**Developments in Business Simulation and Experiential Learning, Volume 28, 2001**

# A FRAMEWORK FOR EVALUATING SIMULATIONS AS EDUCATIONAL TOOLS

Paul L. Schumann, Minnesota State University, Mankato
paul.schumann@mnsu.edu

Philip H. Anderson, University of St. Thomas
phanderson@stthomas.edu

Timothy W. Scott, Minnesota State University, Mankato
timothy.scott@mnsu.edu

Leigh Lawton, University of St. Thomas
L9Lawton@stthomas.edu

## *ABSTRACT*

*This paper explains how a framework consisting of four levels can be used to evaluate business simulations as educational tools. The four levels consist of measuring (1) the reactions of the students, (2) the amount of learning achieved by the students, (3) the degree to which the behavior of students in other settings reflects what they have learned, and (4) the extent to which results are improved. The paper describes each of the four levels, identifies how Bloom's Taxonomy fits into the framework, summarizes the current literature in terms of the four levels, and offers recommendations for future research.*

## INTRODUCTION

ABSEL conferences have maintained a continuing theme of assessing simulations, either as a stand-alone pedagogy or on a comparative basis (e.g., simulations versus lectures). ABSEL researchers have frequently used Bloom's Taxonomy of learning objectives (Bloom, et al., 1956; Krathwohl, et al., 1964) as a framework for making these assessments. This taxonomy of learning objectives ranges from knowledge to synthesis. Most research among ABSEL members has involved the lower levels of learning in Bloom's Taxonomy. Measuring the higher levels has proven to be a difficult task. A lack of reliable and valid instruments has hindered attempts to measure the learning occurring at the higher levels of Bloom's Taxonomy (Anderson & Lawton, 1995). Anderson et al. (1998) looked at the use of simulations as assessment instruments and highlighted the problem of validation. Gosenpud and Washbush (1993, 1994) and Gosen et al. (1999, 2000) have made repeated attempts to answer the call to develop a reliable and valid instrument, but with limited success.

Thus, while Bloom's Taxonomy provides a useful framework for the purpose of *establishing* learning objectives, the framework has not been as helpful for *assessing* student learning. Furthermore, researchers in related disciplines have emphasized that learning is only one dimension among several that should be assessed (Kirkpatrick, 1998).

In particular, Kirkpatrick (1998) has developed a widely used framework for evaluating the effectiveness of training programs. While Kirkpatrick's Framework is best known in the context of evaluating corporate training programs, we believe that his framework can be adapted to the college setting and would be useful in assessing the effectiveness of simulation exercises. This paper presents Kirkpatrick's Framework (1998) as a guide for assessing simulations. As will be seen, the Kirkpatrick framework is broader in focus than Bloom's Taxonomy and can offer another means for assessing the efficacy of simulations.

In this paper we describe the four levels of evaluation in Kirkpatrick's Framework and discuss their application to the evaluation and assessment of business simulation educational experiences. We also suggest how ABSEL researchers can use Kirkpatrick's Framework to guide future efforts to evaluate simulation exercises. The four levels in Kirkpatrick's Framework are *reaction, learning, behavior,* and *results* (1998).

## LEVEL 1: REACTION

Kirkpatrick's first level of evaluation is *reaction,* which measures how the participants in the learning experience feel about the experience (Kirkpatrick, 1998). In the context of evaluating a college course, reaction measures the students' satisfaction with the course. In the context of evaluating a business simulation experience, reaction measures the simulation participants' satisfaction with the simulation experience.

Kirkpatrick argues that knowledge of participants' reactions to programs or pedagogies is important for at least four reasons (1998: 25). First, the reaction data provide

feedback that helps both to evaluate the learning experience and to provide suggestions for the improvement of future learning experiences. Second, it signals the learners that the instructors care about the learners' satisfaction. Third, it provides concrete quantitative data that can be provided to decision-makers such as managers (in a corporate context) or deans and academic vice-presidents (in a college context). Finally, it provides instructors with quantitative information that can be used to establish standards of performance for future learning experiences (for example, specific numerical goals for the average satisfaction of participants in future simulation experiences).

