Developments in Business Simulation & Experiential Exercises, Volume 13, 1986

THE DILEMMA IN EVALUATING CLASSROOM INNOVATIONS

Ernest F. Cooke, Memphis State University

ABSTRACT

This is one of about ten papers that try to address various research opportunities and problems of concern to ABSEL members. This paper focuses on the difficulties for everyone concerned; teacher, researcher, editor and reviewer; in reporting and accepting a new type of experiential learning.

INTRODUCTION

If it had not been necessary to be brief ABSEL's name could have been Association for Business Simulation and Other Forms of Experiential Learning." What the organization's name means is that the primary (and only?) concern of its members is experiential learning, particularly through business simulations. Thus the testing of innovations in experiential learning becomes a key consideration.

Many, if not all, ABSEL members have been trained in the scientific method which places an extra burden on their attempts to introduce new or revised experiential learning methods, procedures and techniques into business school classrooms. Some members feel the need to evaluate innovations in the areas of experiential learning in a way that may not be possible and in many cases may not even be appropriate. The feeling is that the value of a learning innovation is not proven unless it has been tested using a rigorous experimental design. At the other extreme there are people who want to publish their new exercise or program without any evaluation at all. The purpose of this discussion is to examine the problems, inherent difficulties, and pitfalls that come from testing a hypothesis that some specific educational innovation does indeed increase learning.

This is a limited discussion. Books can be written on the subject but this paper will cover, briefly, the points of major interest to people who are trying to evaluate experiential learning in the classroom either as researchers or as reviewers. The paper ends with specific suggestions.

THE PROBLEM

The reason for the questionable quality of various papers on educational innovations is the inherent difficulty associated with designing and conducting experiments for the purpose of measuring the learning effects of such innovations. Many introductions and discussions of educational innovations follow a case- study or antidotal format in which the innovation is described and its benefits lauded in abstract terms and/or with subjective comments and testimonials from students and faculty. An excellent outline for this approach was developed by Richard Nordstrom and is in Appendix A of this paper.

The other extreme in testing educational innovations is the experimental design called non-randomized "before and after" with control group. Two identical classes are selected as to student composition, subject matter, instructor and student ability. Both classes take a pre-test and a post-test. One class is taught using the innovation (the experimental class) and the other is taught without the innovation (the control class). The point is that if the educational innovation increases learning, an analysis of both pre- and post-test scores would reveal the difference. The problems with this approach have been discussed in detail in an earlier paper [1] and also in Appendix B. The biggest problem with this approach is that it is almost, impossible to conduct such an experimental design.

WHY TEST AN EDUCATIONAL INNOVATION?

What follows is a brief discussion of Type I errors versus Type II errors. Which of the two is the worst--accepting an educational innovation which does not significantly increase learning or rejecting an innovation which increases learning? Why test an educational innovation?

The reason for testing an educational innovation is to make a decision as to whether to use it in <u>place</u> of some other teaching method. The possibilities are:

1. The innovation is better than the present method.
2. The innovation is as good as the present method.
3. The innovation is not as good as the present method.

Only the third possibility is harmful to the educational process. The exception is if the innovation is more expensive than the alternative or present teaching method. The amount of increased expense would determine the degree to which proof is required that the innovation is better. An analysis of cost- benefits would be required. Refer to Nagel and Neef [3, Chapter 11 for an excellent discussion of how and why this analysis is conducted.

The confidence levels used to determine statistical significance have a sizable effect on how great the difference in means must be for the rejection of a null hypothesis that there is no difference in learning between an experimental class and a control class. In setting confidence levels, the values are only correct if the null hypothesis is rejected. If confidence levels are set at 90% (.90) and, as a result, the null hypothesis is accepted, it emphatically <u>cannot</u> be said that there is a 90% chance that the means are identical. If, on the other hand, the null hypothesis is rejected, it can be said that there is only a 10% chance a true null hypothesis was rejected (a Type 1 error). If the purpose is to reduce the chance of a Type I error, confidence levels are increased but as confidence levels are increased, the chances of accepting a null hypothesis when it is really false is increased (a Type II error).

