Instructional Product Evaluation Using the Staged Innovation Design

Norman Wagner
Instructional Design Services
Arizona State University
Tempe, AZ 85287

Abstract. In instructional program evaluation, pretest scores typically are compared to posttest scores, and any change is attributed to the instructional program. Such a pretest-posttest evaluation design ignores the possible existence of threats to internal validity. In contrast, the Staged Innovation Design makes use of an experimental and a control-replication group by not introducing the program to all learners at once. Use of the Staged Innovation Design allows for posttest scores for an experimental group to be compared to both pretest and posttest scores of a control-replication group to ascertain the instructional effectiveness of the program. In the study reported here, the Staged Innovation Design was modified with the addition of a simultaneous pretest for both groups prior to the introduction of the program to the control group. As altered, this design controls for threats to internal validity while providing options for multiple comparisons on which to base a contention that the program is instructionally effective.

A typical method of evaluating instructional programs begins with the administration of a pretest to assess subjects' knowledge or competencies related to the objectives of the program. The subjects then are exposed to an instructional program, then given the test again. Scores from the pretest and the posttest are compared to determine whether a significant change occurred in the learners between the two test administrations. Any change usually is attributed to the introduction of the instructional program.

However, there are explanations other than the effectiveness of instruction which can be used to explain the change in learner performance. These explanations, or threats to internal validity, are outlined in the literature (Campbell, 1969). They include factors such as the instability of the test, the effect of multiple testings, regression toward the mean, maturation of the learners over the course of the experiment, and other events that may have occurred in the lives of the learners to influence their test scores (history).

Some of these threats can be controlled through use of evaluation designs such as the various time series designs (Cook & Campbell, 1979) and the Outcome-Consequence Model (Hannafin, 1983). In time series designs, an instructional program is introduced to learners following a series of pretests. Data from these pretests are compared to data from a post-instruction series of tests, to obtain results free from short term or irregular fluctuations in performance. In the Outcome- Consequence Model, variables such as differing levels of prerequisite skills among the learners and retention of learned content over time are examined as part of the evaluation.

In a true experiment, that is, one in which subjects are randomly assigned to treatment groups and groups are randomly assigned to treatment, threats to internal validity are minimized (Asher, 1976). However, given that instructional developers and evaluators seldom have the control necessary to create the conditions required for a true experiment, quasi-experiments often represent the next best choice.

One evaluation design that can satisfy the requirements necessary to qualify as quasi-experimentation is the Staged Innovation Design (Campbell, 1969; Salomon & Clark, 1977). The design is useful for situations in which all subjects must be exposed to a treatment, but it is not necessary that all subjects receive the treatment simultaneously. In staged innovation, subjects are divided into two groups, an experimental group and a control-replication group. The instructional program is introduced to these groups in two stages.

In the first stage, the experimental group is provided with instruction. This group is posttested while the control-replication group is pretested using the same instrument. Results of the posttest administered to the experimental group are compared to results of the pretest administered to the control group (Comparison 1, Figure 1) in order to ascertain the degree of effectiveness of the instruction. In the second stage, the control-replication group is exposed to the treatment. Their posttest results are compared to the posttest results of the experimental group, thus replicating the study (Comparison 2, Figure 1).

Method

This study involved the introduction of a multicultural education unit into an undergraduate educational psychology course at Purdue University. The course was taught in seven class sections of approximately thirty students each. Class sections were assigned randomly to an experimental group (Group I) and a control-replication group (Group II).

The unit consisted of reading materials and small group activities developed for this study using generally accepted principles of instructional design. The test and activities were based on 27 cognitive and two affective objectives. Written materials were distributed to the students for study at home over a weekend. The experimenter supervised three class periods of small group activities for all class sections.

The same multiple choice test was used for both pretesting and posttesting and was administered to students during regularly scheduled class periods. All items had four response choices and were designed to produce norm-referenced results. The test was divided into two sections: cognitive and attitude. The first 20 items constituted an objectives-based test designed to determine the amount of learning between

pre- and posttesting. The last eight items were designed to measure student attitude toward multicultural education and were used to assess the extent of attitude change resulting from the program. All test items related directly to objectives. However, not all cognitive objectives were tested. Achievement of several of these objectives was assessed through performance of the small group activities. Data were analyzed separately for the cognitive and attitude sections of the test. Kuder-Richardson Formula 20 reliability estimates for the 20 cognitive items in the posttest administration to the experimental and replication groups were .58 and .63, respectively. For the eight attitude items, the estimates were .73 and .70, respectively.