A researcher can collect reaction data by administering a satisfaction questionnaire to the students participating in a simulation exercise. The reaction questions could focus on the students' satisfaction with the simulation experience, including satisfaction with the experience as a whole, the appropriateness of the simulation for the course, how much and what students feel they learned from the simulation, and so forth. Since the simulation may well color the students' reaction to the instructor and the course as a whole, reaction questions could also be broadened to include items such as the students' satisfaction with the course, instructor, subject matter, facilities (e.g., location, comfort), schedule (e.g., overall length of the course, speed of progress through the material), and learning aids (e.g., appropriateness, effectiveness).

Using reaction measures (e.g., attitude surveys) to assess a program (or course, or simulation exercise) is not without complications. Respondents who are immersed in any learning endeavor may be incapable of adequately evaluating the value of that experience. In addition, the participant's satisfaction with one aspect of the learning experience may influence how he or she evaluates other aspects of the learning experience, even though they may be separate dimensions. For example, dissatisfaction with the instructor or the facilities may have a negative effect on the respondent's rating of the simulation experience.

Researchers can attempt to confront the confounding influence of this "halo" or "horns" effect by using a control group. For example, if a researcher is attempting to assess reaction to a simulation, the research design could include one section of a course that uses a simulation (the treatment group) and another section of the course that does not (the control group). In this case, reaction data can be collected using a posttest-only control group design (Campbell & Stanley, 1963). The posttest-only control group design would allow the researcher to examine the effect of using a simulation on student satisfaction for the course as a whole by comparing the average satisfaction in the treatment group to the average satisfaction in the control group. The question at issue is, of course, does the inclusion of a simulation result in higher levels of student satisfaction for the course, the instructor, and the subject matter?

In the ABSEL literature, participants' reactions to simulations have been researched since ABSEL's early years. Many researchers have identified a variety of learning outcomes where students rate the simulation over other pedagogies. These range from business specific knowledge and skills to decision-making skills to interpersonal skills (Hemmasi & Graf, 1992; Klabbers, 1996; Miles, et al., 1986; Schellenberger, et al., 1989; Teach, 1990; Teach & Govahi, 1988). Wolfe (1981, 1985, 1987, 1990), in his reviews of the effectiveness of simulation exercises, notes the extensiveness of research on participant reaction. If reaction data were enough to establish the pedagogy's legitimacy, no further research would be necessary. But, as both Bloom and Kirkpatrick identify, there are other dimensions of a pedagogy's effectiveness that need to be assessed to get a true measure of its worth.

## LEVEL TWO: LEARNING

Kirkpatrick's second level of evaluation is learning. He defines learning as the degree to which participants in the program change attitudes, improve knowledge, or increase skill as a result of the program (Kirkpatrick, 1998). For example, a corporate training program on cultural diversity might be designed to teach the participants new attitudes about diversity, to increase the participants' knowledge about diverse cultures, and to increase the participants' skills in managing a diverse workforce. Thus, learning can be said to have taken place when attitudes change, knowledge is increased, or skill is improved as a result of the experience (Kirkpatrick, 1998). Assessing learning involves measuring changes — a change in attitudes, or an increase in knowledge, or an increase in skills.

The learning attributable to a particular program could be assessed by questionnaires that measure attitudes and by tests that measure knowledge or skills. The specific aspects of learning to be measured should relate to the learning objectives for the program, in general, and to the specific aims of the pedagogical experience, in particular. Bloom's Taxonomy of learning objectives has been particularly helpful in establishing learning objectives (Bloom, et al., 1956; Krathwohl, et al., 1964).

Researchers also would like to be able to conclude that the learning experience being assessed *caused* the changes in learning that are observed. To achieve these research goals, a careful choice of research design is important. Kirkpatrick (1998) advocates the use of a control group wherever possible. In particular, Kirkpatrick advocates a pretest-posttest control group design (Campbell & Stanley, 1963) because it allows a statistical comparison of the change observed in the experimental group against the change observed in the control group. The pretest-posttest control group design also allows the researcher to determine the similarities of the control and experimental groups before the learning experience begins by comparing the pretest measures of the experimental group to the pretest measures of the control group. Checking for similarities between the two groups at the beginning of the learning experience might be especially critical when the

participants are not randomly assigned to either the experimental or control group.