The point of this discussion focuses on the avowed efforts not to commit a type II error which is to reject an educational innovation because there does not seem to be a statistically significant increase in learning, when in fact, a real increase in learning

did occur. This means an educational innovation should be considered if the null hypothesis can be rejected at confidence levels of 90% and lower.

A researcher would normally be appalled at the idea of confidence limits of 90% but in this particular experimental situation it would be appalling to reject an educational innovation when there is a strong possibility that there was improvement in learning as measured by the post-test.

Instead of requiring a rejection region (alpha-risk) of 5% or 1%, the researcher should be thinking in terms of 10% or 2% or higher [2] because a Type II error (accepting the null hypothesis which says the classes are not different when in reality the experimental class has learned more than the control class) is worse than a Type I error (accepting the alternate hypothesis which says the classes are different when in reality they are the same). In other words, as supporters of these innovations, we do not want to commit a Type II error and our chances of committing this error increases dramatically as we lower the rejection region of our tests.

All of this ignores the possibility that there is no way to measure real learning until five, ten or twenty years later and how could that be done?

CONCLUSION

The basic problem is that reports on educational innovations can vary between two extremes. One extreme is, "1 tried it, it's great!" (Appendix A represents a meaningful improvement on this extreme). The other extreme is, "After examining the results of a perfect experiment we are 99.44% sure that (Appendix B illustrates the impossibility of this extreme). Neither of these two approaches is satisfactory. The first is vague, the second is impossible. The basic tendency on the part of reviewers is towards the impossible extreme.

This is an appeal to an enlightened community of scholars, reviewers and editors to relax certain so-called standards" for the testing of educational innovations. Otherwise, the use of classroom innovations will be discouraged and even more Type II errors will occur. In many of these areas Type II errors are worse than Type I errors.

What can be done? These suggestions are offered for future discussion:

1. Relax the alpha-risk standard in the more rigid statistical approach.
2. Be much more tolerant of the case-study or antidotal approach as in Appendix A.
3. Try quasi-experimental designs, subjective statistics, and so on.
4. Try to get a large number of educators to test the same innovation in several institutions so that requisite sample sizes will be available for testing Type II errors (beta-risk).

This has been a brief discussion of the problems associated with testing classroom innovations. It is certainly not meant to be comprehensive. In conclusion, four specific suggestions have been presented for future discussion.

APPENDIX A

Guidelines for ABSEL Papers on Teaching Innovations
Developed by Richard D. Nordstrom,
California State University-Fresno

Everyone is interested in experiential learning and/or business simulation and gaming seems to be on the alert for any new opportunity to inject a useful, unique, and interesting situation into the classroom. ABSEL has provided a very good forum for this exchange of ideas. The following guidelines are designed to make this exchange a smoother one. A paper discussing a new game or exercise should include sufficient detail for others to decide upon the applicability of the work to their situation.

1. Details of Class Organization.
   (a) Class Size.
   (b) Number of sections
   (c) Size of team or group.
   (d) Length of class in minutes
   (f) Timing of Assignments.
   (g) Student background.
   (h) Timing of reports.
   (i) Nature of "class discussion."
   (j) Administrative support needed.

2. Details of Feedback or Mechanisms for Debriefing.
   (a) When does the debriefing take place?
   (b) How does the debriefing proceed?

3. Details of the Grading System.
   (a) Is the project graded on letter grade or by assignment of points?
   (b) Which items are evaluated? How are they weighted?
   (c) What percent of total grade is assigned this part of the course?
   (d) Is grading one part of debriefing?
   (e) Is grading a one time assignment or assigned in parts?
   (f) Is the project subject to examination or quiz? If so what type?
   (g) Who grades? (instructor, assistants, class members or business leaders).

