

# Developments in Business Simulation and Experiential Learning, Volume 25, 1998

## GAMES AS INSTRUMENTS OF ASSESSMENT: A FRAMEWORK FOR EVALUATION

Phil Anderson, University of St. Thomas  
Hugh M. Cannon, Wayne State University  
Dolly Malik, SUNY at Geneseo  
Precha Thavikulwat, Towson University

### ABSTRACT

The ABSEL Committee on Assessment was organized to investigate the possibility of establishing registration procedures for the use of simulation games as instruments of student assessment. This paper discusses the issues involved in this initiative, focusing particularly on the problem of validation. It addresses the importance of following rigorous psychometric procedures, and suggests some specific directions for improving future validation work.

### INTRODUCTION

Among the first to note that games might be useful in assessment were Keys and Wolfe (1990), who wrote that "management games will play a more significant role in management development and assessment efforts in business schools as part of the move toward competency-based outcome measurement" (p. 324). Since then, the International Association for Management Education has incorporated assessment into its accreditation standards (AACSB, 1994) and produced a videoconference on the subject (AACSB, 1997). The role that management games might play in assessment, however, was not recognized in either product. Thus, although Keys and Wolfe were correct about the move toward competency-based outcome assessment, their prediction of a significant role for management games has not come to pass.

The Association for Business Simulation and Experiential Learning (ABSEL) itself paid little heed to assessment efforts in business education until 1995, when it organized a Committee on Assessment (Thavikulwat, 1995). The Committee drafted standards and registration procedures for

assessment instruments (Cannon, et al., 1996), presented them at ABSEL's 23rd Annual Meeting in Orlando, Florida, and published the final version in *ABSEL News & Views* ("Standards and Registration Procedure," 1996).<sup>1</sup>

The following year, the Committee accepted three submissions for registration. In accordance with the established procedures, the submissions were accompanied by papers subsequently published in *Developments in Business Simulation & Experiential Learning* (Butler, 1996; Fritzsche, 1996; Thorelli, 1996). Nevertheless, because of concerns about validity expressed at the ABSEL conference session in which two of the three submissions were presented, the Committee decided to withhold registration of all submissions until sufficient documented evidence of validity became available.

The Committee decided that evidence of validity for an instrument would include supporting showing all of the following:

1. Reliability in the measurements obtained by using the instrument;
2. Discrimination by the instrument between individuals within a population with different types and/or degrees of learning;
3. Convergence between the instrument's measure and other reasonable measures of learning;
4. Normative scores for different relevant populations.

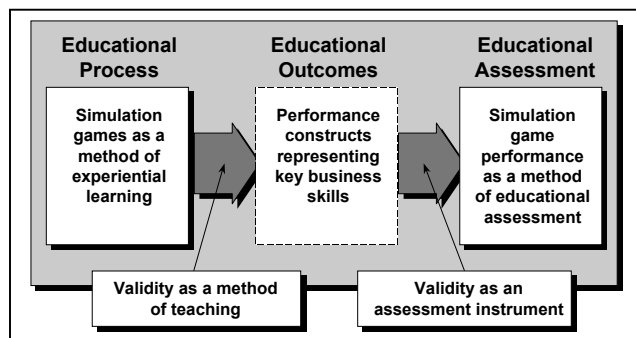
---

<sup>1</sup> The standards and registration procedure cover experiential exercises as well as games.

## Developments in Business Simulation and Experiential Learning, Volume 25, 1998

The Committee was concerned with *instrumental validity*, that is, the extent to which a gaming instrument measures learning, and not with teaching effectiveness, as illustrated in Exhibit 1. Although the Committee's use of the term *validity* departs from common usage in the gaming literature, wherein validity is synonymous with teaching effectiveness (Burns, Gentry, & Wolfe, 1990; Keys & Wolfe, 1990; Stanislaw, 1986), it is consistent with psychometric usage.

**EXHIBIT 1:  
SIMULATIONS AS ASSESSMENT VS.  
EDUCATIONAL TOOLS**



Note that in our discussion we will use the term *simulation game* in the broadest sense. It represents not only computer simulations, but any experiential exercise that is designed to immerse students as live actors in an actual experience that simulates some aspect of business.

