Avoiding Type III Errors in Health Education Program Evaluations: A Case Study

Charles E. Basch, Elena M. Sliepcevich, Robert S. Gold, David F. Duncan and Lloyd J. Kolbe

Health Educ Behav 1985; 12; 315
DOI: 10.1177/109019818501200311

Avoiding Type III Errors in Health Education Program Evaluations: A Case Study

Charles E. Basch, PhD
Elena M. Sliepcevich, DPE
Robert S. Gold, PhD, DrPH
David F. Duncan, DrPH
Lloyd J. Kolbe, PhD

Monitoring the implementation of a program being evaluated can improve the interpretability of data collected and help evaluators to avoid committing a Type III error: evaluating a program that has not been adquately implemented. This article describes an evaluation that analyzed the implementation of a school health education curriculum, assessed cognitive learning outcomes attributable to the curriculum, and examined the relationship between classroom implementation and changes in students' knowledge. Five fifth-grade classes (n = 101) participated in the curriculum, and five classes (n = 84) served as a comparison group. Data collection procedures involved a pretest and posttest of all students' health-related knowledge, daily monitoring of classroom implementation by the five teachers participating, and questionnaires completed by principals and teachers. Analysis methods included descriptive statistics, parametric and non-parametric tests of significance, and qualitative assessment procedures. Results indicated that the curriculum had a positive effect on learning in students; curriculum implementation varied considerably among the five classes participating; teaching/learning activities that were most and least likely to be implemented could be identified and described; both teachers and principals perceived the program favorably; some health instruction was occurring in the comparison classes, so it was not appropriate to consider them as pure controls; and no statistically significant relationship between curriculum implementation and cognitive outcomes was observed. This study provides evidence of the need for and value of measuring implementation of programs being evaluated. Implications for developing implementation measures and the role of formative evaluation in health education practice are considered.

Charles E. Basch is Assistant Professor of Education, Department of Health Education, Teachers College, Columbia University, New York, New York 10027. Elena M. Sliepcevich is Professor, Department of Health Education and School of Medicine, Southern Illinois University, Carbondale, Illinois. Robert S. Gold is Director, School Health Initiative, Office of Disease Prevention and Health Promotion, Department of Health and Human Services, Washington, DC. David F. Duncan is Professor, Department of Health Education, Southern Illinois University, Carbondale, Illinois. Lloyd J. Kolbe is Associate Professor of Behavioral Sciences and Associate Director, Center for Health Promotion Research and Development, University of Texas Health Sciences Center, Houston, Texas.

The contributions of the following individuals were invaluable to this study: Shirly Menendez and Nancy Farr at the Illinois State Facilitator Center, and the teachers and school administrators who participated.

Address reprint requests to Charles E. Basch, PhD, Department of Health Education, Box 114, Teachers College, Columbia University, New York, NY 10027.

Health Education Quarterly,

Vol. 12(4): 315–331 (Winter 1985)
0195-8402/85/040315-17$04.00

INTRODUCTION

When field conditions preclude the use of randomization, monitoring curricular implementation can improve the interpretability of data generated by quasiexperimental designs. Measuring the degree of program implementation and including these measures in data analysis also enhance considerably the validity of randomized field experiments designed to assess program outcomes.[1-3] Failure to include implementation measures in such designs may result in what Scanlon and his colleagues[4] have termed a Type III error. Unlike a Type I error (rejecting a true null hypothesis) or a Type II error (failing to reject a false null hypothesis), a Type III error is defined by the following:

> To spend money to evaluate the effectiveness of a program
> when the program is not measured as implemented,
> when the program has not been implemented, or
> when there is no testable relationship between the program activity carried out and the program objectives being measured is to spend money to measure something that does not exist (p. 264).[4]

After a brief review of literature that considers the functions of and need for research on implementation processes, this article describes an evaluation that was designed to analyze the implementation of a school health education curriculum, assess cognitive learning outcomes attributable to the curriculum, and investigate relationships between the degree of classroom implementation and changes in student knowledge.

