A National Experiment in Curriculum Evaluation

WAYNE W. WELCH
University of Minnesota

HERBERT J. WALBERG
University of Illinois[1]

Although the need for true experiments on broadly defined populations has long been recognized, there are very few local experiments and no national experiments in curriculum research. For example, among 46 government-sponsored course development projects in science and mathematics, a few relied on teacher reports and classroom visits for evaluation, but only four used true experiments in their evaluation strategies (Welch, 1969). The purpose of the present paper is to report the feasibility of a national educational experiment and to present the summative findings regarding Project Physics, a physics course for high school students.

THE COURSE

The developers of Project Physics were originally concerned about the continuing drop in the proportion of students who take physics in high school. To attract students who are not bound for mathematical, scientific, or technical careers, and without compromising on the physics content, they attempted to develop an interest-awakening,

[1] We thank our former colleagues at Harvard; Andrew Ahlgren, James Rutherford, Fletcher Watson, and our outside research consultants, Darrell Bock, Gene Glass, and Les McClean for advice during the research. Mark Bargen and Maureen Flahaven provided valuable editorial assistance in preparing the final manuscript.

module system of course components using a variety of media and methods for learning: a basic text, film loops, programmed instruction booklets, transparencies, laboratory apparatus, special cameras, a student handbook, and other materials. The structure of the course allows students to emphasize aspects which interest them most; for example, rigorous mathematics, laboratory experiments, or historical readings. Perhaps the most distinctive aspect of Project Physics is its humanistic orientation—an attempt to show the place of physics in the history of ideas, and its relation to technology and social development. At the present time (1971-72), the course is being used by approximately 80,000 students in all 50 states. The research reported below was conducted during the final year of the course development. While some 60 other evaluation and research papers are based on Project Physics data (Welch, 1971), the present study concerns only the experimental part of the evaluation.

METHOD

A list of the names and addresses of 16,911 physics teachers was purchased from the National Science Teachers Association (NSTA), which maintains the U.S. Registry of Junior and Senior High School Science and Mathematics Teaching Personnel. The NSTA reported that the list is compiled from responses received from 81 per cent of all secondary schools in the United States. Because of travel costs for teacher training, we limited our population to the 16,702 physics teachers listed for the continental United States. Numbers were assigned to each of the teachers according to his ordinal position on the list and a table of random numbers was used to select a total of 136 names.

Each of the 136 teachers was sent a registered letter describing the curriculum project and inviting him to participate in an experimental evaluation of the course. Each was informed that a teacher agreeing to participate would be randomly assigned to either an experimental group or a control group. The responsibilities of both groups were described in the letter: the experimental group would attend a six-week Briefing Session, take a series of tests, teach the course during the academic year 1967-68, and administer pre, mid, and posttests to their physics students. The control group would attend a two-day briefing session, take a series of tests, and administer the same pre, mid, and posttests to their students: but they would continue to teach their regular physics courses. Travel expenses, summer school stipends, and course materials were to be provided by Project Physics.

A total of 136 letters of invitation were mailed, but only 124 teachers were actually contacted. Nine letters were "returned to sender," and three others could not be reached by telephone follow-up. Of those contacted, 72 agreed to participate according to the conditions specified, while 52 were unable to accept because of prior commitments or lack of interest. The nature and frequency of nonacceptors are listed below:

Continuation of work on Master's degree in summer school	12
No longer teaching physics	11
Summer job commitment	10
Not interested	6
Physics no longer offered at their school	5
Health reasons prevent extended travel	3
Miscellaneous (changing jobs, expecting baby, etc.)	5
	52

Questionnaires were returned by 124 teachers—72 acceptors and 52 non-acceptors. T-tests revealed that teachers who accepted the invitation, when compared to the nonacceptors, are more likely to teach in larger schools and to be currently teaching the Physical Science Study Committee (PSSC) physics course. (PSSC is a recently developed physics course—one of the first of the national curriculum projects.) It seems reasonable to interpret these differences as a greater receptiveness to innovation in larger schools where previous innovations have been accepted. The findings must be interpreted in the light of the sampling limitations; refusals, listing by NSTA, etc. However, the target group for generalization would be the kinds of teachers in the sample, i.e., those willing to try new courses.