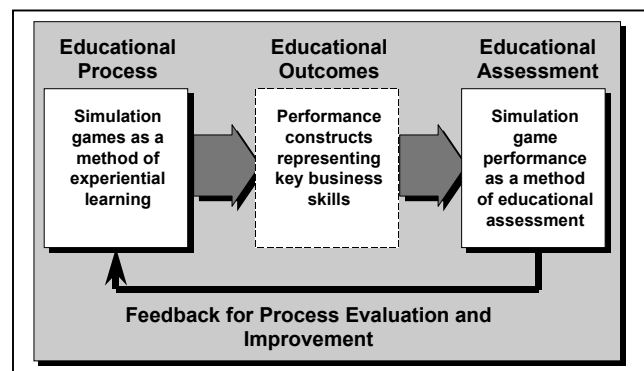
### THE PROBLEM OF VALIDATION

In order to meet minimum standards of registration by ABSEL, an instrument must demonstrate psychometric reliability and validity. The most fundamental requirement is construct validity, confirming the “meaning” of the measurement tool (Kerlinger, 1973). According to Cronbach (1970), the first step in the construct validation process is the identification of the constructs that the instrument is measuring. This suggests that the developers of the instrument cannot validate the instrument, *per se*; rather, they must validate the constructs that the instrument is hypothesized to be measure. With respect to simulations, there-

fore, the developer must establish the validity of the learning objectives that the instrument is designed to measure.

Exhibit 2 puts this problem in perspective. The purpose of the ABSEL Committee on Assessment is to facilitate the use of simulation games as instruments for measuring student learning of key business skills. But, given ABSEL’s commitment to the use of simulations as teaching tools, it is probable that the learning the committee is trying to assess will result, in part, at least, from the use of simulation games. The circularity of the reasoning is obvious from the exhibit. Without some external measure of construct validity, “performance constructs representing key business skills” might really come to mean “performance constructs representing the ability to play business simulation games.” Hence, we conclude, “Games are effective means of teaching, because students who perform well in the games demonstrate high-level learning, as indicated by the fact that they performed well in games.”

**EXHIBIT 2:  
THE CIRCULARITY OF USING GAMES  
FOR TEACHING AND ASSESSMENT**



This problem is by no means unique to the use of simulation results in assessment. In fact, using instruments that are highly related to the teaching process is a long-standing tradition in education. This includes everything from the use of essay exams, performance on which mimics the process of writing and discussion that was used in the class, to the use of analytical problems and exercises, where the ability to perform the analyses is the relevant

criterion of student learning.

The question, of course, is the real educational outcome we are trying to achieve. If, as is the case for business simulations, we are trying to teach a key set of business skills, we must conceptually identify what these are, independent of the teaching process. Once we identify they key business skills we are trying teach, we can begin evaluating potential assessment measures, independent of how the skills are taught. If simulation games can be demonstrated to measure these outcome constructs, then it becomes irrelevant whether students learned them from business simulations.

**EXHIBIT 3:  
MULTITRAIT-MULTIMETHOD MATRIX**

		Measures of Educational Performance					
		Method 1			Method 2		
		T <sub>1,1</sub>	T <sub>1,2</sub>	T <sub>1,3</sub>	T <sub>3,1</sub>	T <sub>3,2</sub>	T <sub>3,3</sub>
Method 1	T <sub>1,1</sub>	C <sub>1,1,1</sub>					
	T <sub>1,2</sub>	C <sub>1,2,1</sub>	C <sub>1,2,2</sub>				
	T <sub>1,3</sub>	C <sub>1,3,1</sub>	C <sub>1,3,2</sub>	C <sub>1,3,3</sub>			
Method 2	T <sub>2,1</sub>	C <sub>2,1,1</sub>	C <sub>2,1,2</sub>	C <sub>1,1,3</sub>	C <sub>3,1,1</sub>		
	T <sub>2,2</sub>	C <sub>2,2,1</sub>	C <sub>2,2,2</sub>	C <sub>2,2,3</sub>	C <sub>3,2,1</sub>	C <sub>3,2,2</sub>	
	T <sub>2,3</sub>	C <sub>2,3,1</sub>	C <sub>2,3,2</sub>	C <sub>2,3,3</sub>	C <sub>3,3,1</sub>	C <sub>3,3,2</sub>	C <sub>3,3,3</sub>