Barriers that have impeded evaluation of school health education curricula are similar to those that have impeded evaluation of health education programs in community and hospital settings and have been discussed frequently in the literature.[5-10] Attempts to produce evidence of the value of curricula largely have focused on summative assessment of outcomes such as changes in knowledge. There have been few formative evaluations conducted to monitor program development and use in order to document program implementation and improve effectiveness.[11] To help ensure the validity of results, summative evaluation research has generally relied on experimental or quasi-experimental designs and their appropriate statistical procedures as described initially by Campbell and Stanley[12] and more recently by Cook and Campbell.[1] These approaches have contributed to the making of objective observations and valid inferences. However, by only studying outcomes (such as changes in IQ scores, attitudes, or mortality rates) without concomitantly studying the extent to which program components designed to modify such outcomes actually were implemented, we cannot unravel how or why such outcomes are influenced. Although traditional summative evaluation research may be able to detect significant differences in outcomes, such approaches do not provide information useful for explaining why differences did or did not occur or about the effectiveness of various components of programs.

This information about program outcomes provides decision makers with very limited information about the program being evaluated, and decisions based on outcome data alone have sometimes been inappropriate.[13-16] Failure to analyze implementation processes has served as a major hindrance both to policy analysis and studies of social policy[17,18] and to evaluation of educational programs.[19-22] Very often, the type of information that is needed by planners and policy makers is related to program implementation.[22] At the minimum, such an assessment will determine if the program does exist. Guttentag and Struening[13] have noted that program evaluators have, in

some instances, falsely assumed that programs did exist when in reality they do not and that this could not occur if program evaluators were familiar with the content and context of the programs they were evaluating. Familiarity with and description of the object (e.g., program, project, material) being evaluated and the context in which the object is used is a quality assurance standard for evaluating educational programs listed by the Joint Commission on Standards for Educational Evaluation.[23] However, determining whether or not a program does exist may not always be a simple matter, because many programs are loosely defined programs, that lack criteria to assess the extent to which planned components are actually implemented and because programs change from one setting to another and over time.

Studying the implementation of a program may be undertaken for at least four reasons: (1) improving understanding about the best techniques for promoting implementation, long-term maintenance, and further program dissemination; (2) providing accountability to agencies that allocated resources; (3) enhancing the validity of summative evaluations; and (4) learning how to modify programs and policies in order to improve their effectiveness and application.[24] Studying implementation can help facilitate strategies to promote implementation by identifying possible problems, solutions, and circumstances under which implementation is likely to succeed,[25] assessing the capabilities for and feasibility of implementation,[26] and improving our understanding of the reasons why an educational innovation fails to be adopted and implemented.[27] Accountability may be provided by monitoring services and producing documentation (e.g., records) that verify implementation.[15] Summative evaluations can be enhanced in a number of ways. For example, monitoring implementation enables the evaluator to determine if negative findings may be attributable to faulty implementation. Goodlad[28] recently noted that many educational programs that were thought to have failed really had not; they really had never been implemented. Also, assessing the nature and extent of instruction in classes that serve as a comparison group can help determine if diffusion of treatment, compensatory rivalry or resentful demoralization, or threats to construct validity of putative causes and effects[1] have occurred. Finally, studying implementation can provide valuable information serving research and development functions and management decisions.[29,30] To conduct implementation analysis often requires that program components be described and the purposes of each in producing desired effects be specified.[17] Measurement of mediating processes can help to test theoretical models from which the programs were designed and to clarify why programs work to the extent that they did and why they failed to be more efficacious.

Since 1970, the growing concern about failure to assess implementation when evaluating programs[2–4,15,16,19,22,31–35] has been accompanied by an increased number of studies of curricular implementation processes.[20,25,27,34,36–50] The number of studies in areas other than education is also expanding.[51] A considerable amount of time, money, and effort has been devoted to the development of model school health education curricula, and sponsors of these programs are attempting to disseminate materials, but relatively little effort or resources have been devoted to determining the extent to which and the reasons why innovations are actually used.

