Group Assessment of Learning: Test the Class, Not the Students

Teko Jan Ernst Bekkering
bekkerin@nsuok.edu
Department of Information Systems and Technology

Julia Kwok
kwok@nsuok.edu
Department of Accounting and Finance

David Kern
kernd@nsuok.edu
Department of Business Administration
Northeastern State University
Tahlequah, OK 74464

Abstract

Assessment of learning has become a major issue in education in recent years. Secondary education has already been confronted with demonstrating learning outcomes. This is predominantly done through standardized tests in multiple-choice format. Higher education has not been confronted with demands for standardized testing, and assessment of learning takes place as part of accreditation and in voluntary alliances. Developing proper assessment methods presents an opportunity to pre-empt regulatory mandates. The current research describes the development of a new methodology to demonstrate student performance in higher education. Results for the first semester application in an Information Systems Design class are presented.

Keywords: assessment, outcomes assessment, course assessment, program assessment, accountability, testing

1. INTRODUCTION

Assessment of learning is the most important factor affecting learning outcomes, and even influences teaching and learning activities themselves (Biggs, 2003). Increasingly, assessment is presented as part of a cycle (AACSB, 2009; ABET, 1998). Unfortunately, a frequent focus of discussion is not about realizing the benefits of learning assessment, but on problems associated with the measuring process. One of these areas of criticism concerns the use of Multiple Choice (MC) tests. Criticism occurs 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 are the tools of choice in standardized testing. They are easy to score, not prone to disagreement between raters, and inexpensive to administer.

At the national level, the No Child Left Behind Act of 2001 is the best-known example of standards-based educational reform, the attempt to improve educational outcomes by requiring states to set standards and measurable goals in secondary education (ED.GOV, no date). Until now, higher education has not been faced with mandatory standardized out-

comes assessment. The current practice does involve voluntary assessment at multiple levels, usually mandated by accrediting agencies. Universities may be required to submit reports to maintain accreditation with regional or national accrediting agencies recognized by the Department of Education (ED.GOV, 2010). Universities join alliances such as the Voluntary System of Accountability (ETS, 2008) in order to meet their assessment needs. Business schools may be accredited by AACSB (AACSB, 2009) or ACBSP (ACBSP, 2010), and Information Systems departments may be accredited by ABET (ABET, 1998). Business 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). In other words, a multitude of agencies and institutions now expect outcomes assessment to be performed; and assessment takes place at many different levels.

One reason for this focus on assessment as an instrument of demonstrating "added value" is the potential financial consequence of failure to improve learning. The National Commission on Accountability in Higher Education has already advocated 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). Assessment projects take place at the international level. 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). Despite all this focus on outcomes assessment as a tool for educational improvement, however, a survey of the National Institute for Learning Outcomes Assessment reported that the main use of assessment is the fulfillment of accreditation requirements (Hebel, 2009). This paper is organized as follows. Part 2 presents past efforts in our department to use MC tests in pretest-post test format. Next, we discuss how randomly selected test items in any format can be used to assess learning for the class as a whole, rather than as the aggregate for all students in the class. Part 4 describes the results for one of the classes in which we have used this assessment methodology. The paper concludes with conclusions and recommendations.

2. ASSESSMENT OF LEARNING IN HIGHER EDUCATION

In secondary education, outcomes assessment is predominantly performed using MC tests. In contrast, in higher education, a variety of instruments are used. In general, evaluation instruments can be classified as direct methods where students demonstrate learning, and indirect methods, which rely on review of documents or on the subjective opinions of other individuals. 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 Information Systems and Technology Department (IST) at Northeastern State University has used objectively scored pretest/post-tests in MC format for course assessment in larger service courses for several years. Some of the problems that surfaced in using this approach include the difficulty using proper statistical methods, limited test-taking time leading to rushed answering, and the development of good MC tests is very difficult.

When comparing the results of the pretest and the post-test, the general practice is comparing the means of the tests with a t-test. Not only may this violate the assumption of equal variances, but moreover, the assumption of independent samples is not met. There is significant overlap between the class before and after, but students who drop the class and students who come in late and miss the pretest cause the overlap to be only partial. Statistically, a more appropriate solution is to use matched pairs (Keller & Warrack, 1999). However, the results are biased because some students, presumably the poorer students, drop the class. The reduced sample number also increases the standard deviation and may cause the means to be not statistically different.

