Applying the "Test the Class, Not the Students" Methodology to Assessment Results and Course Improvement

David Kern
Northeastern State University

Julia Kwok
Northeastern State University

Isaac Diianni
Northeastern State University

Teko Jan Ernst Bekkering
Northeastern State University

Assessment of learning continues to gain in importance. Demonstrating achieving learning outcomes is already mandated in secondary education, and is predominantly done through standardized tests in multiple-choice format. So far, Higher Education has been exempt. In the near future, assessment of learning may be replaced with government-mandated procedures and testing. Developing proper assessment methods and demonstrating achieving learning outcomes presents an opportunity to pre-empt regulatory mandates. The current research describes the use of a new testing method where the focus is on testing of students as a group, not individually. The methodology is applied to a Management course.

INTRODUCTION

Higher education is frequently criticized for failing to deliver quality education. Not only the knowledge and skills of graduates are subject of criticism, but also the ability of higher education institutions to help students finish their degree programs. Nationally, retention rates for first year to second year undergraduate students have declined to 74.7%, and six-year graduation rates for undergraduate degrees to 55.9% in 2008 (National Center for Higher Education Management Systems, 2009). Secondary education has experienced initiatives to quantify the efficacy of the educational process and increasing accountability through measures like No Child Left Behind (ED.GOV, no date). Higher education has not seen a similar initiative yet, perhaps due to the multitude of accreditation agencies. The Office of Postsecondary Education lists 10 national, 14 regional and 20 hybrid institutional accrediting agencies (U.S. Department of Education, undated). However, accrediting agencies have shifted to an outcomes-based approach under pressure from the Federal Government (Palomba, 2001).

Accreditation of business schools involves voluntary assessment at multiple levels. Universities where they are located, may be required to submit reports to maintain accreditation with regional or

national accrediting agencies recognized by the Department of Education (U.S. Department of Education, undated). Some universities join organizations like the Voluntary System of Accountability (ETS, 2008) in order to meet their assessment needs. For business schools, AACSB and ACBSP are the major voluntary accrediting agencies. AACSB (AACSB, 2009) is widely regarded as the highest level of accreditation and only about 25% of US business schools are able to gain this distinction. ACBSP (ACBSP, 2010b) focuses on accreditation of smaller private and public schools that focus on teaching. Schools frequently use the services of third parties, such as ETS (Educational Testing Service, 2009) with its Major Field Tests , and departments can use more specialized tests such as the ISA exam offered by the Institute for Certification of Computing Professionals (ICCP, 2001) or passing rates for the CPA exams for accountants (AICPA, 2006). Like secondary education however, much of the efficacy testing involves standardized testing in multiple choice format. Multiple choice testing is frequently criticized in a wide variety of educational processes: graduation from high school (Amrein & Berliner, 2003), admissions to undergraduate school (Hoover, 2008b), determining eligibility for merit awards (Hoover, 2008a), and graduate admissions (Curry, 2001). Nevertheless, MC tests continue to be tools of choice in standardized testing. They are easy to score, not prone to disagreement between raters, and inexpensive to administer.

A reason for the focus on assessment as a tool to demonstrate "added value" is the potential of negative financial consequences. Nationally, the National Commission on Accountability in Higher Education advocates budget allocations to stimulate performance in its report "Accountability for better results: A national imperative for higher education" (National Commission on Accountability in Higher Education, 2005). Internationally, the Assessment of Higher Education Learning Outcomes (AHELO) project of the Organisation for Economic Cooperation and Development (OECD) aims to measure learning on a global scale (Lederman, 2010). However, despite the focus on outcomes assessment as a tool for educational improvement, the National Institute for Learning Outcomes Assessment reports that the main use of assessment is the fulfillment of accreditation requirements (Hebel, 2009).

A more appealing reason for assessment in Management education is the concept of performance improvement (PI). It can be applied to improvement of individual and organizational performance. Like the assessment cycles of the AACSB and ACBSP (AACSB, 2009; ACBSP, 2010a), a critical component of the cycle is measurement of achieving goals and objectives. The results are then evaluated and new approaches planned.

This paper is organized as follows. The next section presents how a group of faculty members in our business school has started to use randomized testing to assess learning for the class as a whole, rather than administering identical tests to all students. We then describe the results for a Management class in which we have used this assessment methodology. The paper closes with conclusions and recommend-ations.