The most commonly accepted tool for construct validation is Campbell and Fiske’s (1959) *multitrait-multimethod matrix* approach (Exhibit 3) for investigating the discriminant and convergent validities of the instrument (Cronbach & Meehl, 1955; Cronbach, 1971).. Note that the approach requires at least two different assessment instruments, each of which purports to measure the same educational output constructs. These are the *methods* (*Method 1* and *Method 2*) in the multimethod approach. The approach also assumes that learning involves more than a single construct, or dimension. The various dimensions of learning are the *traits* (T<sub>1,1</sub>, T<sub>1,2</sub> and so forth) in the multitrait approach. The numbers in each cell of the matrix represent a correlation of trait measurements. The diagonals (C<sub>1,1,1</sub>, C<sub>1,2,2</sub> through C<sub>3,3,3</sub>) represent correlations of two measurements using the same method to measure the same trait.

The off-diagonals (C<sub>1,1,2</sub> through C<sub>3,3,2</sub>) represent the correlations of measurements for two different traits. Combined, the *multitrait-multimethod matrix* addresses the first three requirements for a valid assessment instrument that were listed earlier:

1. **Reliability** is the degree to which simulation results correlate from one measurement to the next (Schnieder and Schmitt 1992).. For instance, if students were evaluated in two separate games or sets of trials, would their level of performance be similar, relative to that of other students? If so, we can say the simulation is reliable in the assessments it enables us to make. In the multitrait-multimethod matrix, reliability is indicated by the diagonal values in the diagonal line submatrices (i.e. C<sub>1,1,1</sub>, C<sub>1,2,2</sub>, C<sub>3,1,1</sub>, C<sub>3,2,2</sub>, etc.).

2. **Discrimination** actually comes in three varieties. First, the simulation results must discriminate among students with different levels of performance. In the absence of this discrimination, the *multitrait-multimethod matrix* would fail to demonstrate any correlations, since there would be no performance variance for the correlation to explain. Conversely, the presence of a meaningful pattern of correlations implies that this discrimination is present. Second, if simulation performance involves multiple dimensions (traits), the off-diagonal correlations in diagonal submatrices (i.e. the values of C<sub>1,1,2</sub> through C<sub>1,3,2</sub> and C<sub>3,1,1</sub>, through C<sub>3,3,2</sub>) should be relatively low, since they involve correlations between different types of performance as measured by similar methods of measurement. Third, the off-diagonal correlations in the off-diagonal submatrices (in this case, C<sub>2,1,2</sub> through C<sub>2,3,2</sub>) should also be relatively low. These involve correlations between different types of performance as measured by different methods of measurement.

3. **Convergence** represents the degree to which different measures of the same educational outcome correlate (converge) with each other

## Developments in Business Simulation and Experiential Learning, Volume 25, 1998

(reflected in relatively high values for  $C_{2,1,1}$ ,  $C_{2,2,2}$ , and  $C_{2,3,3}$ ). The greater the divergence in the type of measure, the greater confidence we have that the construct is being validly measured, since there is minimal chance of method bias (Kerlinger 1973). That is, there is little chance that the measure is simply reflecting the effects of the teaching method, not the educational outcome the method was designed to achieve. For instance, if our desired educational outcome were the ability to perform complex business analyses, success in a simulation that presumably requires complex analysis, an essay exam in which students have to explain the theory and application of the relevant analysis, and an applied exercise where students are required to perform the analysis would be three relatively divergent types of measures. If they converged, they would provide strong evidence of construct validity.

In the *multitrait-multimethod* framework, simulation game performance becomes one *method* of measuring key business skills. The matrix provides a useful set of guidelines for conducting validation research in support of simulation games as assessment instruments, suggesting a practical structure for validation research. Most important, it highlights the importance of beginning with a clear conceptualization of the constructs – the kinds of business skills (*traits*) we are trying to evaluate – rather than beginning with a discussion of student performance. Once this has been accomplished, it is possible to identify alternative methods by which these might be measured. This, in turn, will provide the external validation we require to avoid the circularity portrayed in Exhibit 2.

