Developments in Business Simulation and Experiential Learning, Volume 29, 2002

THE VALIDITY INVESTIGATION OF A TEST ASSESSING TOTAL ENTERPRISE SIMULATION LEARNING

Gosen, Jerry & Washbush, John
University of Wisconsin-Whitewater
gosenpuj@uww.edu

ABSTRACT

This paper is the fourth in a series dealing with the construction of a test bank of items designed to assess the degree to which learning takes place from playing a total enterprise simulation. It provides data as to whether the test central to this research is valid. In this study, relationships between results on this study's learning test and two criterion variables – self report of learning and forecasting accuracy - were examined. The results revealed no relationship between learning scores on the test and forecasting accuracy, and they showed a correlation of .33 between learning scores and the degree to which students said they learned financial analysis by playing the game.

INTRODUCTION

This study is part of a long-term project to develop instrumentation to assess learning from a total enterprise simulation. The 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 were Anderson and Lawton (1997), Gentry et al. (1998), and Thavikulwat et al. (1998).

In earlier phases of the project, we developed a test bank of 122 multiple-choice and short essay items (Gosen et al., 1999), gathered some initial data on two instruments from the bank (Gosen, Washbush & Scott, 2000), and initiated validity studies on a third version of the test (Gosen & Washbush, 2001). The purpose of the present study was to continue the validity investigation.

Validity. The validity of a test score, according to McDonald (1999), is the extent to which it measures an attribute of the respondents that the test is employed to measure in the population for which the test is used. Alternately, a test is valid if it measures what it purports to measure. Thavikulwat, et al. (1998) have proposed these standards for evidence of validity for assessment instruments: 1) show evidence of reliability, 2) be able to discriminate between individuals with different levels or types of learning, 3) show convergence with other instruments attempting to measure the same constructs, and 4) yield normative scores for different populations.

Given Thavikulwat et al's categories, for this instrument in previous studies, we've shown evidence of reliability (Gosen et al., 1999; Gosen, Washbush & Scott, 2000; Gosen & Washbush, 2001), and we've argued elsewhere (Washbush & Gosen, 2001) that it would be easy to attain normative scores for different populations. The present study focuses on convergent validity. Convergent validity according to McDonald (1999) is when scores on a test are highly correlated with scores on other measures (often called criterion measures) reflective of the same construct.

This investigation attempted to focus on two such criterion measures. The first was forecasting accuracy. This variable has been proposed by Teach (1989, pg. 103) as….(the indicator)…of proficiency with which managers (and student simulation participants) execute a critical management process which is highly associated with a firm's success. He argues further (Teach, 1990, pg.21) that forecasting is a learned skill and that one would expect students to get better with practice. Anderson and Lawton (1988, pg. 242) contend that forecasting accuracy reflects a team's ability to translate its decisions into simulation outputs.

The second criterion was self-reported learning. This is a subjective measure, and its use has been criticized (Gentry et al, 1998), but it makes sense that how much one learns ought to be consistent with how much one thinks she learns.

Background. For a complete description of relevant previous research, see our initial validity study (Gosen & Washbush, 2001) .

The previous validity study. In the summer of 2000 we (Gosen & Washbush, 2001) undertook an initial convergent validity study on the third version of the learning instrument with 23 students. It used two general criterion variables. The first was the accuracy of forecasting units sold. The second was the students' report of what they learned by playing the game. Concretely, this self-report measure consisted of whether a student mentioned a particular type of learning in answering the question, "What did you learn by playing the simulation?" The results regarding forecasting were that learning scores correlated negatively and significantly (p less than .05) with forecasting accuracy. In other words, those who learned the most had the *largest* discrepancy between their sales forecast and actual sales. The results regarding self-report of learning were more positive. They show that those who reported that they learned the complexity of the game had almost significantly higher learning scores than those who did not, and those who reported that they learned financial analysis had almost significantly higher learning scores than those who did not.

The present study. The present study was similar to the 2001 study in that its purpose was to determine the convergent validity of the learning instrument central to this research effort and in that both forecasting accuracy and self-report of learning were used as criterion variables. It was different in that market share and profits after taxes were forecasted in the present study whereas sales in units were forecasted in the previous effort. Also whereas the self-report of learning data came from answers to an open-ended question last year, these data were measured with Likert-type responses this year.

METHOD

Subjects and Procedure. Twenty-seven students taking the capstone policy course at the University of Wisconsin-Whitewater during the Spring of 2001 participated. They played between 14 and 16 quarters of MICROMATIC (Scott et al., 1992), preceded by a practice round. Students played the first six quarters in teams of three, and then each individual entered a new industry, using their team's past performance as the company's history. There were three new industries for the individual players with separate growth curves, but each of these curves was a continuation of the team's original growth curve.