ASSESSMENT OF LEARNING IN HIGHER EDUCATION

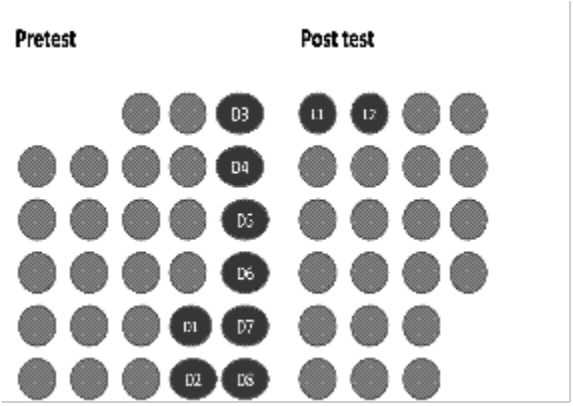
In higher education outcomes assessment, a variety of instruments can be used. Some evaluation instruments measure learning directly, others indirectly through perceptions of stakeholders. White (2007) lists as the most common types of instruments: archival records, behavioral observations, exit interviews, external examiners, focus groups, locally developed exams, oral exams, performance appraisal, portfolios, surveys and questionnaires, and simulations. Even standardized tests can be developed by educators themselves. An example is the IS CORE examination, administered by the ICCP (ICCP, 2001).

The Management Department at Northeastern State University has traditionally relied on student course evaluations and comparisons of average GPAs in the class to assess the efficacy of instruction and instructors. In contrast, the Information Systems and Technology Department (IST) has used multiple choice format pretest/ post-test evaluations for larger service classes. Some problems experienced with that approach follow.

Statistical comparisons of pretests and post-tests are done with t-tests. If this is done as for two samples, the assumption of independent samples is not met. Many students take both tests, some take the

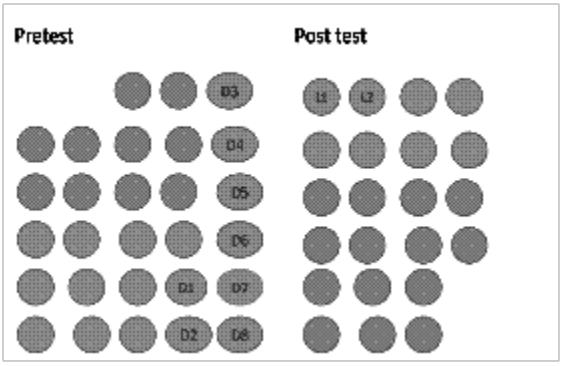
pretest only and drop out, and others end up taking the post-test only when they register late. In the majority of classes, the number of students finishing is lower than the number starting the course. From a statistical point, it is better to use matched pairs (Keller & Warrack, 1999) where the difference scores for each students are used to run a t-test on the gains. From a practical point, this creates bias since weaker students tend to be the ones dropping the class. The reduced sample size will also increase the standard deviation which reduces the power of the test to reject the null hypothesis of no gains. Figure 1 and Figure 2 demonstrate the differences. Students marked with a red circle (D = dropping, L = late) miss the pretest and are not included in the analysis. The bias caused by excluding students in Figure 1, and not meeting the assumption of independent samples in Figure 2 is evident.

FIGURE 1
BIAS CAUSED BY EXCLUDING STUDENTS



Students who register late or drop the class are not included

FIGURE 1
MEETING THE ASSUMPTION OF INDEPENDENT SAMPLES



Overlap between the two samples

Testing time presents another problem. The post-test can be made part of the final exam, but the pretest takes up time at the start of the course. Multiple choice questions with four options have a random probability of scoring at least 25%, and the number of questions has to be substantial enough to measure performance accurately. This can be problematic in classes with only 50 minutes scheduled, where some students will rush through the latter questions and guess at random. Results of identical pretests and post-tests can also be influenced by learning effects and even cheating (e.g. Faulkender et al., 1994)..

Finally, generating custom tests for specific courses tends to produce measurement at the knowledge and comprehension levels, and questions tailored for measuring the application, analysis, synthesis, and evaluation levels as identified by Bloom (1956) was much more difficult. Testing at these higher levels is better done using tests other than multiple choice (MC) formats.