A table of random numbers was used to assign 46 of the teachers to the experimental group and 26 to the control group. Because of transfers and illnesses, the final sample consisted of 53 physics teachers. As shown in Table 1, 34 of these attended the six-week summer Briefing Session and taught the course. Because of the possibility of the so-called (and as yet unreplicated) "Hawthorne effect," the 19 control group teachers were brought to Harvard University for two days, entertained by university physicists, and impressed with the importance of their participation in the experiment. They were asked to teach their regular physics courses during the coming academic year.

Instruments

Nearly 40 instruments were suggested or proposed for construction. Independently, three judges assigned priorities to these tests based on perceptions of the goals of the course, availability and

TABLE 1

Cell Sizes for Analysis of Variance

IQ Group	*HPP*	*Other*	*Total*
Low (112.1 or less)	11	6	17
Middle (112.2 - 119.3)	11	8	19
High (119.4 or over)	12	5	17
Total	34	19	53

usability of instruments, conversations with other Project personnel, and the experience of the evaluators. From the long list of instruments suggested, those described in Table 2 were selected because they were believed to represent the goals of the course, the purposes of the evaluation, sample a broad range of anticipated student outcomes, and fit within the restriction of time and problems of testing on a national basis.

Procedure

The system of randomized data collection employed in the testing plan increases the number of testing instruments that can be used in any given class period (Walberg and Welch, 1967). Briefly, a random half of a class takes one test while the other half is taking a different test. Tests for a given administration were arranged randomly before the tests were sent to the teachers. The teachers were asked to hand the first test to the first student in the first row, the second test to the second student, and so on. By this procedure, the assignment of test to student is random within the room. Thus, in a two-period testing program, mean scores were obtained on four different tests, and individual scores on each test were obtained from one-half the total number of students.

The IQ test (Henmon and Nelson, 1960) and Learning Environment Inventory were given in December of the academic year of the experiment using randomized data collection. The Student Questionnaire was administered in March to all students, and all other criterion instruments were administered using the randomized technique in May.

The unit of analysis used was the posttest teacher-mean, that is, the average score on a test of all physics students taught by a single teacher. The groups were assigned to three levels of mean IQ: Low—less than 112.1; Middle—from 112.2 to 119.3; and High—more than 119.4. Treating IQ as a factor in the design permitted testing for course and IQ interactions. This was particularly of interest in the current evaluation because of the Project's goal of appealing to a broader spectrum of student abilities. Leveling on IQ, of course, also increased the precision of the experiment by reducing the within-cell variance.

Because the cell sizes for corresponding levels are unequal as indicated in Table 1, a nonorthogonal analysis of variance solution was used (Bock and Haggard, 1968). Because main effects are confounded in the statistical analysis of nonorthogonal designs, the order is important when examining the main effects. In this study, the effect of IQ was examined first, followed by the course effect, and then the interaction. This provides an unconfounded test of the course effect which was of major interest in this evaluation study.

The null hypotheses tested were that the mean differences between treatment groups (Project Physics versus Other Physics) equal zero for all variables simultaneously within each cluster. An F-ratio was computed for the multivariate test of the equality of the mean vectors. If the F-value exceeded the two-tailed .10 level of probability, which suggested an overall difference between the two groups, then the univariate F-tests of differences in means for each variable were examined to determine the direction and relative sizes of the course effects on each of the dependent variables. The two-tailed level of significance for the univariate F-tests was also set at the .10 level. The two-tailed .10 level was adopted for three reasons: multivariate and univariate tests were employed; a two-tailed .10 is equivalent to a one-tailed .05 test which could validly be used because the directions were hypothesized; and because the results were to be used for applied decision making.

RESULTS AND DISCUSSION

Table 3 reveals that of the six clusters of criteria, three were significant for the IQ factor, four were significant for the course effect, and one interaction was significant. The results for each significant cluster are discussed successively.

The lack of significant differences in the cognitive criteria may be disappointing to some because of the Project Physics goal to increase science-process understanding. Perhaps the course developers can find some solace in the fact that other objectives were achieved (see below) without a resulting loss in student physics achievement and understanding.