Another problem encountered was the limited time to take the tests. Since some classes only have 50 minutes scheduled, students frequently rushed through the latter part of the test to finish all answers. This could clearly be seen in the analysis of the results, where answers showed more a random pattern and the aver-

age score on a question dropped. The problem of limited time was increased because by their very nature, MC tests produce scores of 25% when taken completely randomly (assuming four answers per question).

Finally, we found that while MC questions appropriately measure cognitive levels of knowledge and comprehension, it was much more difficult to measure at the application, analysis, synthesis, and evaluation levels as identified by Bloom (1956). Testing at these higher levels is better done using tests other than MC format.

The single strength of our approach however, is the provision of a baseline. Whereas exit-only testing can be skewed by prior experience, prior knowledge, and prior skills, repeated tests before and after more clearly demonstrate the effects of the course or the program in between. This is not absolute, since the results can be influenced by effects of learning the test, however minute. To control for these effects, control groups can be used (Figure 1); additional groups may be needed to measure the effect of repeated testing (Figure 2) (Asynchronous Learning Networks, 2001) . Finally, repeated testing with identical tests creates the risk of security breaches (e.g. Faulkender et al., 1994).

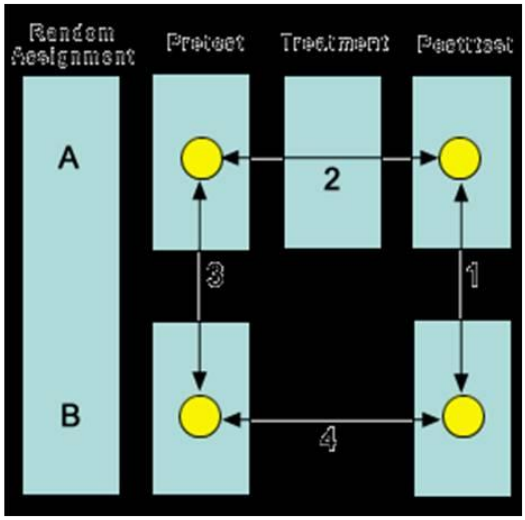

Figure 1 - Use of control groups

All educational testing, whether with baseline testing or without, has one thing in common: testing is done on an individual student basis. This may be necessary for assigning grades or for pass/fail decisions, but course and program assessment is essentially an evaluation of the performance of groups rather than individuals in the group. Instruction is to the group, and individuals are tested with the identical tools. The next section presents an approach that allows assessment of learning at the group level through pooling of randomly created tests from a larger test pool in a pretest/post-test design.

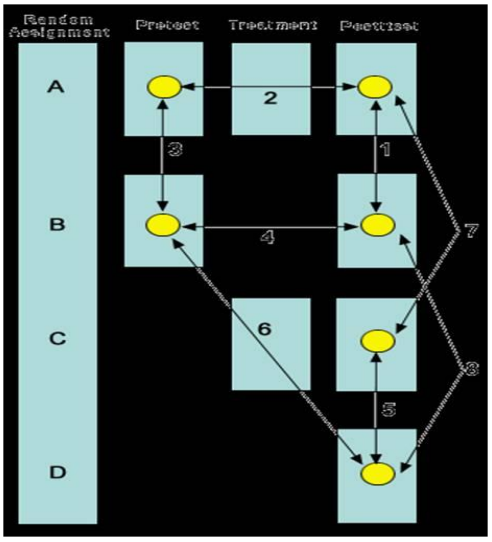

Figure 2 - Solomon four-group design

3. POOLED RANDOMIZED TESTING

In this new testing design, individual students in the group (class, major, program, university) take tests created using random selection of items from a larger test pool. Depending on the size of the group, the number of items used for a student can be larger or smaller in relation to the size of the test pool. Larger groups require fewer items or the pool can be larger, or a combination of both. Smaller groups require more items or a smaller test pool, or a combination. The issue is one of distribution of test items *over the group*. Ideally, all items will be equally represented over the whole group. Of course, this is seldom true.

However, the probability that each test item will be used at least once in the group can be easily calculated with probability theory. Let S be the number of students, I the number of items in the pool, and n the number of items selected for each student. The probability that a student does not have an item selected is $\left(\frac{I-n}{n}\right)$, and that the class does not select an

item $P=\left(\frac{I-n}{I}\right)^S$. The probability that an item has been selected at least once in the group is $P=1-\left(\frac{I-n}{I}\right)^S$, and at least twice $P\left(2^+\right)=1-\left(\frac{I-n}{I}\right)^S-\frac{1}{I}$.