Kreuter, Christenson, and Davis[52] recently challenged school health education researchers to respond to two basic questions (p. 32):

1. What is the best way to get scientifically credible health information and appropriate health skills to children in schools?

2. What must be done to increase the number of elementary and secondary schools which offer adequate health information and appropriate health skills to children and adolescents?[52]

Thus, descriptive and analytical studies are needed to determine which types of programs and program elements are most and least likely to be used in various settings and to determine patterns of utilization. Such research might facilitate the design of curricula and in-service training programs and assist school and community decision makers in selecting programs that will be used.[53]

In research of this nature, the degree (or fidelity) of implementation of the program is the dependent variable. The remainder of this article describes methods that were applied to study the implementation of a health education program. The study is intended to serve as an illustration of ways in which curricular implementation may be defined and measured in a summative program evaluation and how relationships between implementation and learning outcomes may be analyzed.

EFFORTS TO AVOID A TYPE III ERROR WHILE EVALUATING A SCHOOL HEALTH EDUCATION PROGRAM

Since its inception in 1965, the School Health Curriculum Project (SHCP) has expanded to include curricular units for fourth, fifth, sixth, and seventh grades; in 1980 it was reportedly being implemented within approximately 1000 schools in 34 states by more than 4000 teachers. Description of the project and its evolution are described elsewhere.[54-61] Numerous summative evaluations, diverse in scope, purpose, and methods, have been conducted to assess the effects of the SHCP and have been reviewed elsewhere.[8,62] Almost all of these studies, and most other evaluations of health education programs, neglect to assess actual implementation of the curriculum.

Research Objectives

The specific research objectives addressed by this study were to (1) assess impact of the SHCP unit on student knowledge; (2) determine whether the SHCP curriculum activities were implemented as planned, modified or omitted by the five participating teachers; (3) analyze differences in the curriculum activities found least likely and most likely to be implemented as planned; (4) assess participating teachers' perceptions about health, health education, and the SHCP; (5) assess acceptability of the SHCP to participating principals; (6) describe the nature and extent of health instruction in the five classes used as a comparison group; and, the main objective, (7) estimate the magnitude and statistical significance of relationships between fidelity of program implementation and student learning outcomes. Findings and conclusions are presented in subsequent sections for each of these research objectives in the order in which they are listed.

Operational Definitions

Fidelity of implementation was defined as "the extent to which actual use of an innovation corresponds to intended or planned use" (p. 340).[27] Cognitive outcomes were defined as students' performance on a 51-item multiple-choice cognitive test developed specifically for evaluating the fifth-grade unit of the SHCP.

Sample

This study was conducted in southwestern Illinois, in an essentially rural area sparsely populated by white middle-class people. Five classes (101 students) in five separate schools participated in the SHCP, and five classes (84 students) in three schools served as a comparison group.

Instruments and Data Collection

A 51-item cognitive instrument specifically designed for evaluation of the fifth-grade unit of the SHCP by Heit, Sheer, Jurs, and Kolbe was used. Internal consistency was reported to be 0.72 as measured by Kuder–Richardson Formula 20 (KR 20). The cognitive instrument was improved and was administered by regular homeroom teachers immediately before and again after the 16-week implementation period. An effort was made to minimize the differences in administration procedures by requesting that the 10 participating teachers adhere to standardized administration procedures. Internal consistency of the cognitive measure was calculated for each group at pretest and posttest using the KR 20. Internal consistency for the SHCP group was 0.68 at pretest and 0.82 at posttest. For the comparison group, the coefficients were 0.46 at pretest and 0.65 at posttest.