Using a pretest/ post-test design does have one advantage: it creates a baseline for comparison. When standardized norms are not available, the results of testing at exit only can be skewed by prior experience, prior knowledge, and prior skills

Testing students individually may be necessary for assigning grades, admission decisions, and pass/fail decisions; but outcomes assessment is in essence a group assessment. Students are educated in groups, and decisions like accreditation are based on results for groups rather than individual students. Rather than using identical tests for all students to evaluate outcomes, the efficacy of instruction may be better measured at the group level.

In a new group testing design piloted in selected courses over the last two semesters, students take tests created using random selection of items from a larger test pool (Bekkering, Kwok, & Kern, 2010; Kwok, Bekkering, & Kern, 2010). The size of the group, the size of the test pool, and probability of selection at random determine the minimum number of items selected for each student. Larger groups allow for larger test pools with more detailed testing or using fewer items per student. Smaller groups can often still be adequately tested if the number of items in the pool is small enough and the proportion of items selected from the pool per student is large enough. The probability that each test item will be used at least once in the group can be calculated with the formula $P = 1 - \left(\frac{I-n}{I}\right)^S$, where S is the number of students, I the number of items in the pool, and n the number of items selected for each student. The probability of selection at least twice is $P\left(2^+\right) = 1 - \left(\frac{I-n}{I}\right)^S - \frac{1}{I}$ (Bekkering et al., 2010; Kwok et al., 2010). In creating the test pools and deciding on the number of items per student, we used a minimum probability of selection once of .05, since this is generally considered the minimum standard for statistical significance.

Kwok, Bekkering and Kern (2010) used 15 items from a 40-item test pool in two Finance courses. In one class, 12 students took the pretest, 10 took the post-test, and 7 took both. In the other class, 14 took the pretest, and 13 took both tests. Bekkering, Kwok and Kern (2010) used 3 questions per student from a pool of 9 items in a class of 9 Information Systems students. In all cases, each item in the pool was selected by at least one student. The test pool in the Finance courses consisted of multiple choice questions with four options, but the pool in the Information Systems course contained only tasks like creating diagrams, working with database files, and creating actual forms and reports. Especially the latter demonstrates that testing items can take any form. Whereas using a large enough number of testing items virtually dictates the use of multiple choice questions for identical pretests and post-tests, using only part of the pool enables the use of Essay questions require students to demonstrate skills, and even a combination of formats within the same group assessment. The limited number of items allows sufficient time to complete tests that do not rely only on MC format. In this paper, we will describe the results for a Management class where the pool consisted of a mix of True/False questions and multiple choice questions with four options. A future paper will describe the use of short essay questions and drawing diagrams in an Economics class.

To measure student performance at the start and at the end of the course, the scores for all students on a test are added, and the group scores compared. Essentially, this creates a sample of one (1). Students who register late and miss the pretest are assigned a zero (0) score on the pretest; students who drop the class are assigned a zero score on the post-test (Figure 3). As long as the number of students dropping the class exceeds the number of students coming in late, this would appear to be a reasonable correction for partially overlapping samples. A more conservative approach of deleting students joining late can also be used (Figure 4) If the sum of scores on the pretest still exceeds the sum on the post-test, improvement of performance is amply demonstrated.

FIGURE 3
POOLING OF SCORES - REGULAR

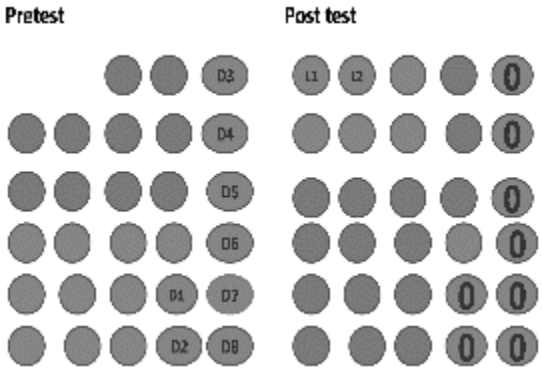
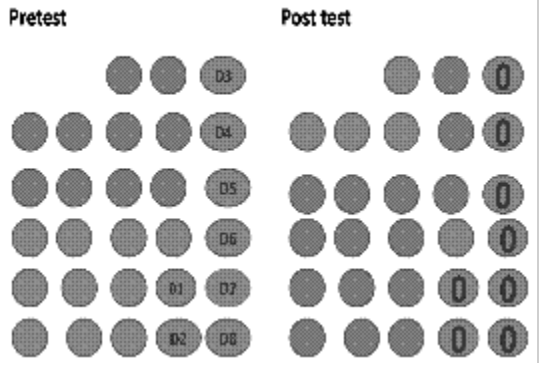