Thus, a researcher could proceed as follows to test the learning associated with a business simulation exercise. Using Bloom's Taxonomy, develop the learning objectives for the course. Next, develop a test of the attitudes, knowledge, and skills that are included in the learning objectives for the course. Create two otherwise equivalent sections of the course: one uses a business simulation (the experimental group), while the other does not (the control group). If possible, randomly assign students to one of the two groups. Administer the test of attitudes, knowledge, and skills at the beginning of the course (the pretest) to both groups. Administer the test again at the end of the course (the posttest) to both groups. Compare the pretest measures in the experimental group to the pretest measures in the control group to check that the groups are starting from a similar level of attitudes, knowledge, and skills. Then compare the pretest to posttest gains in the experimental group to the pretest to posttest gains in the control group. If the experimental group shows larger gains in their attitudes, knowledge, and skills than the control group, then (since the only difference between the learning experiences of the two groups is that one group experienced the simulation while the other did not) we can be fairly sure that the simulation experience accounts for the difference in learning (Campbell & Stanley, 1963).

If the researchers can be sure that the learners in the two groups start at equal levels of attitudes, knowledge, and skills, then a posttest-only control group design (Campbell & Stanley, 1963) might also be considered. In this case, a researcher could proceed as follows. Develop a test of the attitudes, knowledge, and skills that are in the learning objectives for the course. Create two otherwise equivalent sections of the course: one uses a business simulation (the experimental group), while the other does not (the control group). If possible, randomly assign students to one of the two groups. At the conclusion of the course, administer the test to both groups (the posttest). Then compare the posttest measures in the experimental group to the posttest measures in the control group. If the experimental group shows higher levels of attitudes, knowledge, and skills than the control group, then (if we are certain that the two groups started from the same levels) we can be fairly sure that the simulation experience accounts for the difference in learning (Campbell & Stanley, 1963). Note, however, with the posttest-only control group design, researchers are unable to verify empirically that the groups started from the same levels of attitudes, knowledge, and skills since pretest data are not collected.

If it is impossible to use a control group, then the researcher might consider a one-group pretest-posttest design (Campbell & Stanley, 1963). In this case, a researcher could proceed as follows. Develop a test of the attitudes, knowledge, and skills that are in the learning objectives for the course. Administer the test to the learners at the beginning of the course (the pretest) and at the end of the course (the posttest). Compute the gain in attitudes, knowledge, and skills from the pretest to the posttest. Note, however, that with this research design is it not possible to be certain what caused any observed gain in attitudes, knowledge, or skills. Were the observed gains due to a particular pedagogy being assessed, or were the observed gains due to the other aspects of the learning environment? As Campbell and Stanley (1963) conclude, the one-group pretest-posttest design is weak with respect to both internal and external validity. The importance of a control group design places an extra burden on researchers' attempts to assess learning, but it is an issue that cannot be ignored.

In the existing ABSEL literature, many researchers have reported on participants' perceptions of learning outcomes associated with simulation exercises, which Kirkpatrick would classify as reaction (level one). It has long been noted in the ABSEL literature that reaction data is not sufficient. Parasuraman (1981) called for research based on more rigorous measures of learning rather than perceptions 20 years ago. And as Keys and Wolfe stated in their review of simulation research, "… Many of the claims and counterclaims for the teaching power of business games rest on anecdotal material or inadequate or poorly implemented research designs. These research defects have clouded the business gaming literature and have hampered the creation of a cumulative stream of research" (Keys & Wolfe, 1990: 311). More recently, Anderson and Lawton (1997) point out that few studies have been based on objective evidence, relying instead on subjective reaction assessments.