Four instruments used to measure implementation were developed by Kolbe, Kreuter, Iverson, and Christenson and were reviewed by a variety of experts in health education and curricular evaluation. Theoretical foundations first proposed by Fullen and Pomfret[27] and Hall and Loucks[20] were used for development of four instruments. First, a *Principal's Questionnaire* measured administrators' perceptions about (1) the staff's satisfaction with training, (2) problems encountered relevant to implementation, (3) cooperative arrangements among personnel, and (4) plans for the project in the future. This instrument was mailed to principals in schools participating in the SHCP at the end of the implementation period. Second, a *Teacher Activities Analysis Report* (TAAR) was used to measure fidelity of implementation of the fifth-grade unit. The TAAR solicited information from teachers about whether each of the 282 SHCP teaching/learning activities listed in the fifth grade unit was implemented as planned, modified, or omitted. Space was provided for a brief description of any modifications that were made. The instrument was completed by the five SHCP teachers and returned by mail twice a month throughout the implementation period. Concurrent validity of the TAAR was estimated by having an observer note which activities were implemented on 10 randomly selected days (two for each classroom). The agreement between the observer's notes and the respective teacher's responses to the TAAR ranged from 25% to 100% with a mean of 81.5%. Third, a *Materials Analysis Report* (MAR) initially

was used to collect information about whether or not each of the 160 curriculum materials in the fifth-grade unit was used. Similar to the TAAR, this instrument queried teachers about whether they implemented as planned, modified, or omitted recommended materials. Results from this instrument are not reported because materials utilization are reflected in the activities implemented. Fourth, a *Teacher Acceptability Report* (TAAR) solicited teachers' perceptions about health, health education, and the SHCP and its stated objectives. This instrument was distributed by mail to teachers participating in the SHCP at the end of the implementation period.

The *Data Sheet on Health Instruction*, a fifth implementation analysis instrument, was developed by the researchers to compare the extent and nature of health instruction in experimental and comparison classes. The instrument solicited information from teachers about the amount of classroom time devoted to planned health instruction, the frequency with which health instruction was integrated into other subject areas, and the frequency with which teachers covered selected health topics and used specific teaching/learning activities and materials. This instrument was mailed to all 10 teachers at the end of the implementation period.

Analysis Procedures and Results

Means, standard deviations, and skewness of pretest and posttest cognitive scores for the experimental and comparison students are shown in Table 1. Differences in pretest means were determined by applying a one-way analysis of variance (ANOVA), and differences in posttest means were assessed by an analysis of covariance (ANCOVA) with pretest as covariate. In the ANOVA, the student was used as the unit of analysis to guard against committing a Type II error. In the ANCOVA, the class mean was used as the unit of analysis to guard against committing a Type I error. The posttest mean of the SHCP group was 7.52 points greater than the mean of the comparison group. Based on the ANOVA, there was no significant ($p < 0.05$) difference in pretest scores between groups. Posttest standard deviations and skewness were similar for both groups. The F ratio for the ANCOVA was 20.942, which was significant beyond the 0.005 level. The SHCP unit seemed to achieve a statistically significant impact on student's knowledge, though given the magnitude of the posttest difference (7.52) and the ratio of the posttest difference in means (7.52) to the pooled standard deviation of the posttest scores (7.69), the practical significance of this finding is debatable.

Table 1. Cognitive Test Means, Standard Deviations, and Skewness by Groups at Pretest and Posttest for Students in Ten Fifth-Grade Classes in Two Southern Illinois Counties

	Cognitive Pretest			Cognitive Posttest		
	Mean	Standard Deviation	Skewness	Mean	Standard Deviation	Skewness
SHCP Group	18.24	5.62	−0.027	25.28	7.60	−0.097
Comparison Group	16.88	4.38	0.26	17.76	5.49	−0.32
Both Groups	17.63	5.13	0.15	21.89	7.69	0.22