FIGURE 3
POOLING OF SCORES - CONSERVATIVE



APPLICATION

The Operations and Supply Chain Management course is a junior-level core requirement for all business administration students. The course is offered primarily in a blended format, whereas students work both online through the Blackboard virtual learning environment, and in-class with team learning approaches. The content of the course includes qualitative information about principles of operations and supply chain management, and quantitative approaches to problem solving in the same environment. Learning techniques in the course include quizzes for each chapter, multiple application problems (quantitative) for each chapter, substantial Discussion Board participation, in-class review of actual business applications, and three exams during the semester. The quantitative element of the course tends to challenge students with less-developed mathematics skills.

The course is taught by multiple instructors each semester, who utilize a common final exam to assess overall student learning. As a pilot program to improve assessment and course design for this course, one of the instructors developed a pretest/posttest assessment based on the random selection methodology outlined above.

In the spring semester 2010, pre-test and post test assessments were completed in two sections of Operations and Supply Chain Management. Both sections were taught by the same instructor employing identical delivery and feedback approaches, utilizing the facilities of Blackboard. Students were provided detailed instructor notes, which complemented a text that effectively addressed course objectives.

Sixty (60) students took the pretest with nineteen (19) out of sixty-two (62) items selected by a Random Block function in Blackboard. Fifty-five (55) students completed the post test with nineteen (19) items randomly selected from the same sixty-two (62) items used in the pretest. The pretest and post test

included thirty-seven (37) knowledge items, twenty-one (21) comprehension items and four (4) application items taken from the quiz and exam test bank.

Students completed the pretest online during the first week in class prior to being exposed to the course material. The post test was completed during the last week of the sixteen-week semester after student's finished all course work. Items selected for the pre and post tests covered eleven course objectives. The number of items for each objective was determined by a combination of perceived importance, the number of items available from the master test bank items, and the complexity of the item (analytical items take much longer to answer than content items). All items were structured as multiple-choice or true/false, consistent with the examination practices employed for this course in all sections and locations. Three objectives addressed supply chain application; eight addressed specific content areas. The number of items per objective ranged from one (1) to eleven (11).

TABLE 1
GENERAL STATISTICS

PRETEST/POST TEST INFORMATION	**Data**
Items per student (random select)	19
Total Items in Pool	62
No. of Pretests	60
No. of Post tests actually taken by students	55
No. of post tests with zero scores (reflects students dropping the course)	5
No. Application Items	4
No. of Comprehension Items	21
No. of Knowledge Items	37
No. of True/False Items	29
No. of Multiple Choice Items	33
Number of Course objectives	11
Course objectives with 1-3 items	5
Course objectives with 7-11 items	6

There were five (5) fewer students in the post test than in the pretest. This was a result of students dropping the course. A conservative approach is to assume that no items were answered correctly on five (5) post tests that match the pretests of students who dropped the course. The comparative results shown in the following section employ this method.

Before initiating the pretest, we calculated the probability of an item not being selected as less than .001%. The probability that items would be selected at least twice was greater than 99.9%. The probabilities were virtually the same for the post test. Actual item selection resulted in a minimum of nine (9) repetitions for pretest items and a minimum of five (5) repetitions for post test items.

RESULTS

The summary results of the pre and post test using the random sampling methodology are provided in table 2. The pretest scores averaged 51%, which was 13% above the predicted score of 38% (which is associated with guessing on the true/false and multiple choice items). The beginning performance appears to represent prior knowledge of the information required to answer the items. Four students had completed the course in a previous semester, but did not achieve a C or better, a requirement for this core curriculum course. The scores of these students increased the pretest scores modestly; however, a more

detailed review of the results indicates that some items addressed information that has been covered in courses taken by students in prior classes.

As shown in Table 2, post test scores of 79% (including zero correct answers for 5 post test results) showed an improvement of 28% from the pretest score of 51%. Overall, this indicates a positive change in student learning, including zero scores for 5 post tests.