### PRIOR VALIDATION STUDIES

ABSEL's Assessment initiative has created a renewed interest in validation. But the issue is by no means new. Since the early days of gaming, there has been a call for hard evidence to support the teaching effectiveness of simulations (see, for example, Neuhauser, 1976; Snow, 1976). Numerous studies have attempted to assess what students

learn in a business simulation exercise (Greenlaw and Wyman, 1973; Keys, 1976; Parasuraman, 1981; Wolfe, 1981, 1985, 1987; Teach and Govahi, 1988; Whiteley and Faria, 1989; Burns, Gentry, and Wolfe, 1990; Wolfe, 1990; Gosenpud, 1990; Wellington and Faria, 1991; Anderson and Lawton, 1992a; Hemmasi and Graf, 1992; Gosenpud and Washbush, 1993, 1994; Anderson and Lawton, 1995, 1997; Washbush and Gosenpud, 1995).

The length of the bibliography in Keys' and Wolfe's 1990 review of the state of simulation is impressive. Nevertheless, despite the extensive literature, it remains difficult, if not impossible, to support *objectively* even the most fundamental claims for the efficacy of games as a teaching pedagogy. There is relatively little *hard evidence* that simulations produce learning or that they are superior to other methodologies.

As we pointed out in our discussion of Exhibit 1, these studies have tended to look at the validity of simulations as methods of teaching, not as assessment instruments. But the underlying issues are the same. In the end, any discussion of validity must begin with an identification of the educational outcomes we are hoping to achieve and assess. Many studies have glossed over this issue, opting for intuitively derived measures of student performance.

Those studies that have attempted to take a more rigorous approach to identify performance constructs have tended to focus on Bloom's Taxonomy of Learning (Bloom *et al*, 1956). In the late 1940s and 50s, Benjamin Bloom headed a project seeking to develop a systematic taxonomy of educational outcomes. The result was a six-level hierarchy, reflecting progressively higher levels of cognitive learning (Exhibit 4).

Early simulation research focused on students' *perceptions* of what they learned (e.g.; Schellenberger, *et al*, 1989). More recently, paper and pencil tests have been used to assess lower levels of learning on Bloom's Taxonomy (e.g.; Gosenpud

## Developments in Business Simulation and Experiential Learning, Volume 25, 1998

rud and Washbush, 1994). While instructors have used a variety of methods in attempting to determine the level of mastery a student has achieved from exposure to the exercise, financial performance has remained a key measurement tool. A survey by Anderson and Lawton (1992b) found that all respondents, without exception, used financial performance as one of the determinants, and sometimes the sole determinant, of a student's grade for the simulation exercise.

ject's business GPA was found to have a significant relationship with performance on a simulation.

A handful of studies have been conducted to in an effort to determine the relationship between simulation performance and successful performance on-the-job (Norris and Snyder, 1982; Wolfe and Roberts, 1986; Wolfe and Roberts, 1993). They are particularly interesting in this context, since they attempt to address the validity of simulations as assessment instruments rather than as teaching tools. As Wolfe and Roberts' (1993, p. 25) point out, they may serve "as a device for assessing potential managerial talent." The studies are also interesting because of the fact that they use actual on-the-job performance as an external criterion of validity.

While on-the-job performance would seem to be very relevant as an indicator of performance ability, it is less satisfying than Bloom's taxonomy in that it provides little insight into what "performance ability" really is. Is it a multidimensional construct? Is it situationally dependent? How does it relate to the host of individual skills we teach in Schools of Business? To the more general skills taught in other types of courses? To the fundamental thinking skills addressed by Bloom's taxonomy?

This takes us back to the importance of establishing construct validity. Returning to the logic of Exhibit 2, we see why it is important to understand the constructs representing key educational outcomes. Without a clear understanding of what it is we are trying to measure, even the most attractive variable – on-the-job performance, for example – is suspect. How do we know that performance in one situation will have any relationship to performance in another? We can only know by breaking down performance into its relevant components – *traits* in Exhibit 3.