4. Details to Guide in Preparation for Class Use.
   (a) How far in advance should a prospective user start to get ready to use the exercise?
   (b) How much time does each part or phase require?
   (c) What resources are useful?
   (d) Can a person do this alone or is it wise to get some help from other faculty or the business community?
   (e) Based on your knowledge and experience in using the exercise, what can be done to avoid errors?
   (g) Are there any modifications that might be worthy of consideration?
   (h) How long does it take to grade the work?

5. Your experience with the game or exercise.
   (a) Number of uses.
   (b) Do you intend to use it in the future?
   (c) What are its learning objectives?
   (d) Is it effective?

In summary, it is our view that papers are written for the purpose of widening the application of experiential learning concepts. The free exchange of ideas at an ABSEL conference is proof of that statement. Incorporation of these guidelines through careful documentation in the report or by use of an appendix may improve the opportunity for others to have similar experiences.

APPENDIX B

This material is an update on part of an earlier paper by the author [1].

EXPERIMENTAL DESIGN

A popular method of testing educational innovations is the experimental design called non-randomized "before and after with control group. Two identical classes are selected as to student composition, subject matter, instructor and ability. Both classes take a pre-test and post-test. One class is taught using the innovation (the experimental class) and the other is taught without the innovation (the control class). The point is that if the educational innovation increases learning, then an analysis of both pre- and post-test scores will reveal a difference.

For an ideal experiment these conditions must be met:

(1) There must be two sections of the same course offered during the same semester; furthermore, the material must be presented using the same syllabus and taught in the same way (except for the educational innovation.) More specifically, this means you must use the same instructor and that the instructor must make sure her/his every action or statement is duplicated exactly in the other class (except for the educational innovation.) The two sections must be the same days of the week and the same length of time and at similar times.

(2) Each class must be identical on all unmeasured factors that could possibly explain learning difference. For example: ability, motivation, age, future educational plans, time available for study, prior course work and so on. Randomization or matching of students satisfies this constraints but these processes are usually not possible; therefore, pre-test scores must provide the basis for group identification and measurement. In other words, pre-test scores must have identical means and standard deviations. (See note in Table 1).

(3) To avoid a "Hawthorne" or placebo effect, the students in the class with the educational innovation should not be aware that the other class is taught differently or vice-versa.

(4) The students must make every effort to score as high as possible on both the pre-test and the post-test.

(5) The tests, the educational innovation and the subject matter of the course should be highly related to each other.

(6) The tests should be designed so that pre-test scores are relatively low but should contain no more than one zero and post-test scores should be higher but contain no more than one perfect score. This is to avoid a floor and ceiling effect.

(7) The pre-test and the post-test in both classes should be administered under identical and ideal conditions.

Obviously, it is impossible to meet all the above conditions. Some are more important than others and the expected deviations and the ramifications of these deviations will be discussed below.

As a matter of practicality, the educational innovation must be compared to some type of teaching technique except on the rarest occasions, when it can be introduced and used without taking any class time. If the innovation requires class time, then the experimental class will have less time for whatever other teaching methods are used than the control class. This means that the two classes will have some dissimilarity that cannot be avoided. However, the innovation can be compared to what is extra in the control class. Since lecture is frequently considered the least effective method of teaching, the innovation could be compared to lecture. Lecture is probably the minimum benchmark.

STATISTICAL TESTS

Assuming that all of the conditions outlined above have been satisfied, we end up with four Sets of mean values and standard deviations, specifically, the means and standard deviations of the pre-tests and post-tests for both classes. The number of students is the number who took both pre-test and post-test in a given class. In Table 1 the statistics available from the experiment are shown.

The successful statistical test requires rejecting a null hypotheses which states that the true mean of both populations are equal and accepting an alternate hypotheses which says that the true mean of the experimental population is greater than the true mean of the control population. This is a one-tail test. This test would be considered sucessful, because the results show higher test scores in the experimental class. If the null hypotheses is not rejected, then this reduces but does not eliminate the likelihood that the educational innovation has improved learning.