Because the multivariate test of the affective criteria was significant, the univariate F-tests were examined. Only the Course Satisfaction scale was found to be significant on this test (see Figure 1 for all significant univariate F-tests for the course effects). The standardized course contrast (obtained by dividing the least-square estimates of course differences by the within-group standard deviation) is shown in

TABLE 2

Test Information and Reliabilities

Instrument	Reliability
Cognitive	
Test on Understanding Science. Assesses students' understanding of the scientific enterprise, scientists, and the methods and aims of science (Cooley & Klopfer, 1961).	.76[a]
Physics Achievement Test. Locally-developed test of general topics in physics. Derived from the six unit achievement tests developed for Project Physics (Winter & Welch, 1967).	.77[a]
Science Process Inventory. Assesses students' knowledge of the activities, assumptions, products, and ethics of science (Welch & Pella, 1967-68).	.86[a]
Course Grade. Final grade received by students.[d]	
Affective	
Physical Science Interest Measure. One of six subject matter interest measures (Halpern, 1965).	.93[a]
Pupil Activity Inventory. An operational measure of science interests (Walberg, 1967). Derived from Reed Science Activity Inventory (Cooley & Reed, 1961).	.90[b]
Course Satisfaction. Assesses students' satisfaction in course. Derived from cluster analysis of twenty items from Student Questionnaire (Welch, 1969).	.80[c]

Learning Environment	
Fourteen cluster scores from the Learning Environment Inventory. Used as a substitute for direct classroom observation in determining the social climate of the classroom (Walberg and Anderson, 1968).	.58-.86[b]
Course Reaction	
Twenty item scores from Student Questionnaire. Score obtained by computing percentage of students agreeing with each of twenty statements about physics courses in general.[d]	
Semantic Differential	
Fourteen cluster scores from the Semantic Differential Test. Assesses students' attitudes relating to physics (Geis, 1969).	.60-.86[c]
Physics Perception	
Fifteen item scores from Special Semantic Differential. Forced choice semantic differential instrument. Assesses students' perceptions of physics.[d]	

[a] Kuder-Richardson Formula 20 reliability

[b] Cronbach Alpha reliability

[c] Stepped-up mean item intercorrelation (equivalent to Cronbach Alpha reliability)

[d] Reliability not computable (single item scores used)

TABLE 3

Multivariate F-Values for Analysis of Variance

Cluster	(df) and Multivariate F-Value IQ	Course	Interaction
Cognitive	(8/88) 4.32***	(4/44) .82	(8/88) .48
Affective	(6/90) 1.18	(3/45) 3.36**	(6/90) 1.54
Learning Environment	(28/68) 1.13	(14/34) 1.74*	(28/68) .65
Course Reaction	(40/56) 2.06***	(20/28) 7.15***	(40/56) .79
Semantic Differential	(28/68) 2.50***	(14/34) .69	(28/68) 1.77**
Physics Perception	(30/66) 1.40	(15/33) 4.74***	(30/66) 1.52

NOTE: F-values significant at the .10, .05, and .01 levels are indicated, respectively, with one, two, and three asterisks.

Figure 1. Project Physics students scored nearly one standard deviation higher than the other students on this criterion, and the contrast is highly significant.

The Learning Environment scales were significant on the multivariate test, and three scales were significant on the univariate tests. Project Physics students scored higher on (perceived their classes as having more) Diversity, while students in other courses saw their classes as having more Favoritism and Difficulty.

On the Course Reaction items, Project Physics students found a historical approach interesting, thought physics could be understood without an extensive mathematics background, found their test enjoyable to read, hoped the course would not change, and finished the course during the year in contrast to students in other courses. Students in other courses more often found physics to be one of the most difficult courses they had taken in high school, and concluded that physics has to be difficult.

The course effect on the Semantic Differential scales was not significant on the multivariate test; however, it interacted significantly with IQ on the multivariate test and on nine univariate tests: Doing Laboratory Work as Valuable and Interesting; Learning about Science as Interesting; Physics as Interesting, Valuable, Safe, Orderly, and Understandable; and Universe as Interesting. Plots of the significant interactions revealed that, for the low IQ group (teacher-mean IQ less than 112.3; see Table 1), students in other courses responded significantly more favorably to these scales. For the middle IQ group, the reverse holds: Project Physics students responded more favorably. There were no significant differences between the two groups in the high IQ classification.

