies. The next phase of test bank item development will be based upon the current 4th Edition of the Education Competencies, and as new editions are released the item pool will be reevaluated.[11]

Reliability

Items were grouped into 12 subscales based on the domain they were designed to assess. Internal reliability analyses were conducted on each subscale. (Assessment and Evaluation, Psychosocial, and Modalities items were not analyzed, as each student responded to a random sub-grouping of those items and a reasonable estimate of reliability could not be determined.) From the large initial pool of items, 73 items which lowered reliability as estimated by coefficient alpha, were removed from the subscale. Final internal reliability estimates for subscales ranged from .23 to .44. These alphas indicate very low internal reliability for the 12 subscales, suggesting that individual subscale scores produced by these groupings of items are not reliable. A total score across all items would likely be a more reliable score.

Excluding incomplete items for which data values were missing, all items were analyzed to develop a reliable single total score. This technical requirement for reliability analysis resulted in

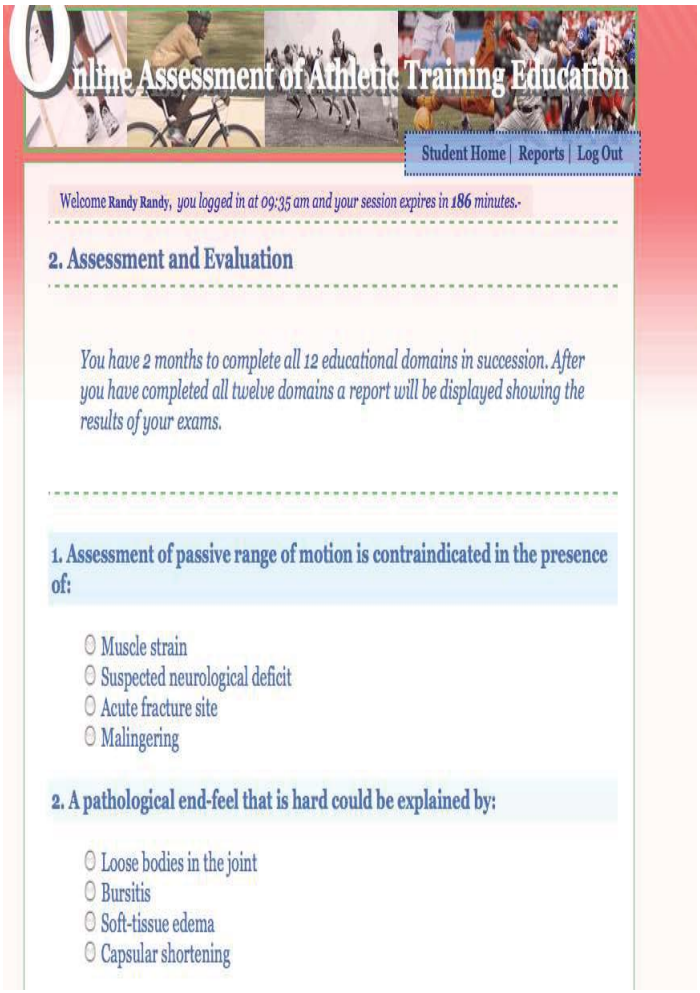

Figure 1. Example Format of Test Items

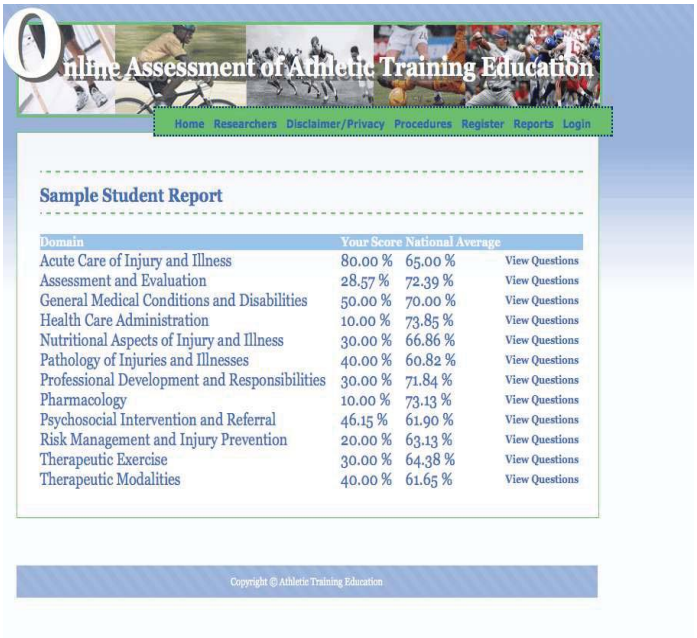

Figure 2. Sample Student Report Illustrating Order of Domains

the exclusion of items meant for the Assessment and Evaluation, Psychosocial, and Modalities subscales. A revised single scale of 57 items that represented all domains, albeit unequally, produced an internal reliability of alpha= .84 This may be a limitation in the use of the current scale's total score. As future revisions attempt to produce independent reliable subscales for each domain, greater equality in sampling across the domains will be achieved. The mean total score for the 57-item test was 74% correct, with a standard deviation of 8%.