TABLE 1
SUMMARY OF SCORES

	Max score	Raw score	Avg Score
Total Pretest	1140	587	51%
Post test	1140	899	79%
Improvement			28%

We analyzed the results based on the type of item asked, by the learning category, by course objective and by individual item. The value of completing this review is in identifying specific weaknesses in the structure of the assessment tool. One of the benefits of employing the pooled randomized testing approach is that the number of items in the pool can be increased for any objective or type of item, even including items with greater time requirements. The following evaluation identifies specific weaknesses to be corrected in future assessments. Additionally, the weaknesses identified can be generalized to other courses and sections.

An evaluation of the results (table 3) indicates some issues with specific items, the type of items and the limited number of application items included in the item pool. Eight (8) items resulted in post-test scores under 70%, a performance level below the targeted threshold. Notably, only 50% of the application items resulted in scores over the 70% threshold, indicating a need for more effective learning processes.

TABLE 3
DETAILED ANALYSIS

RESULTS OF DETAILED ANALYSIS	Data
No. of Items below 70% post-test	7
No. of Items with lower post scores than pretest	2
Average Improvement Application Items (all multiple choice)	16%
Average Improvement in Comprehension Items	26%
Avg Improvement Knowledge Items	40%
Avg Improvement True/False	17%
Avg Improvement Multiple Choice	39%
No. of Course Objectives with no improvement	1
No. of Course Objectives with improvements of 10-25 percentage points	3
No. of Course Objectives with improvements of 26 -39percentage points	3
No. of Course Objectives with improvements of 40-47 percent points	4

The overall improvement of 16 percentage points for application items is substantially below the improvement registered for knowledge items. The improvement for comprehension items is in between the two. Together these results indicate lower performance on items involving more complexity (comprehension and application).

It is not surprising that the improvement in multiple choice items is more than double that of true/false items, because random selection would result in a pretest score of 50% on true/false items versus 25% on multiple choice items.

As exhibited in Table 4, which details results for course objectives, only one objective (1) failed to show at least a 13% improvement from pre to post-test. The results for four objectives indicate very substantial improvement (in excess of 40 percentage points). These results provide evidence of meaningful learning in most of the course objectives.

TABLE 4
RESULTS BY OBJECTIVE

Course Objective	No. of Items	% Correct Pretest	% correct Post test	Improvement
Productivity Application	1	40%	54%	14%
Purchasing Application	1	63%	84%	22%
Inventory Application	2	37%	69%	32%
Supply Chain Management	8	52%	78%	26%
Process Improvement	1	83%	83%	0%
Quality Management	8	45%	87%	42%
Forecasting methodology	7	44%	91%	47%
Sourcing Management	3	43%	86%	43%
Purchasing Management	9	53%	94%	41%
Logistics Management	11	41%	88%	47%
Inventory Management	11	73%	85%	13%

Three objectives with the smallest improvements were assessed with only one (1) item. All but one (1) high improvement objective included seven (7) or more items in the test pool. These results may indicate that greater emphasis was placed on the high improvement objectives during the course.

CONCLUSION

By employing the random selection approach reviewed in this paper, an effective pretest was completed by students that provided a baseline for posttest results at the end of the semester. By identifying the weaknesses of assessment reviewed in this paper, improvements in the structure of the assessment tool have been initiated for future classes.

The major issues relative to the current assessment approach include the following:

- Use of true/false questions limits the potential improvement over the 50% random success rate. The number of true/false questions will be limited or completely eliminated from future assessments.
- Items that have high pretest scores indicate prior knowledge by students, and may not be meaningful measures of learning in this course. These items will be replaced with items that will focus on new learning.
- Student's application skills at the beginning of the course are at a lower level than comprehension skills, which in turn are lower than knowledge skills. Moreover, application learning improves less than comprehension, while comprehension learning improves less than knowledge. This suggests that there is a need to increase comprehension and problem solving learning opportunities in both upstream courses and in this course.

- The size of the item pool for at least five (5) of the course objectives included 3 or fewer questions. More questions will be added to the pool for each objective.
- Items that show regression or low post-test scores should be re-evaluated and improved, replaced with more effective items, or require a different approach to learning.

