INITIAL DATA ON A TEST BANK ASSESSING TOTAL ENTERPRISE SIMULATION LEARNING

Jerry Gosen , University of Wisconsin- Whitewater
John Washbush, University of Wisconsin- Whitewater
Tim Scott, Mankato State University

ABSTRACT

This paper is the second in a series dealing with the construction of a test bank of items designed to assess whether or not learning takes place from playing a total enterprise simulation. It describes the procedure for and the results of an initial assessment of a selected number of items of the bank. The results show post test scores significantly higher than pre-test scores in three of the four classes sampled and reliability scores near .66. Difficulty and discrimination correlations for the items used are also reported.

BACKGROUND

This effort was initiated partially in response to a seminar for instrument design assessing the effectiveness of simulations at the ABSEL conference in New Orleans in 1997 (Anderson and Lawton, 1997b). In the following year, five scholars, namely Dick Cotter, Jerry Gosen, Alan Patz, Tim Scott, and John Washbush created a list of learning objectives for simulations, and an overlapping set of scholars, namely Jerry Gosen, Al Patz, John Wasbush, and Joe Wolfe, furnished items which could be used to measure whether an objective could be accomplished. The result was 119 multiple choice and short essay items designed to measure 40 learning objectives, which ranged from the simple and concrete, such as 'understand the consequences of specific decisions such as issuing stock,' to the abstract and comprehensive, such as 'adapt strategy and decision making to changing circumstances.' Of the items, 102 were multiple choice, eight were short essay, six involved analysis of simulation-generated financial statements, four involved analysis of a hypothetical income statement, and one asked students to study hypothetical marketing information[1]. The procedure to construct the bank and a pilot to provide limited reliability and validity statistics was presented at the ABSEL conference in Philadelphia last year (Gosen et al., 1999).

This project was proposed in the context of criticism of the simulation field for not defining or properly measuring the learning that takes place from simulation play. Among the critics are Anderson and Lawton (1997a) Gentry et al. (1998), and Thavikulwat et al. (1998)

According to Anderson and Lawton (1997a):

> There is relatively little *hard evidence* (emphasis theirs) that simulations produce learning or that they are superior to other methodologies. Much of the reason for the inability to make supportable claims about the efficacy of simulations can be traced to the selection of dependent variables and the lack of rigor with which investigations have been conducted…. Virtually all research designed to measure the outcomes produced by engaging in an activity requires by necessity assumptions concerning the expected outcomes produced by performing that activity. We cannot construct an assessment activity without knowing what it is we expect to measure.

In other words, these critics have maintained that the field has failed to produce objective devises for measuring learning from the simulation, that objective measures must come from measurable objectives conceptualizing expected outcomes, and that the field has not provided agreed upon expected outcomes from simulation play.

There have been attempts to measure simulation learning, but the measuring devices in some studies have been indirect. These devices include course grades (Comer and Nichols, 1996) and course exams, (Raia, 1966 and Wellington and Faria, 1991). There have been measures that are more direct but stem from very general learning objectives such as attaining quantitative skills (Faria and Whiteley, 1990 and Whiteley and Faria, 1989), company self concept development, (Pearce, 1978-90), and goal setting abilities (Wheatley, Horneday and Hunt, 1988). One study in which learning measures emerged from specific learning goals was performed by Wolfe (1976). His focus was on the effects of game participation on learning strategic management and organizational goal setting. His more specific objectives included 'administer a preconceived strategy' and 'create the components of a business policy system.' After reviewing the above studies it appears that in only one study (Wolfe, 1976) were specific objectives used to guide the development of an instrument measuring simulation related learning, and in none were measurement devices developed from specific objectives emerging from the simulation itself. The present research was designed to fill the void. For the test bank central to this research, the items created were developed from specific objectives emerging from the simulation. The long-term result of this effort is intended to be a test bank of usable items, the objectives from which they emerge, and reliability and discrimination statistics. The intention is also to create simulation-learning related scales and validity statistics for each scale.

THE PRESENT PAPER

The present paper describes the first attempt to collect relatively extensive data on a limited set of items in the bank. More specifically two versions of the test were constructed, one of 35 items and the other of 38. In all, fourteen of the objectives were assessed using 66 of the items (as seven of the items were used in both versions). One hundred and sixty-two students from four classes took the test, two classes in the Fall of 98 and two in the Spring of 99. Two classes were from Mankato State University, one of 52 students and one of 38 and two were from the University of Wisconsin-Whitewater, one of 26 students and one of 46. Three instructor/administrators participated, one from Mankato and two from Whitewater.