Prior to the practice round,, students were administered version 4 of the learning test central to this study as a pre-test. After the game ended, students completed version 5 of the test as a post-test and also responded to the self-report-of-learning questionnaire. With each game decision, they were required to forecast company market share in each of the game's three markets and profits after taxes. Game performance was worth 15% of the student's course grade, the learning instrument post-test score was worth 10%, forecasting accuracy was worth 5%.

Variable Measurement. Forecasting Accuracy was the average difference between predicted and actual market share in percentages for each of three market areas for each quarter and the average difference between predicted profits after taxes and the actual profits after taxes for each quarter.

Learning for each participant was defined as post-test percent score minus pre-test percent score.

We used a common scoring key to ensure uniformity of measurement for the learning instrument.

Performance in the simulation was measured at the end of play using the game's scoring procedure and was based on net income (40%), return on assets (30%), and return on sales (30%).

Self-Report of Learning was measured with a ten item, Likert-type questionnaire. Each item reflected a type of learning. The items were modifications of answers to the open-ended question, "What did you learn by playing the simulation?", asked of students in our 2001 validity study (Gosen & Washbush, 2001). Concretely, students were asked the degree to which they learned each of the following following by playing the simulation:

that businesses are very complex
cause and effect between decisions and results
financial analysis
balance between sales, capacity, production and workers
forecasting accurately
how to compete
strategic planning and analysis
the need to and how to correct mistakes
how to come back from poor performance
stability in managing a company.

Exhibit 1: Correlations between, learning, performance and forecasting accuracy

	Learning Score	Post Test	Cumulat-ive Profits	Last 4 Forcsts (Market Share)	Last 4 Forcsts (Profits)	All Forcasts (Market Share)	All Forcasts (Profits)
Learning Score		.20	.12	-.26	.00	-.26	.14
Post Test			.02	.07	.04	-.16	-.13
Cum Profits				.24	**.41**	.05	**.38**
Last 4 Forcsts (Mk Sh)					**.49**	**.68**	.30
Last 4 Forcsts (Profits)						.32	**.84**
All Forcasts (Mk Sh)							.28
All Forcsts (Profits)							

Correlations significant at the .05 level or below in bold

RESULTS

Exhibit1 contains the results that pertain to the use of forecasting accuracy as a criterion variable. Correlations between learning scores and the accuracy of forecasting profits were very nearly zero, suggesting no relationship between learning and the accuracy of forecasting profits. Correlations between learning scores and the accuracy of forecasting market share were negative and insignificant, suggesting no relationship between learning and the accuracy of forecasting market share. On the other hand, the accuracy of forecasting profits correlated significantly and positively ($r = .38$; $p = .05$) with the actual profits (the dominant component of the performance index of the simulation in this study) for all forecasts, and even higher ($r = .41$; p less than .05) when only the last four forecasts were considered. Correlations between the accuracy of forecasting market share and profit performance were near zero for all forecasts and .24 (p - NS) for the last four forecasts. Finally learning and performance correlated near zero.

Exhibit 2 contains the results relevant to the use of the self-report of learning as a criterion. This exhibit contains correlations between learning scores and the degree to which students said they learned each of the ten categories of learning listed at the end of the method section of this paper. This exhibit shows one correlation of .33 (p-NS) between those who reported the learned financial analysis and learning and nine correlations of very near zero.

DISCUSSION

To a reasonable degree, these results are consistent with those of our first validity study undertaken last year (Gosen & Washbush, 2001). In both studies, there was a nearly significant tendency

Exhibit 2:
Correlations with Self -Report of Learning

Learning Type	Correlation with learning score
Strategic planning and analysis	-.11
That businesses are very complex	.04
Financial analysis	.33
Cause and effect between decisions and results	-.07
Balance between sales, capacity, production and workers	-.07
Stability in managing the company	.25
The need to and how to correct mistakes	.10
How to compete	.14
How to come back from poor performance	.23
Forecasting accuracy	. 22

Correlations significant at the .05 level or below in bold

(p about .10 for a t-test in last year's study and for a correlation in this year's) for those who reported they learned financial analysis by playing the game to have learned more as measured by this study's learning instrument. In both, there was the lack of a positive association between forecasting accuracy and learning in the game. In both studies, there was a positive relationship between forecasting accuracy and game performance. These performance-forecasting accuracy results are consistent with results found by Teach (1989), who also found positive associations between forecasting accuracy and game profits.