Improvements have been implemented for future assessments by expanding and altering the item pool. All true/false questions have been eliminated. The number of items for each course objective has been increased, particularly application questions, which require more time to complete. In order to address lower performance on application questions, the number of in-class sessions devoted to review has been increased. A key consideration in the changes already completed and those contemplated is the benefit of employing the random sampling technique discussed in this paper.

REFERENCES

AACSB. (2009). Eligibility Procedures and Accreditation Standards for Business Accreditation. from www.aacsb.edu/accreditation/business/STANDARDS.pdf

ACBSP. (2010a). ACBSP Conference 2010: Pre-Conference workshops. Retrieved Sept 14, 2010, from http://www.acbsp.org/p/cm/ld/&fid=119

ACBSP. (2010b). Accreditation home. Retrieved June 14, 2010, from http://www.acbsp.org
AICPA. (2006). American Institute of CPAs. Retrieved Mar 19, 2011, from http://www.aicpa.org/BecomeACPA/CPAExam/Pages/CPAExam.aspx

Amrein, A. L., & Berliner, D. C. (2003). The Testing Divide: New Research on the Intended and Unintended Impact of High-Stakes Testing. *Peer Review, 5*(2), 31-32.

Bekkering, E., Kwok, J., & Kern, D. (2010). *Group assessment of learning: Test the class, not the students*. Paper presented at the Information Systems Educators Conference.

Bloom, B. (1956). *The Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook I: Cognitive Domain*. New York: Longmans, Green.

Curry, D. (2001, July 13). Texas Law Limits Use of Standardized Tests in Graduate Admissions. *Chronicle of Higher Education*.

ED.GOV. (no date). NCLB - overview. Retrieved January 31, 2010, from http://ed.gov/nclb/landing.jhtml

Educational Testing Service. (2009). Educational Testing Services. Retrieved January 31, 2010, from http://www.ets.org

ETS. (2008). *Measuring Learning Outcomes in Higher Education Using the Measure of Academic Proficiency and Progress (MAPP)* (No. ETS RR-08-47).

Faulkender, P., Range, L., Hamilton, M., Strehlow, M., Jackson, S., Blanchard, E., et al. (1994). The Case of the Stolen Psychology Test: an Analysis of an Actual Cheating Incident. *Ethics & Behavior,* 4.

Hebel, S. (2009, June 15). Many Colleges Assess Learning but May Not Use Data to Improve, Survey Finds. *Chronicle of Higher Education*.

Hoover, E. (2008a, April 20). Admissions Group Seconds Testing Commission's Call for Change. *Chronicle of Higher Education*.

Hoover, E. (2008b, September 29). At Admissions Conference, 3 Questions About Standardized Tests. Retrieved January 31, 2010, from http://chronicle.com/article/At-Admissions-Conference/1201

ICCP. (2001). IS CORE (outcome assessment) Examination. Retrieved June 14, 2009, from http://www.iccp.org/iccpnew/outlines.html#22

Keller, G., & Warrack, B. (1999). *Statistics for management and economics* (5th ed.). Pacific Grove, CA: Duxbury Resource Center.

Kwok, J., Bekkering, E., & Kern, D. (2010, May 27-30). *Assessment of learning: Test the class, not the students.* Paper presented at the Hawaii International Conference on Business, Honolulu, HI.

Lederman, D. (2010). Measuring Student Learning, Globally Retrieved January 31, 2010, from http://www.insidehighered.com/news/2010/01/28/oecd

National Center for Higher Education Management Systems. (2009). Graduation Rates. Retrieved Mar 19, 2011, from http://www.higheredinfo.org/dbrowser/index.php?submeasure=27&year=2008&level=nation&mode=graph&state=0

National Commission on Accountability in Higher Education. (2005). *Accountability for better results: A national imperative for higher education*.

Palomba, C. A., & Banta, T. W. (Ed.). (2001). *Assessing student competence in accredited disciplines*. Sterling, VA: Stylus Publishing.

U.S. Department of Education, O. o. P. E. (undated). The database of accredited post-secondary institutions and programs. Retrieved Mar 19, 2011, from http://www.ope.ed.gov/accreditation/

White, B., & McCarthy, R. (2007). The Development of a Comprehensive Assessment Plan: One Campus' Experience. *Information Systems Education Journal, 5*(35).