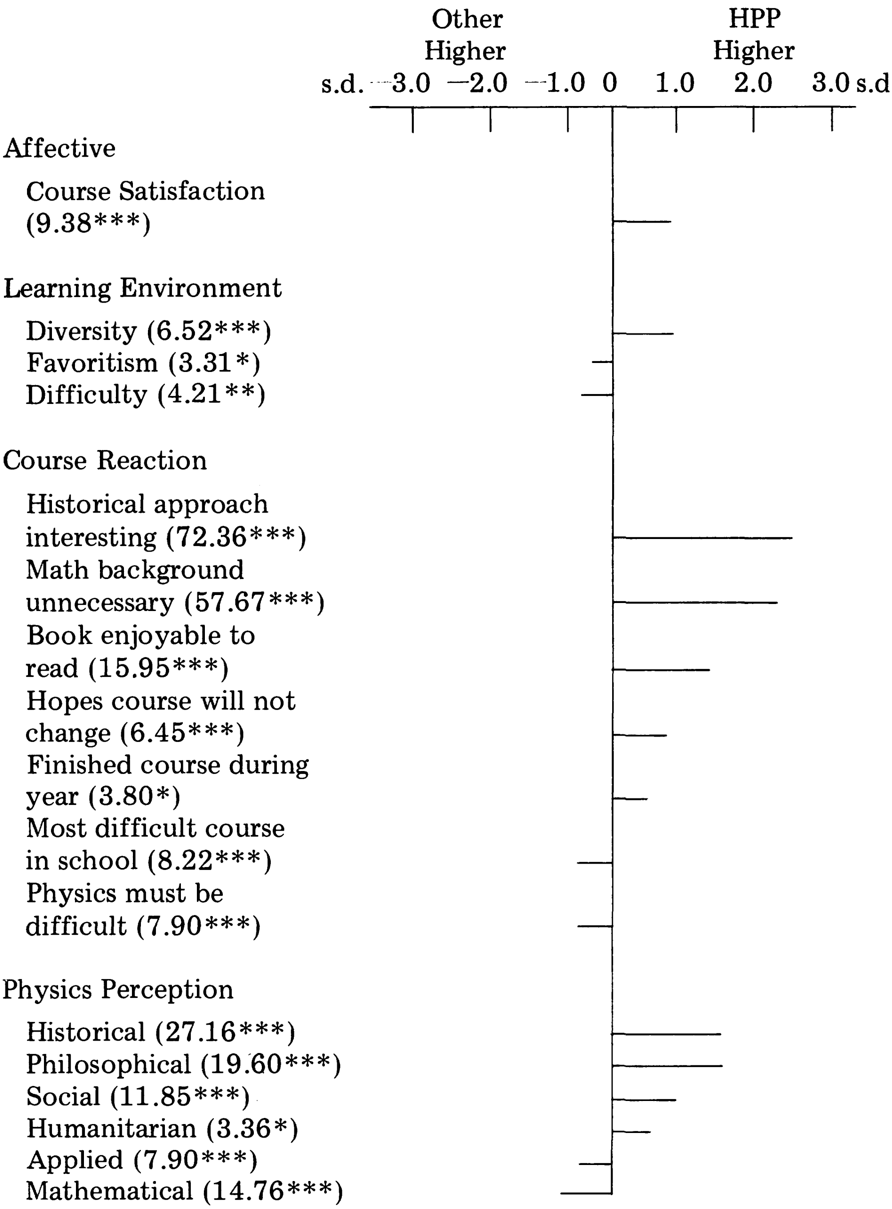

FIGURE 1
Significant standardized course contrasts and f-ratios

For the Physics Perception scales, the multivariate and six univariate tests were significant. Project Physics students rated the concept Physics as more Historical, Philosophical, Social, and Humanitarian and less Mathematical and Applied than did students in other courses.