Using these formulas, test creators can decide on the size of the test pool to be developed and the number of items to be selected since the number of students S is a given (the number of students in the class at the time of the pretest). Different test developers can determine their own norms, but we decided to use a cutoff for the probability of test items not selected to be less than .05 . This is an arbitrary level, but has the precedence of being used for statistical use. We also consider overlap between pretest and post-test for individual students, but will not address this issue due to space limitations. The interested reader is referred to Kwok (2010).

For the purpose of this research, the actual test items can take any form. Test creators can use MC questions, use essay questions, require students to demonstrate a skill, and even a combination of formats within the same group assessment. Since only a selected number of items from the larger pool is used for each student, there is sufficient time to complete tests that do not rely only on MC format.

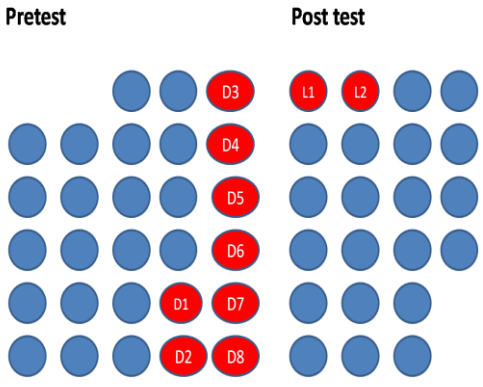

Figure 3 - Paired t-tests

Analysis for both pretest and post-test is done at the group level. Rather than comparing paired scores for individual students or the average scores for pretest and post-test, the scores for all students on a test are added. The group score on the pretest and the group score on the post-test are compared. This essentially creates a sample of one (1). The traditional approaches of paired t-tests and t-tests of groups are shown in **Error! Reference source not found.** and **Error! Reference source not found.**. Students marked with a red circle are not included in the analysis, D indicates students who drop the class, and L indicates students who register late and miss the pretest. The bias caused by excluding students in the first case, and the violation of the assumption of independent samples in the second, are evident.

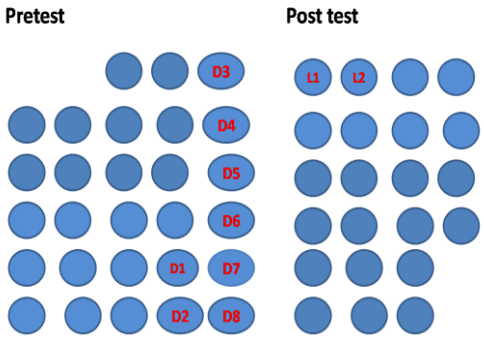

Figure 4 - T-test of groups

In our new testing design, students who register late are assigned a zero score on the pretest and students who drop the class are assigned a zero on the post-test. This assumes that the number of students dropping the class exceeds the number registering late, but it does create equal samples before and after without introducing bias by excluding poor students. An even more conservative approach can be used by including the students who drop with a zero score on the post-test, but excluding the post-test scores of the students registering late. If the sum of scores on the pretest still exceeds the sum on the post-test, improvement of performance is amply demonstrated.

One crucial element is the proper formulation of test pool items. If students can score sufficiently high on the pretest due to prior skills or prior knowledge, the combined scores on the post-test may well be lower than the pretest due to sheer force of numbers. Ideally, test items in the pool are sufficiently difficult that no students can do them at the time of the pretest, and the instruction targeted and effective enough to allow students to score high on the post-test. We accomplish this by basing items in the test pool on the course goals, and by focusing on achieving the course objectives

in class – something which should be done in any case. Currently, we have piloted this approach in Information Systems (IS), Finance, and Management classes. In the following section, we will describe how this approach has worked in one IS class, IS4213 – Systems Design.