The Staged Innovation Design was modified somewhat for the study reported here. In the first stage of this study, written materials were distributed to Group I students following pretesting of both groups on a Thursday. On Monday, Tuesday, and Wednesday of the following week, the experimenter supervised three periods of small group activities in each Group I class. On Thursday of that week, after Group I finished the unit, both groups were retested. This administration, which served as a posttest for Group I, served as a second pretest for Group II. The following comparisons, which are diagrammed in Figure 2, were made using *t* tests.

1. Comparison 1 tested for differences between the results of the pretest and the posttest administered to Group I, in order to obtain a preliminary estimate of the degree of instructional effectiveness of the unit.

2. Comparison 2 tested for differences between the posttest administered to Group I and the second pretest administered to Group II. This comparison was also designed to confirm findings from Comparison 1 regarding the instructional effectiveness of the unit.

3. Comparison 3 tested for differences between the results of the two pretests administered to Group II, in order to determine whether students who had not been exposed to the unit also gained knowledge. The addition of this second pretest for Group II provided an opportunity to assess the extent to which testing effects and test reliability existed in this study.

After the second testing, the unit was distributed to Group II students. The following Monday, Tuesday, and Wednesday, the experimenter supervised three periods of small group activities in each Group II class. These students were tested for the third time on Thursday of that week. Scores from this administration served as posttest results for Group II. Repeated measures analyses of variance with unweighted means solutions for unequal group sizes were used to compare the pretest and posttest scores of both groups.