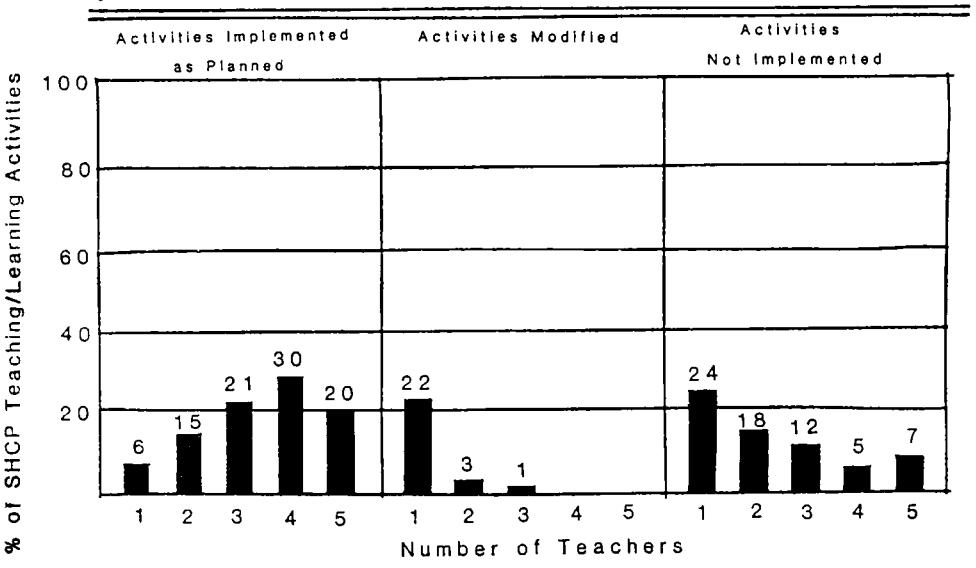
Table 2. Percentage of SHCP Activities Implemented as Planned (IAP) Modified (M) and Not Implemented as Planned (NI) in Each Phase of the Curriculum

	Mean	Median	Minimum (%)	Maximum (%)
(1) Introduction				
IAP	54.8	59	27	71
M	14.6	10	10	24
NI	30.2	29	19	54
(2) Phase I				
IAP	73.8	79	48	86
M	13.2	7	0	42
NI	13.0	14	7	20
(3) Phase II				
IAP	63.8	70	26	79
M	5.4	3	0	19
NI	30.2	24	19	55
(4) Phase III				
IAP	67.0	71	30	84
M	5.8	3	0	21
NI	27.2	26	11	49
(5) Phase IV				
IAP	66.2	66	28	92
M	2.8	2	0	8
NI	31.4	34	8	70
(6) Phase V				
IAP	55.0	75	17	83
M	0.6	0	0	3
NI	44.4	25	17	80

Percentages of curricular activities implemented as planned, modified, or omitted by the five fifth-grade SHCP teachers were calculated for each of the six curricular phases that comprise the fifth-grade unit. In Table 2, fidelity data are shown for each of the six curriculum phases. Overall, Phase I material tended to be implemented as planned to a greater degree than material in the other curriculum phases. This same phase also tended to have the fewest activities that were not implemented at all by these teachers. In other phases, participating teachers implemented from 54.8% to 67.0% of the activities as planned, at the same time disregarding 13.0% to 44.4% of the proposed curricular activities. The mean percentage of activities implemented as planned across the six curriculum phases was about 63%. Based on the average percentage of activities implemented as planned by each teacher, there was a relatively high degree of implementation fidelity (over 75%) by two of the teachers, a moderately high degree of implementation fidelity (50–70%) by two teachers, and a low level of implementation fidelity (less than 30%) by one teacher.

