Insights into Experiential Pedagogy, Volume 6, 1979

TRIALS AND TRIBULATIONS IN TESTING EDUCATIONAL INNOVATIONS

Ernest F. Cooke, University of Baltimore

ABSTRACT

The author has conducted a number of experiments for the purpose of testing the hypothesis that the use of simulations *in* marketing classes will increase learning. A paper has been published describing the results of the initial experiment [l]. The author has also reviewed the results of a number of similar experiments, both published and unpublished. The ideal experimental design, the insurmountable difficulties in achieving that ideal and the compromises that have to be made are discussed in this paper.

INTRODUCTION

A key question for members of ABSEL is whether educational innovations such as simulations and Other experiential exercises increase learning. The purpose of this paper is to discuss the problems, inherent difficulties and pitfalls in testing the hypothesis that some specific educational innovation does indeed increase learning.

Every ABSEL proceedings published so far (five from 1974 to 1978) contained one or more papers describing a specific experiment designed to determine if some educational innovation has increased learning; yet, these papers comprised only about five per-cent of over 200 papers published in these five ABSEL proceedings. Of equal concern is the fact that these articles have represented different levels of quality.

The reason for the questionable quality and probable scarcity of such papers is the inherent difficulty in designing and conducting experiments for the purpose of measuring the learning effects of an educational innovation. Most introductions and discussions of educational innovations follow a case-study format in which the innovation is described and its benefits lauded in abstract terms and/or subjective comments from students and faculty.

The purpose of this paper is to present a realistic discussion of the most popular experimental design for the purpose of testing educational innovation. This discussion is based on the author's experience with this methodology as well as the experiences of other researchers,

EXPERIMENTAL DESIGN

A popular method of testing educational innovations is the experimental design called "before and after" with non-randomized 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-rest 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 instructor must make sure 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 constraint 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.

(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 then 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 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, the innovation should be compared to lecture. Lecture would be 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 I the statistics available from the experiment are shown.

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 above in 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.

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 set. This test would be considered successful because the results show higher test scores in the experimental class. If the null hypotheses is not rejected, then this reduces 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 accepted but if it is greater, then 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 I 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.

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	Increases
or	
Confidence Level	Decreases
n_e or n_c	Increases (degrees of freedom)

Obviously if $\bar{x}_{pe}$ increases 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, this 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 rejection. If the confidence is increased (reduced alpha risk), the value of the right-hand boundary is increased which requires a higher ratio for rejection. 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 e) is 70.0. The mean of the post-test scores in the experimental class (type) shown in Table 3 is the minimum value required to reject the null hypothesis under the conditions indicated.

TABLE 3

NUMERICAL EXAMPLES SHOWING NULL HYPOTHESIS REJECTION

Key to the headings:

A = Example Number
B = Number of Students in Each Class (held equal 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 Reject-

ion.
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 = Per-cent Increases of Experimental Class Scores over Control Class Scores.

A	B	C	D	E	F	G	H	I
1	40	15	.99	.01	1.960*	70.0	76.7	9.6%
2	40	15	.95	.05	1.645*	70.0	75.6	8.0%
3	40	15	.90	.10	1.282*	70.0	74.4	6.3%
4	40	10	.90	.10	1.282*	70.0	73.0	4.3%
5	30	10	.90	.10	1.282*	70.0	73.4	4.8%
6	20	10	.90	.10	1.328**	70.0	74.3	6.1%
7	20	8	.90	.10	1.328**	35.0	38.4	9.7%

* Z-test
** t-test

1) As confidence levels increase from 90% to 99%, the difference in mean scores increases from a difference of 6.3% to 9.6%. 2) As the sample standard deviation increases from ten to fifteen, the difference in mean scores increases from 4.3% to 6.3%. 3) As sample size goes from twenty to forty students in each class, the difference in mean scores decreases from 6.1% to 4.3%. Finally if we have post-test scores in the thirties instead of the seventies with a slight reduction in standard deviation, the differences in mean scores in- creases from 6.1% to 9.7%. This last example illustrates one reason why it is important to have higher post-test scores. If post-test scores are lower, we do not get a corresponding reduction in sample standard deviation and 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, we are emphasizing large classes, lower confidence levels and high post- test scores.

WHY TEST AN EDUCATIONAL INNOVATION

This is, of course, a discussion of Type I errors versus Type II errors. Which is worse, accepting an educational innovation which does not significantly increase learning or rejecting an educational innovation which does increase learning? What is a significant difference between sample means in the test of an educational innovation? Why test an educational innovation?

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

(1) The innovation is better.
(2) The innovation is as good.
(3) The innovation is not as good.

Only the third possibility is harmful to the educational process. The exception is of the innovation is more expensive than the alternative (traditional) teaching method. The degree 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.