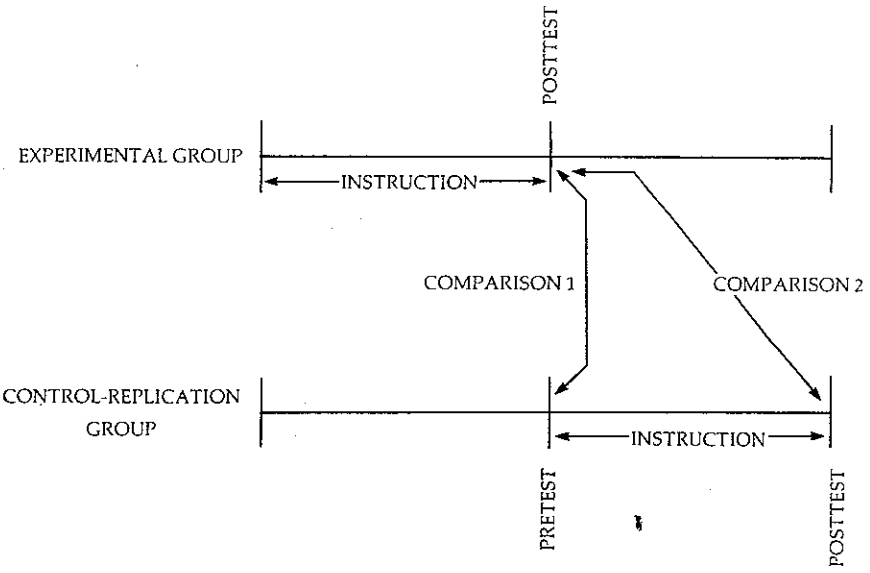

Figure 1
Instructional Product Evaluation Using the Staged Innovation Design

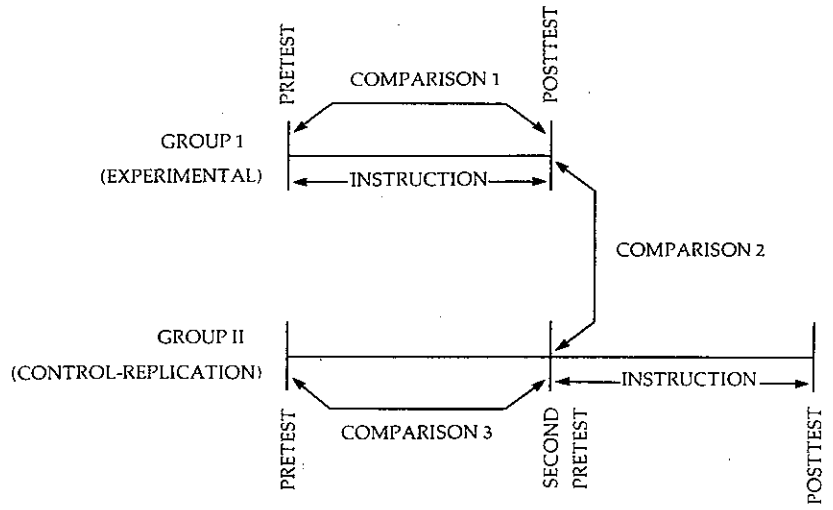

Figure 2
Instructional Product Evaluation Using the Staged Innovation Design

Results

Pretest and posttest means and standard deviations for the cognitive and attitude test items are reported in Table 1. Separate *t* tests and analyses of variance were conducted for each test section (see Tables 2 and 3).

Cognitive. The ANOVA conducted using data from the cognitive items yielded no main effect for group, indicating that no significant difference existed between the knowledge of the two groups either prior to or following instruction. However, a significant difference was found between the pretest and posttest results for both groups. The presence of this main effect was interpreted as evidence that the unit was effective in transmitting knowledge about multicultural education, as measured by this test.

Comparison 1 was conducted using a *t* test for correlated data to compare the pretest and posttest results for Group I.

Table 1
Pretest and Posttest Means and Standard Deviations for the Performance and Attitude Test Items

	n	Pretest	Second Pretest	Posttest
Performance Scores				
Group I	103			
Mean		7.26		9.96
S.D.		2.44		3.00
Group II	90			
Mean		7.11	6.77	11.06
S.D.		2.30	2.11	3.21
Attitude Scores				
Group I	103			
Mean		4.74		5.23
S.D.		1.70		2.00
Group II	90			
Mean		4.91	5.04	5.98
S.D.		1.66	1.63	1.85

Table 2
ANOVA Summary Table for Pretest and Posttest Scores on the Performance Test Items

Sources of Variation	df	MS	F
Between Subjects	192		
Group Main Effect	1	20.8753	2.0652
Subjects Within Groups	191	10.1083	
Within Subjects	193		
Test Main Effect	1	1059.9458	206.8753*
Interaction	1	37.7468	7.3662*
Tests by Subjects Within Groups	191	5.1243	

*$p < .05$

Table 3
ANOVA Summary Table for Pretest and Posttest Scores on the Attitude Test Items

Sources of Variation	df	MS	F
Between Subjects	192		
Group Main Effect	1	9.1442	1.6088
Subjects Within Groups	191	5.6840	
Within Subjects	193		
Test Main Effect	1	83.3341	98.7647*
Interaction	1	1.7579	2.0835
Tests by Subjects Within Groups	191	.8438	

*$p < .05$

This analysis indicated that a significant gain in knowledge occurred between the two test administrations. Comparison 2 consisted of a test for differences between the posttest administered to Group I and the second pretest administered to Group II. Results of a *t* test for two sample means revealed that the scores on the cognitive items for Group I students after receiving instruction were significantly better than the scores of Group II students who had not received instruction. Comparison 3 involved a test for differences between the scores on the two pretests administered to Group II. Results of a *t* test for correlated data demonstrated that no significant difference existed between these two sets of scores. That is, the performance of Group II students was not significantly different on these two administrations, between which the students received no instruction related to multicultural education. However, Comparison I demonstrated that a significant pretest to posttest gain occurred for Group I students, who received instruction between test administrations. Considered together, the results of these *t* tests lend strength to the contention that the unit was successful at increasing student knowledge about multicultural education.

Attitude. The analysis of variance conducted on data from the attitude items yielded no main effect for group, indicating that the groups were similar in attitude both prior to and following instruction. However, the main effect for test was significant. This was interpreted to mean that the unit was effective at creating more positive student attitudes.