To test the null hypothesis, some degree of desired significance is established and this factor determines the right-hand boundary (one-tail test) between acceptance and rejection of the sampling distribution. This is compared to a ratio of the difference between sample means and the unbiased estimator of the standard error of the difference between means.

If the ratio is less than the value of the right hand boundary, the null hypothesis is rejected. Rejection of the null hypothesis means acceptance of the alternate hypothesis, that is the higher mean score of the experimental class is statistically significant.

The ratio which is compared to the boundary is made up of the six factors shown in Table 1 as well as the two factors which determine the boundary condition. The relationship of all of these factors determine acceptance or rejection of the null hypothesis. See Table 2.

Obviously, if $\bar{X}_{pe}$ increase and/or $\bar{X}_{pc}$ decreases, the numerator of the ratio is larger and more likely to fall to the right (rejection region) of the boundary.

The denominator of the ratio is a fraction with the sample standard deviations in the numerator and the class size in the denominator; therefore, any decrease in standard deviations or increase In sample size will decrease the denominator of the ratio, thus increasing its value and making it more likely to fall to the right (rejection region) of the boundary.

The right-hand boundary is determined by the alpha risk or confidence level and by the class sizes. If either class falls below thirty students, it is necessary to use a t-test instead of a Z-test. Using the t-test means the value of the right-hand boundary is increased, which requires a higher ratio for rejec-

tion. If the confidence is increased (reduced alpha risk), the value of the right-hand boundary is increased, which requires a higher ratio for rejection.

TABLE 1

STATISTICS AVAILABLE FROM EXPERIMENT

Class	Pre-Test	Post-Test
Experimental	n_e	n_e
	$\bar{x}_e$	$\bar{x}_{pe}$
	s_e	s_{pe}
Control	n_c	n_c
	$\bar{x}_c$	$\bar{x}_{pc}$
	s_c	s_{pc}

n = number of cases (students)

$\bar{x}$ = mean value of test scores

s = standard deviation of test scores

The reason for testing the educational innovation is to see if it will improve learning.

The measure of the true gain in learning is:

$$(\bar{x}_{pe} - \bar{x}_e) > (\bar{x}_{pc} - \bar{x}_c)$$

When $\bar{x}_e = \bar{x}_c$ (see condition number two (2) above under Experimental Design) we have a special case which reduces to:

$$\bar{x}_{pe} > \bar{x}_{pc}$$

If $\bar{x}_{pe}$ is greater than $\bar{x}_{pc}$, indicating increased learning due to the educational innovation, then we must test for statistical significance.

TABLE 2

FACTORS DETERMINING REJECTION OF NULL HYPOTHESIS

Factors in Ratio	Null Hypothesis is Rejected as Factor
$(\bar{x}_{pe} - \bar{x}_{pc})$	Increases
n_e or n_c	Increases
s_e or s_c	Decreases
Factors in Boundary Condition	
Alpha risk or Confidence Level	Increases
	Decreases
n_e or n_c	Increases (degrees of freedom)

In Table 3 some numerical examples are shown to illustrate these points. In these examples, it is assumed the mean of the post-test scores in the control class (xpc) is 70.0. The mean_ of the post- test scores in the experimental class (xpe) shown in Table 3 is the minimum value required to reject the null hypothesis under the conditions indicated. The pre-test scores are assumed identical and lower than 70.0.

The significance of these examples are as follows:

1. As confidence levels increase from 90% to 99% (examples 3 and 1), the differences in mean scores increases from a difference of 6.2Z to a difference of 11.3%.

2. As the sample standard deviation increases from ten to fifteen (examples 4 and 3), the difference in mean scores increases from 4.1% to 6.2%.

3. As sample size goes from forty to twenty, students in each class (examples 4 and 6), the difference in mean scores increases from 4.1% to 6.2%.