Therefore, learning resulting from the use of a simulation exercise (level two) is an area in need of additional research. While studies have shown the effectiveness of simulations relative to learning on the lower levels of Bloom's Taxonomy, there is still a paucity of objective evidence relating simulations and Bloom's higher levels of learning (Anderson & Lawton, 1995, 1997). The principal obstacle to assessing these higher levels of learning is the lack of suitable assessment instruments. Gosenpud and Washbush (1993, 1994) and Gosen et al. (1999, 2000) have spent several years attempting to develop an instrument, but have yet to come up with one that meets the standards outlined by Anderson et al. (1998) as necessary to provide reliable and valid results. Until instruments are developed that can directly measure specified learning outcomes, Anderson and Lawton (1997) argue that researchers cannot make legitimate claims of the learning effectiveness of simulation exercises. Successful assessment of simulation learning at these levels is contingent both on the development of reliable and valid instruments and on the use of proper research designs. Development of reliable and valid tests of the knowledge contained in the learning objectives, and the use of proper research designs, should be the focus of future research in this area.

## LEVEL THREE: BEHAVIOR

Kirkpatrick's third level of evaluation is behavior. He describes behavior as the degree to which learners have changed their behavior outside of the learning environment because of their participation in the learning activities (Kirkpatrick, 1998). In other words, behavior refers to whether the learners are actually using what they learned. In a corporate training environment, for example, behavior would refer to whether the trainees are applying on their jobs what they learned in the training program.

While learning for its own sake is valuable, instructors frequently desire that students are able to transfer what they have learned in a classroom (or other learning environment) to other classes, to their jobs, and to their lives. An instructor may conduct lectures, discussions, role-plays, and simulations to teach about, for example, leadership. But does the students' newly learned knowledge and skills about leadership transfer to other settings so that students are better leaders? Thus, assessing the effect of different pedagogies on behavior is important.

Simulations may be particularly effective in enhancing the transfer of learning from the learning environment to other settings. Simulations are based in part on the idea that by creating a learning environment that matches the job as closely as possible, learners will be better able to transfer their learning to their jobs (Goldstein, 1993; McGehee & Thayer, 1961; Miller, et al., 1998). Since business simulations are designed to be a more realistic environment in which to learn, one might therefore hypothesize that the use of simulations will result in improved behaviors.

To test the hypothesis that simulations improve behavior more than other pedagogies, it is necessary to collect data on the degree to which learners are using what they learned in one course in other classes and on their jobs. This might require conducting the assessment of behavior following the completion of the course. That is, if the full length of the course is necessary to inculcate behavior, expecting participants to exhibit that behavior before the course ends may be unrealistic. If this is the case, then a learning objective in a particular course (e.g., leadership skills of the students) would have to be assessed by professors in courses that follow the course in question, or by employers of the students.

When evaluating business simulations, behavior could be defined as the degree to which learners are *exhibiting* the attitudes, knowledge, and skills taught in one class to subsequent classes and non-academic settings (such as on their jobs). In this context, behavior could be measured by surveying individuals who can observe the behavior of the learners in the other settings. For example, a questionnaire could be developed that asks the students' other professors to rate the degree to which the students are using the desired attitudes, knowledge, and skills. Similarly, the students' employers could be surveyed to measure the degree to which the students are using the attitudes, knowledge, and skills on the students' jobs.

To evaluate adequately the effectiveness of business simulations in terms of behavior, we must assess whether the simulation experience has resulted in greater improvements in behaviors related to the course than do alternative pedagogies. This would necessitate measuring the degree to which learners who participated in the simulation experience display *better* desired behaviors than the learners who did not participate in the simulation experience. As with our previous discussion of the assessment of learning (level two evaluation), the goal of determining if simulations result in better behaviors (level three evaluation) suggests that a pretest-posttest control-group design or a posttest-only control group design would be needed.

In our review of the ABSEL literature, we were unable to find studies that examined the hypothesis that simulations will result in better behaviors than other pedagogies. Even if simulations do not result in higher levels of learning than other pedagogies (level two evaluation), it may turn out that simulations result in better behaviors than other pedagogies (level three evaluation) because the similarity between the simulation experience and the real world allows students to better transfer what they learn from the classroom to their lives. Thus, research on the effect of simulations on behavior is a promising area for future research. However, as was noted in the discussion of learning (level two), the ability to develop instruments that are reliable and valid is critical to successful assessment at this level. Whether development of instruments for behavior will be easier or more difficult than for learning is not known, but should be an important focus of researchers of simulation assessment.