Table 1 shows when each of the two versions were administered. It shows that version 1 was the pre-test in three of the four classes in this study, while version 2 was the post-test for three of the four classes. It also shows that each of the tests was administered four times.

Table 1: Test Version Administration

Class	Pre	Post
UWWf98	version 1	version 2
UWWs99	version 1	version 1
MSUf98	version 1	version 2
MSUs99	version 2	version 2

Classes differed with respect to size, team size, and percent of course weight devoted to game performance and post-test performance. Table 2 shows the information for these design variables for each of the four classes. In each of the classes, students played MICROMATIC (Scott et al., 1992) and in each of the classes, students took two forms of the learning test. At UWW, students took one version of the test (a pre-test) before simulation play began. At MSU, the pre-test followed practice rounds consisting of students playing the game as individuals against computer developed, virtual competitors. In all classes, a second (post-test) was administered just about the time the game ended. Learning was defined as the difference between the percent correct score from the post-test minus the percent correct score from the pre-test. Percent correct score was determined for each student for each test by dividing the points awarded by points possible (i.e., raw scored divided by 52 in version 1 and 50 version 2). At UWW for all classes, game performance was weighted 40%

net income, 30% return on sales and 30% return on assets. At MSU, performance was 20% sales, 10% net income, 5% return on sales, 20% earnings per share, 25% return on assets, 10% return on equity, and 10% stock price.

Table 2: Game Design Variables by Class

School	Semester	Students	Team Size	Rounds	% Grade on Performance	% Grade on Post Test
UWW Fall	98	26	2-4	11	12.5	12.5
UWW Spring	99	46	2-4	12	15	10
MSU Fall	98	52	2-4	14	5	9
MSU Spring	99	38	3-4	14	5	9

We collected four kinds of data. The first concerned the validity of the simulation. We tested the validity of the simulation by using test scores from the pre and post-tests. The simulation would be judged valid if scores on the post-test for a given class were significantly higher, by t-test, than pre-test scores of that class. This validity test would be, also, an indirect assessment of the validity of the tests in this study. Previously (Washbush and Gosen, under review), we have found that with similar test items, post-test scores were consistently and significantly higher than pre-test scores (by an average of about 10%). If differences between pre and post-test scores in this study were not significant, then perhaps these items are less valid than those used in the above mentioned previous research. The other three kinds of data emerged directly from the test and its items. These were reliability coefficients, difficulty coefficients, and correlations of discrimination. Reliability coefficients of versions of the test were calculated with the Kudor Richardson 20 formula used in the test scoring software at the University of Wisconsin-Whitewater. Difficulty coefficients were the percent of those attaining the correct answer. Discrimination correlations were the correlation between a given item score with that of the total score of that version of the test. The purpose for obtaining discrimination and difficulty scores is to assess item viability. Items that were too easy or too hard or fail to correlate with other items were candidates for elimination from future versions of the bank.

RESULTS

Validity. T-tests reveal that in three of the four classes, post-test scores were significantly higher than pre-test scores. Table 3 shows those data.

Table 3: t-test Results

University and Semester	Pre-test % scores	Post-test % scores	t and Significance	
UWW-fall 98	45	56	2.34	.03
UWW-spr.99	54	66	5.90	<.001
MSU - fall 98	47	59	2.56	.02
MSU - spr. 99	49	52	1.18	ns

Reliability. Kudor Richardson coefficients were .663 for version 1 (n = 170) and .668 for version 2 (n = 154).

Item Analyses. Table 4 contains a list of the learning goals measured in this study, the number of items used to measure that particular goal, the number of items for that goal with low discrimination correlation coefficients (r < .25), the number of items measuring that goal with high difficulty coefficients (% wrong >60), and the number of items with low difficulty coefficients (% wrong <10). Table 4 results show that almost two thirds of the items did not vary to a great degree with the total test scores. This suggests that the test may be measuring more than one construct. These results also show that almost half (32 of 66) of the items proved difficult for theses students, in that more than 60% of the students taking this test answered these items incorrectly. Only ten of the 66 item proved so easy that 90% or more of the students got them right.