Another analysis assessed the extent to which each individual activity was implemented. The proportion of activities implemented as planned, modified, or omitted by the five teachers participating in the SHCP are shown in Figure 1. For each activity, the number of teachers who reported implementing the activity as planned, modifying it, or omitting the activity was determined. Then, the proportion of activities implemented as planned, modified, and omitted by various teachers was calculated. This analysis revealed that of the 282 activities: 66% (n = 185) were not implemented as

Figure 1. Proportion of SHCP activities implemented as planned, modified, or not implemented by five teachers in two southern Illinois counties.



planned by one or more teachers. Thus, 34% were implemented as planned or modified by all of the teachers, 8% (n = 24) were not implemented as planned by any of the teachers, and 74% (n = 207) were not modified by any of the teachers.

Those activities that were least and most likely to be implemented were identified and reviewed to assess commonalities. A total of 33 activities were not implemented or modified by more than three teachers. Review of these activities revealed that slightly more than one-third involved a resource person. The remaining activities were diverse and did not share a common trait. 20% (or 56) of the activities were reported to be implemented as planned by all five teachers. These highly implemented activities had two characteristics in common: (1) they involved children in viewing an audio–visual aid (usually a film) and discussing it, and (2) they involved children in highly participatory "learning centers."

Teacher Acceptability Report (TAR) ratings about health, health education, and the SHCP and its stated objectives revealed favorable responses. The response format for the first component was a five-point Likert-type scale, with 1 being *agree strongly* and 5 being *disagree strongly*. All five teachers agreed strongly that a person's health could be influenced significantly by personal behavior. Three of the five agreed strongly and two agreed that acquiring knowledge about health can help people to make decisions that will affect their health. All five of the teachers agreed strongly that schools should provide education about health for children. Three of the five agreed strongly and two agreed that the SHCP is an effective method for teaching children about health. Three of the five agreed strongly, one agreed and one disagreed that as a result of participating in the SHCP they felt prepared to teach children about health.

These five teachers were also asked to rate the importance of SHCP objectives on a five-point Likert-type scale ranging from 1 (*extremely important*) to 5 (*extremely unimportant*) and similarly to rate the importance of the curricular activities designed for achieving the SHCP objectives. Although all objectives received high ratings of importance, the objective rated most important was "To arouse curiosity and wonder

in children." This was followed closely by "To motivate children to learn," and "To provide learning situations in which children interact with their environment." Objectives rated as least important were "To stimulate success experiences for children" and "To develop in children a scientific approach to problem solving." Ratings of the importance of teaching/learning activities for accomplishing the stated objectives were similar. Teaching/learning activities were rated most important for the objective "To teach children to question 'Why?'" "To challenge ideas," "To read critically," and "To evaluate information."

Results from the principal questionnaire were analyzed using descriptive statistics and qualitative reporting for open-ended items and showed that principals were generally satisfied with the SHCP. The principals' level of satisfaction with the SHCP project were scaled on a Likert-type format ranging from 1 (*not at all satisfied*) to 5 (*extremely satisfied*). The mean rating for workshop training was 4.2 with a range from 4—5. Concerning the SHCP materials, the mean rating was 4.2 with a range from 3—5. The SHCP activities received a mean rating of 4.4 with a range from 4—5. The mean rating for evaluation methods was 3.6 with a range from 3—4. The mean rating for principals' general overall feeling about the project was 4.4 with a range from 4—5. Principals' impressions of students' satisfaction with the project resulted in a mean rating of 4.4 with a range from 4—5, and impressions regarding parents' satisfaction with the project resulted in a mean rating of 4.2 and a range from 4—5. In response to the question of whether or not principals anticipated continuing the project in the future, two indicated that they were "uncertain," two responded "definitely," and one responded "probably."

When asked to indicate what, if any, were major problems their participating teacher experienced in implementing the SHCP in the classroom, the principals reported a variety of problems. Problems cited could be grouped into five categories: sharing films among the five schools, amount of time required for curriculum implementation, lack of coordination among the five schools, nature of the curriculum, and snow. Sharing films among the five schools was cited as a problem by four of the five principals.