## LEVEL FOUR: RESULTS

The fourth level of evaluation in Kirkpatrick's Framework is results. Results refer to the degree to which the output of the participant's workgroup or organization has improved because of the learning program (Kirkpatrick, 1998). In a corporate training environment, results might refer to the effects of the training program on productivity, quality, costs, accidents, sales, turnover, profits, and so forth.

One challenge in applying a level four evaluation of results to settings outside of corporate training is to decide which results are relevant to examine. We can approach this question from different perspectives. From the student's perspective, the relevant results might include grades in other classes, the number and quality of job offers, salary offers, the speed and frequency of promotions, and so forth. Thus, a researcher could try to measure whether simulation participants receive higher grades in other classes, more job offers, higher salaries, and better promotions than nonparticipants. In addition, results can be defined from the employer's perspective. Thus, a researcher could try to measure if the hiring of employees who participated in a simulation while they were in school result in higher productivity, higher work quality, lower costs, fewer

accidents, higher sales, lower turnover, and higher profits for the employing organization than the hiring of employees who did not participate in a simulation while in school.

There is a void in the literature that reports the effect of simulation exercises on results either from the student's or from the employer's perspectives. The closest approximations to this assessment are the longitudinal studies conducted by Norris and Snyder (1982) and by Wolfe and Roberts (1986, 1993). Future longitudinal research in this area could explore the effect of simulations on results from both the student's and employer's perspectives. This direction of research would need to determine the degree to which the simulation experience accounts for different results. Here again, measurement and research design problems arise. That is, were the improved results that may be observed the consequence of the simulation experience, or other educational variables such as other modes of instruction experienced by the student? Thus, once again, a careful choice of measurement instruments and research design is critical in order to establish that participation in simulations lead to improved results.

## CONCLUSIONS

Since each level of evaluation examines the effectiveness of the business simulation from a different perspective, the four levels are complementary — by using all four levels, we get a more complete picture of the effectiveness of the simulation experience. Currently, ABSEL researchers have done an effective job of assessing students' reactions to simulations at level one of Kirkpatrick's Framework. Unfortunately, measurement problems and research design challenges make similar achievement at the other three levels difficult. Perhaps a holistic approach using all four levels in concert can lessen this handicap. That is, assessing the collective results of research across all levels of Kirkpatrick's Framework may provide a means for making a generalized assessment of the effectiveness of simulations; while the assessment of simulation exercises using any one level of the framework may yield inconclusive results due to measurement issues, when the results of research for the four levels are combined into one holistic assessment, researchers may be able to draw inferences and make tentative conclusions. The combined research across the range of the framework could provide a sufficient pool of evidence to render a judgment regarding the efficacy of simulations in educational programs.

## REFERENCES


Anderson, P.H., H.M. Cannon, D. Malik, & P. Thavikulwat (1998). Games as Instruments of Assessment: A Framework for Evaluation. *Developments in Business Simulations and Experiential Exercises*, 25: 31–37.

Anderson, P.H., & L. Lawton (1997). Demonstrating the Learning Effectiveness of Simulations: Where We Are and Where We Need to Go. *Developments in Business Simulations and Experiential Exercises*, 24: 68–73.

Anderson, P.H. & L. Lawton (1995). The Problem of Determining an "Individualized" Simulation's Validity as an Assessment Tool, *Developments in Business Simulations and Experiential Exercises*, 22: 43–48.

Bloom, B.S., M.D. Englehart, E.D. Furst, W.H. Hill, & D.R. Krathwohl (1956). *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, & J. Wolfe (1990). A Cornucopia of Considerations in Evaluating the Effectiveness of Experiential Pedagogies. In J.W. Gentry (ed.), *Guide to Business Gaming and Experiential Learning,* (pp. 253–278). East Brunswick: Nichols/GP Publishing.

Campbell, D. T. & J.C. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Goldstein, I. (1993). *Training in Organizations* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Gosen, J., J. Washbush, & T. Scott (2000). Initial Data on a Test Bank Assessing Total Enterprise Simulation Learning. *Developments in Business Simulations and Experiential Exercises*, 27: 166–171.