Implicit is the fact that much of our lack of progress in validating simulation games can be traced to the selection of dependent variables. Rigorous

### EXHIBIT 4: COGNITIVE LEARNING OBJECTIVES

Learning Objectives	Description of the Learning	Evidence of Learning
1. Basic knowledge	Student recalls or recognizes information	Answer to direct questions /multiple-choice tests
2. Comprehension	Student changes information into a different symbolic form	Ability to act upon or process information by restating material in his own words
3. Application	Student discovers relationships, generalizations and skills	Application of knowledge to simulated problems
4. Analysis	Student solves problems in light of conscious knowledge of principles and relationships	Identification of critical assumptions, alternatives and constraints in a problem
5. Objective synthesis	Student goes beyond what is known, providing new insights	Solution of a problem that requires original, creative thinking
6. Objective evaluation	Student develops the ability to create standards to judge, to weigh, and to analyze	Logical consistency and attention to detail

At present, there are few objective measures for assessing learning at the higher levels of Bloom's Taxonomy. In the absence of these measures, financial performance has been relied on as a proxy for student learning at all levels of Bloom's Taxonomy (Anderson and Lawton, 1992b). Unfortunately, research by Anderson and Lawton (1992a, 1995, 1997) found the relationship between financial performance and other measures of student mastery to be weak or non-existent. This lack of a relationship exists regardless of whether simulation performance is based on group-managed or individually managed companies. No significant relationship was found between financial measures on a simulation and independent variables which included: the grade received on a case study write-up; the grade received for class participation during the course; the grade received on an assessment of a managerial scenario; overall GPA; a peer group assessment of the subject's strategic management skills; and a self-assessment of managerial skills. Only the sub-

**Developments in Business Simulation and Experiential Learning, Volume 25, 1998**

research design is also important, but even if researchers are assiduously attentive to good experimental design, useful research results will not be achieved if the measure of learning is invalid.

**SUMMARY AND CONCLUSIONS**

Many people believe that business simulation games have enormous potential as assessment instruments for evaluating the skill of job candidates. After all, what more logical candidate for measuring student potential in a real work situation than performance in an exercise designed to simulate a real situation?

Unfortunately, this logic fails us, when we consider the fact that we don't really know what student potential for work success really is. The problem is suggested in Exhibits 1 and 2, where the validity of both the educational approach and the assessment measures are dependent on the educational outcomes -- the key skills needed for success. Without knowing what these are, we have no way of knowing whether the educational approach and assessment measures are valid. Conversely, once we know what the skills are, we can look for different ways of measuring them, using the logic of the *multitrait-multimethod matrix* presented in Exhibit 3 to evaluate convergent and discriminant validity.

In essence what is missing is a theory of simulation game performance. What is it that causes some students to succeed in simulation games and others to be less successful? Similarly, what causes people to be more versus less successful in real life business situations? We have noted that much of the work in this area has drawn on Bloom's taxonomy of educational objectives as a basis for conceptualizing educational outcomes. This assumes that general intellectual skills will help people in specific decision-making situations. This is very different from the view that there are a specific set of skills – forecasting ability, the ability to project cash flow, the ability to use market research, and so forth – that are generalizable across business situations.

Largely missing from these discussions are factors that address the affective dimension of learning, or the way people attend to and value different kinds of business activities (Krathwohl et. al 1964) – the way they are motivated to behave. It may well be that success is more related to issues relating to motivation than intellectual skills. Or it may be that the failure of previous research to show a relationship between intellectual skills and performance is because there is an interaction effect between the cognitive and affective dimensions. That is, either general abilities such as analysis, synthesis and evaluation, or specific skills such as forecasting ability and the ability to project cash flow, may be necessary but not sufficient conditions for success. Rather, they would depend on the recognition that the skills are important and necessary, either in general, or in specific situations.

**EXHIBIT 5:  
A FRAMEWORK FOR DEVELOPING  
THEORIES OF SIMULATION  
PERFORMANCE**

		Cognitive Skills	
		General	Specific
Affective Orientation	Specific	Students learn to apply general problem-solving skills in those situations where they are most needed.	Students learn to apply specific problem-solving skills in those situations where they are most needed.
	General	Students learn the importance of mastering and applying general problem-solving skills to business situations.	Students learn the importance of mastering and applying specific problem-solving skills to business situations.