As can be seen by looking at the first three examples in Table 3 the confidence levels used to determine statistical. significance have a sizable effect on how great the difference in means ($x_{pe} - x_{pc}$) must be for rejection. In setting confidence levels, the values are only valid if the null hypothesis is rejected. If confidence levels are set at 90% (.90) and, as a result, the null hypothesis is accepted, we emphatically can not say that there is a 90% chance that the means are identical. If, on the other hand, the null hypothesis is rejected, we can say that there is only a 10% chance we rejected a true null hypothesis (Type I error). If we want to reduce the chance of a Type I error, we increase our confidence levels BUT as we increase our confidence levels, we increase the chances of accepting a null hypothesis when it is really false (Type II error).

The point of this discussion focuses on our avowed efforts not to commit a Type II error which is to reject an educational innovation because we do not see a statistically significant increase in learning when in fact there does occur a real increase in learning. This means we should consider an educational innovation if we can reject the null hypothesis at confidence levels of 90% and maybe even lower.

Normally, a researcher would be appalled at the idea of confidence limits of 90% but in this particular experiment 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.

FURTHER STATISTICAL DIFFICULTIES

The ratio used to determine if the difference in post- test mean scores ($x_{pe} - x_{pc}$) is statistically significant is calculated by dividing the difference in sample means by 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 motivation; 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 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. It 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. The whole

study hinges on rejection of the null hypothesis since rejection proves the classes are different.

DISCUSSION

In six experiments using simulation at the University of Baltimore, we have learned a great deal through our errors. The following comments should provide some help to other researchers.

To set up a successful before and after experiment with control group, the researcher needs to start with two large classes which are identical in all possible respects. To illustrate one problem, that occurred in an experiment we tried, the following happened: The pre- test showed the two classes to be sufficiently close but, during the semester, we had students withdraw frog both classes. Due to this attrition factor, when we looked at the pre-tests of the survivors at the end of the semester, the two classes were no longer sufficiently close and unfortunately the experiment was invalidated.

Another reason the study needs to have large classes with the sample size above 30, is to eliminate the use of a t-test. Using a t-test makes it harder to accept the alternate hypothesis that the classes are different.

Pre-test scores should be low and the post-rest scores should be high. Thus the pre-test must be difficult for students starting the course and the post-test not too difficult for students finishing the course. This way the test will reflect a maximum amount of learning and any differences in the amount of learning will be more apparent. If there are too many zeroes on the pre-test, the test was too hard and is not valid because it is truncated • There is a floor effect. The same applies if there are too many prefect scores on the post-test, resulting in a ceiling effect. Obviously the test must reflect what is being taught during the semester and be relevant to the educational innovation.

If the researcher is able to get through the semester with enough surviving students whose pre-test scores are close and whose post-test show more learning in the experimental class, then the researcher must carefully think through the basis for accepting or rejecting the hypothesis that greater learning has taken place in the experimental class.

Instead of requiring a rejection region (alpha-risk) of 5% or 1% the researcher should be thinking in terms of 10% or possibly even 20% because a type II error (accepting the null hypothesis which says the classes are not different, when in reality the experimental class has done better 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 educational innovations, we do not want to commit a Type II error and our chances of committing a Type II error increases as we lower the rejection region (alpha-risk).

CONCLUSION

This paper has gone into some detail concerning the statistics involved in testing the hypothesis that an educational innovation has increased learning. This is important because normal researchers will tend to accept the standard statistical test without questioning
the implications for what they are trying to determine. As believers in simulations and other types of experiential learning, we are trying to prove that these educational innovations do increase learning.

If the experimental design, “before and after” with control group, is used, it requires considerable good luck to be successful. The luck comes in finishing the semester with two classes whose survivors started Out essentially equal and whose respective test scores show a predicated difference at an acceptable level of statistical significance. Many of the experiments, both ours and others, published and not published, reviewed for this paper, suffered from insufficient attention to this point. If you read a paper which says the classes started out about equal according to some vague criterion set down by the researcher, you can be reasonably sure that the results are suspect. If the class (sample) sizes are small, the results should be even more suspect. The importance of starring out with essentially identical classes and knowing this class factor as the result of a suitable pre-test can not be minimized.

If the researcher is fortunate enough to wind up with post-tests for a reasonably large experimental and control class, the next step is to intelligently, not blindly, apply the required statistical tests. Depending on how close together the pre-test scores are, this could mean acknowledging that more learning has occurred in the experimental class even though statistical significance is not shown at the 0.05 alpha risk level. The importance of not blindly accepting traditional levels of alpha risk (0.05 or 0.01) cannot be minimized.

REFERENCES

[1] Cooke, Ernest F. and Maronick, Thomas J., “Simulations Do Increase Learning,’ in Barnett A Greenberg and Danny N. Bellinger (editors), Contemporary Marketing, 1977 Educators’ Proceedings. (Chicago: American Marketing Association, 1977).