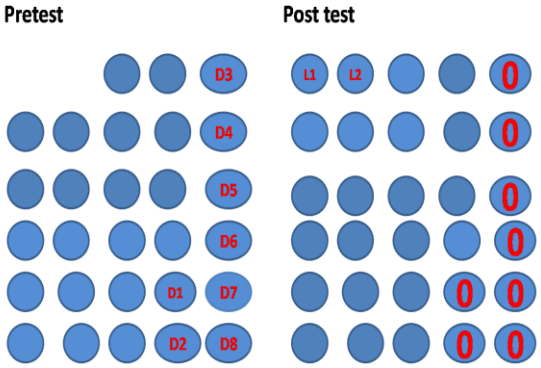

Figure 5 - Pooling of scores - regular

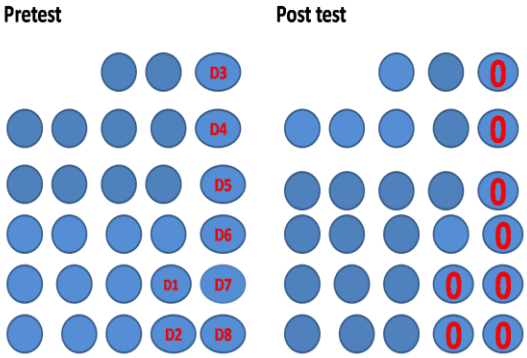

Figure 6 - Pooling of scores - conservative

4. APPLICATION

In the IST department, Systems Analysis and Systems Design are separate courses. Whereas Systems Analysis focuses on analyzing business requirements and logical modeling, the Systems Design class focuses on physical models and implementation. Course objectives for the course are as follows:

1. Students can use forms, reports, and other sources to create ERDs (Entity Relationship Diagrams).
2. Students can use simple data integrity controls
3. Students can design forms , reports, interfaces, dialogues , screens, network diagrams
4. Students can develop a simple test plan
5. Students can create user documentation

Based on the course objectives, the test items in

Table **1** were developed.

The Design class is a senior level class, and students have to pass Systems Analysis as a prerequisite. Consequently, the class is taken mostly by IS majors. The class size is usually limited, and we therefore determined to use three items per student from a pool of nine items. This is a relatively small number due to the limited class size and using tasks for testing. In other classes, we have used 40 to 100 items. With an enrollment of 18 students (one registered late), we calculated the probability that each item would be selected at least once as 99.93% and that it would be selected at least twice as 88.82%. For the post-test, we used all nine items for the final exam but only used the first three as post-test. Neither the pretest nor the final exam was a timed test, and the items were randomly selected using the "random block" feature on Blackboard.

Not all students ended up taking the pretest. The course is an internet course, and some students tend to start late due to procrastination. From the eventual 19 students we used 12 students who did either the pretest or the post-test, or both. Since the number of dropping students (7) far exceeded the number of late registrants (1), we considered this justified. The number of students dropping was significantly higher than in other years, and should not be considered to be a normal condition. Students were cautioned that the pretest was a reflection of the level of performance expected at the end of the course, and that anything in the pretest was going to be taught during the semester. The rigor of the test, based on the principle that students should not be able to do the test before, and all students should be able to do it afterward, is reflected in the lack of file submissions on the pretest (marked in red in Appendix 1). One student wrote in the Word file to be submitted "I have no idea how to do this" (student02, item 4). Five students opened the pretest but did not submit any files at all. When asked for the reason, they indicated that they did not think that they would get any points for their answers anyway.

Table 1 - Test items

	Item description	Deliverable
1	Use screen shots of input forms and reports to create an ERD of the underlying database	Visio 2007 ERD
2	Use a description of a system to create a logical database design in third normal form with all keys	Word file with description in 3NF
3	Use an existing database to create proper integrity constraints in the database tables	Word file with list of constraints and database with constraints implemented
4	Use an existing database with only tables to construct input forms for all tables	Visio 2007 Windows XP User Interfaces
5	Use an existing database to create specific detail reports and summary reports	Visio 2007 Windows XP User Interfaces
6	Use the dashboard in an existing database to create a diagram of the dialogue flow	Visio 2007 Flow Diagram
7	Use a case description to create a network diagram	Visio 2007 Network Diagram
8	Use an existing database with integrity constraints to create a test plan	Word file with test plan
9	Use an existing document in plain text to format a user manual	Formatted Word 2007 file

Another observation on the pretest was the submission of the wrong files or the wrong diagrams. For instance, an ERD might be submitted instead of a flow diagram (student12, item 6). On the final exam (and consequently the post-test), two students used the "Save" button instead of the "Submit" button, and consequently files were not available for grading. Both had been doing very poorly submitting any work in the course, were assigned zero points for the post test, and included in the analysis.