Exhibit 5 summarizes this perspective. It suggests a framework for developing theories of simulation performance, which combine cognitive and affective objectives. To the extent that these theories are able to explain performance, they can be used to guide the development and validation of simulation-based assessment measures. Again, the key will be twofold: First, theorists must identify the key types of cognitive and affective skills (*traits*) thought to be essential to business success. Second, they must identify a variety of different,

## Developments in Business Simulation and Experiential Learning, Volume 25, 1998

maximally dissimilar, measures (*methods*) of these skills. These provide the components required for validation studies, as suggested by the multitrait-multimethod matrix discussed in conjunction with Exhibit 3. By following this approach, we anticipate that validation research regarding the use of simulation games as assessment instruments will begin to make much greater progress.

### REFERENCES

- AACSB - International Association for Management Education. (1994). *Achieving Quality and Continuous Improvement Through Self-Evaluation and Peer Review: Standards for Accreditation, Business Administration and Accounting*. St. Louis, MO: Author.
- AACSB - International Association for Management Education (Producer). (1997, April 24). Program assessment: Strategies and practices for achieving teaching and learning outcomes [Interactive video conference].
- Anderson, P.H., and L. Lawton (1989) "An Evaluation and Application of an Instrument for Measuring Pedagogical Effectiveness." *Developments in Business Simulation and Experiential Exercises*, eds. T. Pray and J. Wingender, 16: 92-96.
- Anderson, P.H. and L. Lawton (1988), "Assessing Student Performance on a Business Simulation Exercise." *Developments in Business Simulations and Experiential Exercises*, 15, 241-244.
- Anderson, P.H. and L. Lawton (1990) "Measuring the Learning Outcomes of Management Training Activities," *Proceedings of the Association of Management*, August.
- Anderson, P.H. and L. Lawton (1992a), "The Relationship Between Financial Performance and other Measures of Learning on a Simulation Exercise", *Simulation & Gaming*, 23 (3), 326-340.
- Anderson, P.H. and L. Lawton (1992b), "A Survey of Methods Used for Evaluating Student Performance on Business Simulations", *Simulation & Gaming*, 23 (4), 490-498.
- Anderson, P.H. and L. Lawton (1995), "The Problem of Determining an "Individualized" Simulation's Validity as an Assessment Tool", *Developments in Business Simulations and Experiential Exercises*, 22, pp.43-48.
- Anderson, P.H. and L. Lawton (1997), "Performance on a TE Simulation: What Does It Represent?", *Developments in Business Simulations and Experiential Exercises*, 24.
- Bloom, B. S., M. D. Englehart, E.D. Furst, W.H. Hill, and D. R. Krathwohl (1959), *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain*, New York: David McKay Company, Inc.
- Burns, A. C., J.W. Gentry, and J. Wolfe (1990), "A Cornucopia of Considerations in Evaluating the Effectiveness of Experiential Pedagogies," in *Guide to Business Gaming and Experiential Learning*, J.W. Gentry (ed.), New York: Nichols/GP Publishing, 253-278.
- Butler, J. K., Jr. (1996). Assessing negotiators' proficiency with a negotiation role play. *Developments in Business Simulation & Experiential Learning*, 24, 88-89.
- Campbell, D. and D. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin* 61, 81-105.
- Cannon, H. M., J. R. Frazer, J. Groff, D. R. Markovich, R. Stevens, and P. Thavikulwat (1996). "Draft standards and registration procedure for assessment instruments." *Developments in Business Simulation & Experiential Exercises*, 23, 94.

Note that some references have been omitted for the sake of brevity. For a complete manuscript, please contact:

Hugh M. Cannon  
Adcraft/Simons-Michelson Professor  
Department of Marketing  
Wayne State University  
5201 Cass Avenue, Suite 300  
Detroit, MI 48202-3930  
(313) 577-4551(o)  
(313) 577-5486(f)  
hughcannon@aol.com  
<http://cannon.busadm.wayne.edu>