4. Finally, if post-test scores are in the thirties instead of the seventies with a slight reduction in standard deviation (examples 7 and 6) the differences in mean scores increases from 6.2% to 9.8%. This last example illustrates one reason why It is important to have higher post-test scores. If post-test scores are lower, and there is no corresponding reduction in sample standard deviation the result is the need for more of a percentage difference in mean scores to be statistically significant.

It may not seem that significant, but it takes a lot of something extra to get a class that was destined to have average grades of 70.0 raised up to 75.0 or more. After all, there has to be some value to the existing teaching method. For this reason, there is a need for large classes, lover confidence levels and high post- test scores to get any kind of decent statistical results (example 4 in Table 3 which is a 4.1% increase). Otherwise, the difference in scores is just impossible to achieve (example 8 in Table 3 which is a 18.8% increase).

FURTHER STATISTICAL DIFFICULTIES

The ratio used to determine if the difference in post- test mean scores ($\bar{X}_{pe} - \bar{X}_{pc}$) is statistically significant is calculated by dividing the difference in sample means by the unbiased estimator of the standard error of the difference between means. The denominator in the ratio is a function of sample size and the standard deviation of the sample.

As can be seen in Table 3 with the third and fourth examples, an increase in the sample standard deviation requires a much higher difference in post- test means to be statistically significant.

Recognize that if a class is normally distributed in ability and motivation from A students to F students, you would expect a large deviation as compared to a class of all B students. The usual large class contains a group of students with a wide range of knowledge, ability and motivations; consequently, the standard deviation of grades on any given test can be expected to be high and this factor is reflected in the denominator of the ratio used to determine statistical significance in the difference between two means. Therefore, the difference in means must be

TABLE 3

NUMERICAL EXAMPLES SHOWING NULL HYPOTHESIS REJECTION

Key to the headings:

A = Example Number
B = Number of Students in Each Class (held equal from pre-test to post-test in each class for these examples).
C = Standard Deviation of Each Class (assumed equal in each class for these examples).
D = Confidence Level
E = Alpha Risk
F = Right-Hand Boundary Between Acceptance and Rejection.
G = Mean Value of All Post-test Scores in the Control Class.
H = Minimum Possible Mean Value of All Post-test Scores in the Experimental Class if the Null Hypothesis is Rejected.
I = Percent Increases of Experimental Class Scores over Control Class Scores.

A	B	C	D	E	F	G	H	I
1	40	15	.99	.01	2.327*	70.0	77.9	11.3%
2	40	15	.95	.05	1.645*	70.0	75.6	8.0%
3	40	15	.90	.10	1.282*	70.0	74.4	6.2%
4	40	10	.90	.10	1.282*	70.0	72.9	4.1%
5	30	10	.90	.10	1.282*	70.0	73.4	4.8%
6	20	10	.90	.10	1.328**	70.0	74.3	6.2%
7	20	8	.90	.10	1.328**	35.0	38.4	9.8%
8	20	8	.99	.01	2.539**	35.0	41.6	18.8%

* Z-test
** t-test

greater to be statistically significant than would be necessary if we had a class that was close together in ability and motivation.

These comments are based on having two classes with at least 30 students in each class. If the class size is less than 30, the researcher must use a t-test, because it can not be assumed that the sampling distribution is normally distributed. Defined further, this means that the right-hand boundary between rejection and acceptance will increase, a fact complicating the problems outlined above. In addition, the class size enters the denominator in such a way that the ratio is reduced and rejection is still harder to achieve. This type of study hinges on rejection of the null hypothesis since rejection proves the classes are different, but if the null hypothesis is not rejected, that does not mean that the classes are identical. In fact, if alpha is at 20%, there is still an 80% chance that the classes are different.

REFERENCES

[1] Cooke, Ernest F. "Trials and Tribulations in Testing Educational Innovations," in Samuel C. Certo and Daniel C. Brenenstuhl (Editors), Insights into Experiential Pedagogy, Tempe: Arizona State University, 1979, pp. 209-212.

[2] Nagel, Stuart S. and Marian Neef, Policy Analysis in Social Science Research. Beverly Hills, California: Sage, 1979.