The TAAR and TAR were developed from the activity and materials lists of the SHCP and therefore were not applicable to teachers in the comparison schools. The *Data Sheet on Health Instruction* was intended to provide information about the extent and nature of health instruction practices of teachers in the comparison group as compared to teachers in the SHCP group. The first item on the *Data Sheet on Health Instruction* concerned how many class periods per week were devoted to planned health instruction. Planned health instruction was defined as teaching of specific health units at scheduled times reserved for such instruction. The second item on the *Data Sheet on Health Instruction* concerned how many minutes per class period were devoted to planned health instruction. The response format included six ordinally scaled response categories (e.g., 0 to 9 minutes, 10 to 19 minutes) plus a category for an irregular length of class period. In order to determine the amount of time per week devoted to planned health instruction, teacher's responses to items 1 and 2 were combined. Combined responses for each of the 10 teachers are shown in Figure 2 (excluding one of the comparison teachers because of incomplete reporting). On the basis of midpoints of the minutes-per-class-period scale, the mean number of minutes per week devoted to planned health instruction was 208 for the SHCP group, whereas the mean number of minutes per week devoted to planned health instruction for the comparison group was 72.5. Although it appears that the students participating in the SHCP received substantially more health instruction, this analysis also revealed that it would be inappropriate to consider the students not participating in the SHCP as a pure control group.

Figure 2. Amount of time devoted to planned health instruction in the ten health classes (SHCP and comparison groups).

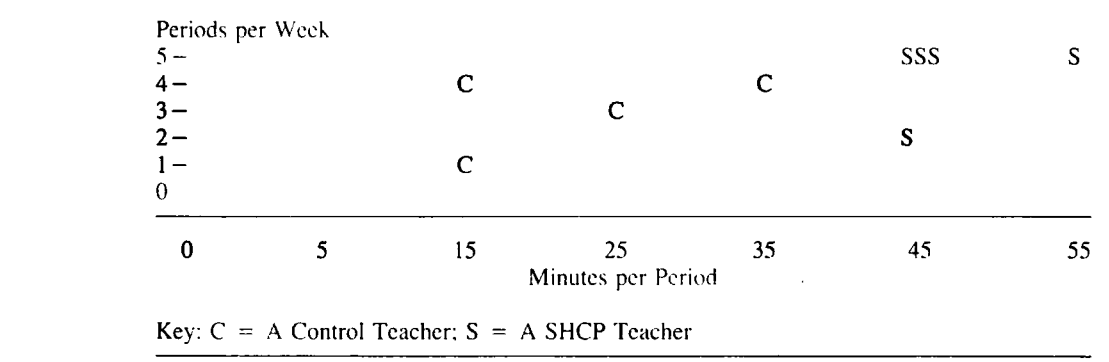


Table 3. Class Periods (Planned or Informal) Devoted to Health Instruction on Selected Topics and Informal Health Instruction under Subject Areas as Reported by Five SHCP and Five Comparison Teachers

	SHCP Teachers					Comparison Teachers				
	Daily	Weekly	Monthly	Never	No Response	Daily	Weekly	Monthly	Never	No Response
Topics										
Alcohol Education	0	1	3	1	0	0	1	2	1	1
Consumer Health Education	0	1	2	1	1	0	0	2	1	2
Cigarette Smoking Prevention	0	4	1	0	0	0	1	2	1	1
Dental Health Education	0	0	2	3	0	0	0	4	0	1
Drug Use and Abuse	0	0	2	3	0	0	0	2	1	2
First Aid	1	2	0	2	0	0	0	3	1	1
Human Ecology and Health	0	1	1	3	0	0	0	3	1	1
Human Growth and Development	0	3	1	1	0	0	0	4	0	1
Nutrition	0	1	2	1	1	0	0	4	0	1
Personal Health	1	1	2	0	1	0	0	4	0	1
Prevention and Control of Disease	0	3	1	1	0	0	0	4	0	1
Safety Education	0	2	1	2	0	0	0	4	0	1
Subject areas										
Art	0	1	1	2	1	0	0	1	3	1
Mathematics	0	1	1	2	1	0	0	1	3	1
Physical Education	0	1	1	2	1	0	0	2	2	1
Reading	1	1	0	2	1	0	0	3	1	1
Science	1	2	0	2	0	0	2	2	0	1
Social Studies	0	2	0	2	1	0	0	2	2	1
Spelling	1	0	1	2	1	0	0	1	3	1
Writing	0	0	1	2	2	0	0	1	3	1
Other (Religion)	0	0	0	0	5	1	0	0	0	4