Difficulty and discrimination indices

Overall difficulty indices (proportion of students getting an item correct) and discrimination indices (correlation of each item score with the total score) on the 57 item scale were also produced. Item difficulty indices for these items ranged from .03 to .99 (Mean = .74 ± .25). This level of difficulty for the test is appropriate for a measure of professional knowledge. For most students, there will be both easy and difficult items. It will allow for variability in performance and for the eventual setting of a variety of cut scores for different purposes. Item discrimination indices for these items ranged from -.01 to .41 (Mean = .21 ± .09). Item developers typically prefer item difficulty indices between .25 to .90 and discrimination indices greater than .20. These standards will guide future revisions of the instrument.

Discussion

The completion rate of 36.7% is acceptable for survey research, but well below our goal of 100%. For program personnel to draw conclusions about curriculum issues, they need complete information. To address this, we have made two modifications: 1) program personnel will have access to results only when all students for a given graduating class have completed the assessment and, 2) student progress will be displayed on the web interface so that program personnel can determine their status.

Internal reliability scores for the domains were low (.23 to .44) suggesting that the initial test blueprint categories and the items developed for them did not produce internally consistent dimensions. The difficulty in creating and assessing reliable subscales was compounded by three domains having more than ten questions thus students only responded to a random sub-set of questions. A new test blueprint has been developed and items created or assigned to the theoretically stronger domains will be assessed for internal reliability.

Item difficulty is the proportion of test-takers who got an item correct. An item difficulty mean of .74 ± .25 would suggest the test items are of moderate difficulty which should result in psychometrically sound scales. Item discrimination is an indication of how well each individual item correlated with the total test score and is a rough indicator of the validity of an item. In other words, does an individual item measure the same construct as the total test? Discrimination indices above .20 are considered very good, while indices below 0.0 are poor. An item discrimination mean of .21 ± .09 suggests that the items have good validity by this standard.

Limitations

From our initial development phase we have identified a limitation to the current instrument. The individual item scales are not reliable. A revised test item development scheme has been adopted. New items will be solicited from content experts around the nation and undergo a vetting process. Items will be written for the test blueprint that is based upon the latest edition (4th) of the Education Competencies. As new editions of the Education Competencies are developed the test blue print will be revised accordingly. The cognitive competencies from each domain will be used as the test blue print. Several multiple choice items will be developed for each cognitive competency. An initial draft has been submitted and analyzed by the Advisory Board for input, review, editing, and evaluation. Items will be screened for cultural, gender, racial, geographic, and religious biases. A representative sample of programs will be selected for initial reliability testing of the new items. Diagnostic statistical data will be collected and items will be evaluated. New items will be administered nation-wide and reliability studies will be conducted as target numbers of students complete the assessment.

Conclusions

This study has illustrated that drawing conclusions based upon the individual domain scores is not reliable but the use of the overall scale score is reliable. Program personnel need to consider this when looking at possible revisions to their curriculum. We anticipate that the adoption of the new test blueprint and item development scheme will raise the individual domain reliability scores. Subsequent studies will be conducted to determine if this indeed occurs. This outcome measure could be adopted as one piece of a comprehensive assessment system. We are currently developing additional outcome measures (alumni survey, employer survey, and clinical hours tracking) that should improve the marketability of this system and encourage more widespread usage. It is important that program personnel consider all available data sources when deciding when and how to modify a curriculum.

'Continuous measurement' for 'continuous improvement'[4] requires the same measures be completed year after year. The longitudinal design of this study will allow for continuous measurement. It will be left to the program personnel to implement the 'continuous improvement' aspect.

References

1. Tierney, W. *Building the Responsive Campus: creating high performance colleges and universities*. Thousand Oaks, California: Sage Publications; 1999.
2. *Improving the college experience: National benchmarks of effective educational practice.* Bloomington: Indiana University; 2001.
3. Tierney, W. Building the Responsive Campus: creating high performance colleges and universities. Thousand Oaks, California: Sage Publications; 1999.
4. Seymour, D. TQM: focus on performance, not resources. *Educational Researcher*. 1993; 74: 6-14.
5. Commission on Accreditation of Athletic Training Education. *Standards for the accreditation of entryl-level athletic training education programs.* Revised December 2007. Round Rock, Texas: CAATE; 2005.
6. Raehl, C. AACP pharmacy education assessment services: outcomes, assessment, accountability. *Am J of Pharm Ed*. 2008; 72: 1-2.
7. Educational Benchmarking Inc. Available at http://www.webebi.com/. Accessed December 12, 2008.
8. Assessment Technologies Institute. Available at http://www.atitesting.com/. Accessed December 12, 2008.
9. Health Education Systems Inc. Available at http://www.hesitest.com/. Accessed December 12, 2008.
10. National Athletic Trainers' Association. *Athletic Training Educational Competencies*. 3rd Ed. Dallas, Texas: National Athletic Trainers' Association; 1999.
11. National Athletic Trainers' Association. *Athletic Training Educational Competencies*. 4th Ed. Dallas, Texas: National Athletic Trainers' Association; 2005.