Table 4: Item Analysis

Learning Objective	# of items in both versions of tests measuring	# of items with low discrimination statistics	# of items with high difficulty statistics	# of items with low difficulty statistics
Attend to detail, such as ordering raw materiel, accounting for employee turnover, so that poor performance does not result	5	1	0	0
Understand consequences of specific decisions, such as ordering materials at a discount	31	10	8	3
Apply models involving cash flow, growth, profits, assets and dividend payments.	9	3	3	1
Apply models involving cash flow, growth, profits, assets and dividend payments.	2	0	0	0
Effectively interpret game's financial statements	17	8	4	3
Distinguish between market structure, rivalry and other economic forces that affect the firm.	2	2	2	0
Understand and distinguish between market characteristics, competitor behavior, and other economic forces that influence decisions	15	7	5	2
Derive and implement effective decisions which address situations, problems, and opportunities that arise during the simulation	4	2	2	0
Appropriately use financial statements in decision making.	3	3	3	0
Create and implement internally consistent strategies.	2	2	0	1
Use pro forma or what if analyses to assess the probable impacts of decisions.	2	2	2	0
Use financial statements to enhance decision making.	2	2	2	0
TOTALS	66	43	32	10

DISCUSSION

It's difficult to draw conclusions with so little data. Reliabilities are neither good nor awful, given N's of 160 and about 35 items per version of the test. It is hard to explain why learning scores in one of the classes were so low. This result and the variability of results across instructors raise questions about how instructors prepare students for and administer these tests. Do instructors encourage students to study for the post-test? Does the timing of the pre-test coincide with students preparing for the initial rounds of the game (or as in the case of Mankato State coincide with practice rounds)? Are there statements by the instructor that encouraging students to take the tests more or less seriously.

This project has only just begun. The following are necessary for this project to be successful.

1. Studies must be undertaken on the validity of the tests which means developing criterion measures.
2. More items need to be used, partially so there are more items for practitioners to choose from and partially so that more of the learning goals can be measured with a sufficient number of items.
3. For the same reasons as in #2, there should be longer tests.
4. Factor analyses should be undertaken so help determine what is being measured by these tests.
5. If statistics suggest that the tests measure more than one learning dimension, then validity studies should be undertaken on each dimension.
6. More studies are needed at more universities to substantiate generalizability.

1. The marketing analysis question was also a short essay question.

REFERENCES

Anderson, P.H. & Lawton, L. (1997a). Demonstrating the learning effectiveness of simulations: where we are and where we need to go. Developments in Business Simulation & Experiential Exercises, 24, 68-73.

Anderson, P.H. & Lawton, L. (1997b). Designing instruments for assessing the effectiveness of simulations. Developments in Business Simulation & Experiential Exercises, 24, 300.

Comer, L.B. & Nichols, J.A.F. (1996). Simulation as an aid to learning: How does participation influence the process? Developments in Business Simulation & Experiential Exercises, 23, 8.

Faria, A.J. & Whiteley, T.R. (1990). An empirical evaluation of the pedagogical value of playing a simulation game in a principles of marketing course. Developments in Business Simulation & Experiential Exercises, 17, 53-57.

Gentry, J.W., Commuri, S.F., Burns, A.C. & Dickenson, J.R. (1998). The second component to experiential learning: A look back at how ABSEL has handled the conceptual and operational definitions of learning. Developments in Business Simulation & Experiential Exercises, 25, 62-68.

Gosen, J., Washbush, J., Patz A., Scott T.W., Wolfe, J., &Cotter, D. (1999). A test bank for measuring total enterprise simulation learning. Developments in Business Simulation & Experiential Exercises, 25, 82-92.

Pearce, J.A.II (1978-1979). Developing business policy skills: A report on alternatives. Journal of Educational Technology Systems, 7, 361-371.

Raia, A. P. (1966). A study of the educational value of management games. Journal of Business, 39. 339-352.

Scott, T.W., Strickland, A.J., Hofmeister, D.L. & Thompson, M.D. (1992). Micromatic: A Management Simulation. Boston: Houghton Mifflin.

Thavikulwat, P., Anderson, P.H., Cannon, H., & Malik, D. (1998). Games as instruments of assessment: A framework for evaluation. Developments in Business Simulation & Experiential Exercises, 25, 31-37.

.Washbush, J. & Gosen, J. (under editorial review). Learning in total enterprise simulations. Simulation & Gaming: An International Journal.

Wellington, W.J. & Faria, A.J. (1991). An investigation of the relationship between simulation play, performance level and recency of play on exam scores. Developments in Business Simulation & Experiential Exercises, 18, 111-115.

Wheately, W.J., Hornaday, R.W. & Hunt, T.G. (1988). Developing strategic management goal setting skills. Simulation & Games, 19, 173-185.

Whiteley, T.R. & Faria, A.J. (1989). A study of the relationship between student final exam performance and simulation game participation. Simulation & Games, 20, 44-64.

Wolfe, J. (1976). Correlates and measures of the external validity of computer-based business policy decision making environments. Developments in Business Simulation & Experiential Exercises, 7, 411-438.