Considering forecasting accuracy as a validity criterion, the results of this study do not support claims for instrument validity. Over two studies, the relationship between forecasting accuracy and learning has been either near zero or negative. Although not established as a valid criterion for the construct learning, there are claims (Teach, 1990) that forecasting improves with practice and varies as participants learn the game. That learning scores on a test do not vary positively with the accuracy of forecasts suggests that either the test, the forecasting variable, or both are not valid representations of the learning construct we have been attempting to measure in this series of studies.

While not quite significant, the results showed that those who learned from the game reported that they learned financial analysis, and the fact that the same result has occurred over two samples provides reason for increased confidence in this not-quite-significant result. Since the game involves analyzing financial statements and since the learning instrument involves assessing financial analytical ability (Gosen et al., 1999), the correlation between self-report of learning financial analysis and learning provides weak evidence that the learning instrument in this study is valid.

Then, this study provides very weak evidence, at best, for the validity of the instrument focal to the present stream of research. One could rationalize that a small sample size and the lack of an established objective criterion measures are partly responsible. In an ideal research world, a course totally devoted to the simulation with the grade based to a very great degree on learning the principles inherent in the simulation might yield consistently positive associations between learning scores and some criterion. But neither that criterion nor the research situation seems to exist in our reality.

In addition, the construct we are trying to measure is at least relatively obscure and complicated. Instruments such as the one focal to this study ought to be available. And they are available. In fact there are at least three, the one from the present study, one by Wolfe and Guth (1975) and one by Feinstein (2001). However, the present instrument has gone through sufficient iterations, has face validity, and shows extremely weak (if any) evidence of convergent validity. The validity of Feinstein's learning instrument (Feinstein, 2001) was based on judgments from hospitality and education graduate students. As written, there is no evidence of any validity studies performed on the instrument from Wolfe & Guth's (1975) research. This series of studies emerged from criticism of the simulation field for not being able to measure the learning that emerges from playing the simulation. If this ability depends on the availability of valid instruments, then the criticism still holds.

On the other hand, to conclude that the instrument used in the present study or the instruments from Wolfe & Guth (1975) and Feinstein (2001) are not valid is premature. There is no evidence at present that these instruments are valid. That does not mean that such evidence is impossible to develop.

REFERENCES

Anderson & Lawton (1997a) "Demonstrating the Learning Effectiveness of Simulations: Where We Are and Where We Need to Go." *Developments in Business Simulation & Experiential Exercises*, Volume Twenty Four, 1997, 68-73.

Feinstein (2001) "An Assessment of the Effectiveness of the Simulation as an Instructional System for Food Service." *Journal of Hospitality and Tourism Research*, Volume Twenty Five, 2001, 421-443.

Gentry, Commuri, Burns & Dickenson (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*, Volume Twenty Five, 1998, 62-68.

Gosen, Washbush, Patz, Scott, Wolfe & Cotter (1999) "A Test Bank for Measuring Total Enterprise Simulation Learning." *Developments in Business Simulation & Experiential Exercises*, Volume Twenty Five, 1999, 82-92.

Gosen, Washbush & Scott (2000) "Initial Data on a Test Bank Assessing Total Enterprise Simulation Learning." *Developments in Business Simulation & Experiential Exercises*, Volume Twenty Six, 2000, 166-171.

Gosen & Washbush (2001) "An Initial Validity Investigation of a Test Investigating Total Enterprise Simulation Learning." *Developments in Business Simulation and Experiential Learning*, Volume Twenty Eight, 2001, 92-95.

McDonald (1999) *Test Theory: A Unified Treatment*. London: Lawrence Erlbaum Associates.

Scott, Strickland, Hofmeister & Thompson (1992) *Micromatic: A Management Simulation*. Boston: Houghton Mifflin.

Teach (1989) "Using Forecasting Accuracy as a Measure of Success in Business Simulations." *Developments in Business Simulation & Experiential Exercises*, Volume Sixteen, 1989, 103-107.

Teach (1990) "Profits: The False Profit of Business Gaming." *Simulation & Gaming: An International Journal*, Volume Twenty One, 1990, 12-26.

Thavikulwat, Anderson, Cannon & Malik (1998) "Games as Instruments of Assessment: A Framework for Evaluation." *Developments in Business Simulation & Experiential Exercises*, Volume Twenty Five, 1998, 31-37.

Washbush & Gosen (2001) "Learning in Total Enterprise Simulations." *Simulation & Gaming: An International Journal*, Volume Thirty Two, 3.