The results are listed in Appendix 1. Even though a significant number of students (7) dropped the class, the improvement in group score was significant from 28 to 128. The post-test average score per item was relatively low (3.6/10) due to the high number of students dropping the class, however. The average score on the test items also indicated that all test items improved over the course, as demonstrated in Appendix 2. The same appendix demonstrates that even though not all items were selected an equal number of times, all were represented both in the pretest and the post-test (range: 1-5).

5. CONCLUSIONS

Based on the results with a relatively small number of students, our new approach of using a small number of randomly selected test items from a larger pool appears to work well. In several courses, distribution of the test items has been as expected – not exactly equal but all items in the pool were used in the class. Despite several handicapping factors (students dropping the class, small class size, and students missing points on the post-test), the combined score for the class still improved significantly. As expected, the use of non-MC test items has significantly improved the face validity of the assessment instrument, and we can test at a higher level of Bloom's cognitive levels: application and analysis. With larger classes and fewer items, we may be able to test at even higher levels.

We are currently working on a more detailed description of our approach and the underlying statistical principles. This includes, but is not limited to, a more complete description of the generation of test items based on course objectives (and program objectives), differences between our approach and Computer Adaptive Testing (CAT), segmentation of test items based on course objectives in analysis, mixing items of varying difficulty levels, and mixing items of different types within the same test pool. Moreover, expanding the use of the approach to the major (IS, FIN, MGMT, and other) level will be pursued. The results we obtained in the spring 2010 semester will be presented at other conferences within our disciplines, and published in journals within our respective disciplines.

6. REFERENCES

AACSB. (2009). Eligibility Procedures and Accreditation Standards for Business Accredi-

tation. from www.aacsb.edu/accreditation/business/STANDARDS.pdf

ABET. (1998, June 14, 2010). Computing accreditation criteria. Retrieved June 14, 2010

ACBSP. (2010). Accreditation home. Retrieved June 14, 2010, from http://www.acbsp.org

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.

Asynchronous Learning Networks. (2001, 12/2005). Pretest/Posttest Comparison. from http://www.alnresearch.org/HTML/AssessmentTutorial/Strategies/PrePostComparison.html

Biggs, J. (2003). *Teaching For Quality Learning at University.*: Open University Press.

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. (2010, 6/10/2010). Regional and National Institutional Accrediting Agencies Retrieved June 14, 2010, from http://www2.ed.gov/admins/finaid/accred/accreditation_pg7.html

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 Measuring Student Learning, Globally

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

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

Appendix 1.

	pretest items			pretest scores			post test items			post test score			Grade
student01													F
student02	9	3	4	6	0	*0*	6	4	5	1	7	*0*	B
student03	5	6	4	2	3	2	6	8	4	8	8	7	A
student04													F
student05													W
student06	5	7	8	*0*	*0*	*0*				*0*	*0*	*0*	W
student07				0	0	0	1	9	3	8	7	8	C
student08	1	5	9	*0*	*0*	*0*	7	1	8	*0*	*0*	*0*	F
student09	3	2	6	*0*	*0*	*0*				*0*	*0*	*0*	W
student10													W
student11	4	6	7	*0*	*0*	*0*	7	8	9	*0*	*0*	*0*	F
student12	5	8	6	*0*	*0*	1	5	4	3	3	2	4	A
student13													F
student14													W
student15	8	7	4	6	3	1	3	2	4	9	10	9	A
student16	5	3	8	2	2	*0*	7	3	6	6	8	4	A
student17													W
student18	6	1	4	*0*	*0*	*0*				*0*	*0*	*0*	W
student19	8	1	3	0	0	0	5	3	6	7	6	6	A
				Total Points		28				Total Points		128	
				Avg/item		0.8				Avg/item		3.6	

Appendix 2

Pretest						Post-test		
item	#	avg				item	#	avg
1	3	0.0				1	2	4.0
2	1	0.0				2	1	10.0
3	4	0.5				3	5	7.0
4	5	0.6				4	4	6.3
5	5	0.8				5	3	3.3
6	5	0.8				6	4	4.8
7	3	1.0				7	3	2.0
8	5	1.2				8	3	2.7
9	2	3.0				9	2	3.5
avg	3.7	0.8				avg	3.0	3.6