Gosen, J., J. Washbush, A. Patz, T. Scott, J. Wolfe, & D. Cotter (1999). A Test Bank for Measuring Total Enterprise Simulation Learning. *Developments in Business Simulations and Experiential Exercises*, 26: 82–85.

Gosenpud, J. & J. Washbush (1994). Simulation Performance and Learning Revisited. *Developments in Business Simulations and Experiential Exercises*, 21: 83–86.

Gosenpud, J. & J. Washbush (1993). The Relationship Between Total Enterprise Simulation Performance and Learning. *Developments in Business Simulations and Experiential Exercises*, 20: 141.

Hemmasi, M. & L.A. Graf (1992). Managerial Skills Acquisition: A Case for Using Business Policy Simulations, *Simulation & Gaming* 24: 298–410.

Keys, B. & J. Wolfe (1990). The Role of Management Games and Simulations in Education and Research, *Journal of Management*, 16: 311.

Klabbers, J.H.G. (1996). Problem Framing Through Gaming: Learning to Manage Complexity, Uncertainty, and Value Adjustment. *Simulation & Gaming* 27: 74–92.

Kirkpatrick, D. L. (1998). *Evaluating Training Programs: The Four Levels* (2nd Ed.). San Francisco: Berrett-Koehler.

Krathwohl, D.R., B.S. Bloom, & B.B. Masia (1964). *Taxonomy of Educational Objectives: The Classification of Educational Goals; Handbook II: Affective Domain.* NY: David McKay.

McGehee, W. & Thayer, P. (1961). *Training in Business and Industry.* NY: John Wiley and Sons.

Miles, W.G., Jr., W.D. Biggs, & J.N. Schubert (1986). Student Perceptions of Skill Acquisition Through Cases and a General Management Simulation: A Comparison. *Simulation and Games* 17: 7–24.

Miller, H.E., P.L. Schumann, P.H. Anderson, & T.W. Scott (1998). Maximizing Learning Gains in Simulations: Lessons from the Training Literature. *Developments in Business Simulation and Experiential Exercises,* 25: 217–223.

Norris D.R. & C.K. Snyder (1982). External Validation of Simulation Games. *Simulation & Games* 13: 73–85.

Parasuraman, A. (1981). Assessing the Worth of Business Simulation Games. *Simulation and Games* 12: 189–200.

Schellenberger, R.E., J.A. Hill, & R.B. Keusch (1989). An Exploratory Study of the Effect of Strategic Emphasis in Management Games on Attitudes, Interest, and Learning in the Business Policy Course. *Developments in Business Simulations and Experiential Exercises*, 16: 178.

Teach, R. (1990). Designing Business Simulations. In J.W. Gentry (ed.), *Guide to Business Gaming and Experiential Learning* (pp. 93–116), East Brunswick: Nichols/GP Publishing.

Teach, R. & G. Govahi, (1988). The Role of Experiential Learning and Simulation in Teaching Management Skills. In P. Sanders and T. Pray (eds.), *Developments in Business Simulation & Experiential Exercises*, 15: 65–71.

Wolfe, J. (1981). Research on the Learning Effectiveness of Business Simulation Games: A Review of the State of the Art. *Developments in Business Simulations and Experiential Exercises*, 9: 72.

Wolfe, J. (1985). The Teaching Effectiveness of Games in Games in Collegiate Business Courses: A 1973–83 Update. *Simulation and Games*, 16: 251–288.

Wolfe, Joseph (1987). The Teaching Effectiveness of Games in Collegiate Business Courses: A 1973–83 Update. *Simulation & Games*, 16: 251–288.

Wolfe, J. (1990). The Evaluation of Computer-based Business Games: Methodology, Findings, and Future Needs. In J.W. Gentry (ed.), *Guide to Business Gaming and Experiential Learning* (pp. 279–300), New York: Nichols/GP Publishing.

Wolfe, J. & C.R. Roberts (1986). The External Validity of a Business Management Game: A Five Year Longitudinal Study. *Simulation & Games,* 17: 45–59.

Wolfe, J. & C.R. Roberts (1993). A Further Study of the External Validity of Business Games: Five-Year Peer Group Indicators. *Simulation & Gaming* 24: 21–33.