CONCLUSIONS

From the experimental part of the evaluation, it appears that Project Physics has reached several main goals which were established for it. Students exposed to the course perform as well as students in other courses on cognitive measures. In keeping with the humanistic, affective, and multimedia elements of the course, they perceived their classroom environments as more diverse and egalitarian and less difficult. They found their textbook more enjoyable, a historical approach more interesting, and physics less difficult. Reflecting the way the subject was to be portrayed in the course, they saw physics as more historical, philosophical, and humanitarian and less mathematical. Finally, the course does seem to have a special appeal to the middle-range IQ group, 112 to 119, which has increasingly tended to elect not to take physics in high school in the last decade.

With respect to national curriculum experiments, we concluded that they are feasible and necessary. Not only do they meet canons of broader inference, but they also are more convincing to evaluation consumers. Our liberally-estimated extra costs of a national, over a regional experiment, are $1,000 for random sampling, long distance calls, and mailing and $8,500 in transportation expenses for bringing the 57 teachers to Cambridge, Massachusetts from various parts of the country. Compared to the developmental costs of a government-sponsored high school science course, which often runs into the millions of dollars, the benefits of national experiments (and other evaluation activities) seem worthwhile.

REFERENCES

Bock, R. Darrell and Haggard, Ernest A. The Use of Multivariate Analysis of Variance in Behavioral Research. In Dean Kay Whitla (Ed.) *Handbook of Measurement and Assessment in Behavioral Sciences.* Boston: Addison-Wesley Publishing Co., 1968.

Cooley, William W. and Klopfer, Leo E. *Test on Understanding Science.* Princeton, N. J.: Educational Testing Service, 1961.

Cooley, William W. and Reed, H. B. The Measurement of Science Interests: An Operational and Multidimensional Approach. *Science Education,* 1961, 45, 320-326.

Geis, Fred, Jr. *The Semantic Differential Technique as a Means of Evaluating Changes in "Affect."* Doctor's thesis. Cambridge, Mass.: Graduate School of Education, Harvard Univ., 1968.

Halpern, Gerald. *Scale Properties of the Interest Index.* Princeton, N. J.: Educational Testing Service, 1965.

Henmon, V. A. C. and Nelson, M. J. *The Henmon-Nelson Tests of Mental Ability Manual.* Boston: Houghton Mifflin Company, 1960.

Walberg, Herbert J. Dimensions of Scientific Interests in Boys and Girls Studying Physics. *Science Education,* 1967, 51, 111-116.

Walberg, Herbert J. and Anderson, Gary J. The Achievement-Creativity Dimension and Classroom Climate. *Journal of Creative Behavior,* 1968, 2, 281-291.

Walberg, Herbert J. and Welch, Wayne W. A New Use of Randomization in Experimental Curriculum Evaluation. *School Review,* 1967, 75, 369-377.

Welch, Wayne W. Curriculum Evaluation. *Review of Educational Research,* 1969, 39, No. 4, 429-443.

Welch, Wayne W. A Review of the Research and Evaluation Program of Harvard Project Physics. ERIC Center for Science and Mathematics, Ohio State University, Columbus, Ohio (mimeo), 1971.

Welch, Wayne W. Some Characteristics of High School Physics Students: Circa 1968. *Journal of Research in Science Teaching,* September, 1969.

Welch, Wayne W. and Pella, Milton O. The Development of an Instrument for Inventorying Knowledge of the Processes of Science. *Journal of Research in Science Teaching,* 1967-68, 5, 64-68.

Winter, Stephen S. and Welch, Wayne W. Achievement Testing Program of Project Physics. *The Physics Teacher,* 1967, 5, 229-231.

(*Received November, 1971*)

(*Revised January, 1972*)

AUTHORS

WELCH, WAYNE W. *Address:* College of Education, University of Minnesota, Minneapolis, Minnesota 55455 *Title:* Associate Professor and Assistant Dean *Degrees:* B.S., Wisconsin State University; M.S. (Education), University of Pennsylvania; M.S. (Physics), Purdue University, Ph. D. University of Wisconsin *Specialization:* Curriculum evaluation and measurement; research in science education.

WALBERG, HERBERT J. *Address:* College of Education, University of Illinois, Box 4348, Chicago, Illinois 60608 *Title:* Professor *Degrees:* B.E., Chicago State College; M.E., University of Illinois; Ph.D. University of Chicago *Specialization:* Social psychology and individual differences; measurement, evaluation, and statistical analysis.