Comparisons 1, 2, and 3 were conducted on the attitude data using the same types of *t* tests that were used for the data from cognitive items. Comparison 1 yielded a significant improvement in the attitude of Group I students from pretest to posttest. For Comparison 2, the attitude scores on the posttest administered to Group I were significantly more favorable than the scores on the second pretest administered to Group II. Comparison 3 revealed no significant difference between the attitude scores on the two pretests administered to Group II. That is, the attitudes of students who were not exposed to the instruction did not change between test administrations. The results of Comparisons 1 and 2 were interpreted as indication that the unit was effective in producing more favorable attitudes toward multicultural education as measured by the test items. The lack of significant difference for Comparison 3 provided evidence that the pretest to posttest changes in scores were due to the treatment and not to other variables.

Discussion

Analyses of variance and multiple comparisons using t tests were conducted separately on the data from the cognitive and attitude test items for both the experimental and the control-replication groups. The results of these analyses revealed statistically significant increases in knowledge about and favorable attitudes toward multicultural education. Ordinarily, statistical significance alone is not an adequate measure of the value of an instructional program. However, in this case, results were interpreted as support for the contention that the unit was effective.

The Staged Innovation Design, when employed as outlined here, represents an improvement over the commonly used pretest-posttest design for two reasons. First, the comparison of scores obtained from the two pretests administered to the control-replication group (Comparison 3, Figure 2) can produce evidence regarding the existence of the five threats to internal validity that were

mentioned earlier.

Two of these threats to internal validity involve events other than instruction that may be responsible for a difference in subjects' pretest and posttest scores: history and maturation. Often, these events do not represent major threats in short-term evaluation studies. However, if available groups are randomly assigned to the experimental or control-replication conditions, and a change occurs in the experimental but not in the control subjects, then these threats to internal validity need not be of concern.

The other threats to internal validity mentioned here are the instability of the measuring instrument, the effect of testing, and the regression toward the mean from pretest to posttest by individuals selected for instruction because of their extreme scores. Using the Staged Innovation Design, these occurrences would be manifest equally in the control and experimental groups. Thus, if no change in scores is apparent between the two pretests given the control group, it can be assumed that any change in the experimental group between the pretest and the posttest can be attributed to instructional intervention.

A second reason why the Staged Innovation Design as described here as superior to the pretest-posttest design is that introducing a program in stages provides opportunity for multiple comparisons of instructional effectiveness. These opportunities include the use of analysis of variance or *t* tests to compare pretest to posttest scores, and the comparison of scores from a posttest administered to the experimental group to scores from a pretest administered to the control-replication group at the same time.

The Staged Innovation Design, as modified for this study, could be further expanded to accommodate additional groups. For example, during the second stage, a new control-replication group (Group III) could be added. This new group could serve as a control group during the delivery of instruction to Group II. Following the posttesting of Group II, the delivery of instruction to Group III would constitute a second replication of the study. A design such as this would be practical for implementation of programs at multiple sites. Comparisons similar to those made between Groups I and II could be conducted between Groups II and III, or among all three groups. Analysis of variance could be used to establish pretest and posttest equivalence among all groups. A second posttest administered later to any of the groups could be used to establish the long-term instructional effectiveness of the program.

In many cases, the benefits derived from the Staged Innovation Design as modified for this study may not be justified in terms of the extra cost, instructor time, and student effort required by the multiple test administrations. In these cases, the design as proposed by Campbell (1969) and by Salomon and Clark (1977) is more practical. Using this design, the experimental group is posttested only, and the control-replication group is both pretested and posttested (Figure 1). Whether employed in this manner or as modified for this study, the Staged Innovation Design can provide a practical and experimentally sound framework for the evaluation of instructional programs.

References

Asher, J. W. (1976). *Educational research and evaluation methods.* Boston: Little, Brown and Company.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24,* 409-429.

Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally College Publishing Company.

Hannafin, M. J. (1983). Measuring the importance of learning from instruction. *Journal of Instructional Development, 6*(3), 14-18.

Salomon, G., & Clark, R. E. (1977). Reexamining the methodology of research on media and technology in education. *Review of Educational Research, 47,* 99-120.