Three remaining implementation phenomena were measured by the *Data Sheet on Health Instruction*. First, the amount of time devoted to informal health instruction was measured. Informal health instruction was defined as teaching health with another subject (correlation, integration) or when a teachable moment arises (incidental). Second, the amount of time (planned or informal) devoted to health instruction on selected topics was measured. Third, the frequency with which various materials and activities were used for planned health instruction was measured. These data are summarized in Tables 3 and 4. These data support the finding above that there was some health

Table 4. Frequency with Which Materials and Activities were Used for Planned Health Instruction as Reported by Five SHCP and Five Comparison Teachers

	SHCP Teachers					Comparison Teachers				
	Daily	Weekly	Monthly	Never	No Response	Daily	Weekly	Monthly	Never	No Response
Materials										
Booklets/Leaflets/ Pamphlets	1	4	0	0	0	0	0	3	1	1
Books	1	3	1	0	0	2	1	1	0	1
Cassettes	0	4	1	0	0	0	0	1	3	1
Films	0	4	1	0	0	0	0	2	2	1
Filmstrips	0	5	0	0	0	0	0	2	2	1
Microscopes and Slides	0	1	3	1	0	0	0	3	1	1
Models	0	2	1	2	0	0	0	4	0	1
Photographic Slides	0	2	1	2	0	0	1	0	3	1
Posters/Charts/ Graphs	1	3	1	0	0	0	0	3	0	2
Records	0	2	1	2	0	0	0	0	3	2
Relia (Games, Simulations, Ads, etc.)	0	1	3	1	0	0	0	2	1	2
Transparencies	0	1	4	0	0	0	0	2	1	2
Other	0	0	0	0	5	0	1	1	0	3
Activities										
Exhibits	0	2	2	1	0	0	0	2	1	2
Field Trips	0	0	1	4	0	0	0	0	4	1
Guest Speakers	0	0	2	3	0	0	0	1	3	1
Independent Class Study	1	0	3	0	1	0	0	2	1	2
Large Group Discussions	2	2	1	0	0	2	0	1	0	2
Lectures	1	1	3	0	0	2	0	0	1	2
Problem Solving	2	1	2	0	0	0	1	3	0	2
Small Group Activities	1	2	2	0	0	0	1	2	1	1
Team Teaching	0	0	1	4	0	0	0	0	3	2
Values Clarification	0	1	3	1	0	0	3	1	0	1
Other	0	0	0	0	5	0	0	0	0	5

Table 5. Spearman Rank-Order Correlations between Cognitive Test Score Gain and Percent of Activities Implemented as Planned, Modified, or Not Implemented in Each Phase

	Implemented as Planned		Modified		Not Implemented	
Phase	rho	*p*	rho	*p*	rho	*p*
Introduction	0.72	0.086	−0.34	0.285	−0.57	0.156
I	0.63	0.127	−0.79	0.056	0.00	0.500
II	0.63	0.127	−0.26	0.334	−0.56	0.019
III	0.63	0.127	−0.24	0.351	−0.56	0.161
IV	0.31	0.307	0.29	0.318	−0.50	0.196
V	0.67	0.109	−0.54	0.172	−0.67	0.109

education occurring in the comparison group classes. Use of multiple measures was a strength of the study. Information from the TARR about the percent of activities implemented was verified by data obtained with the *Data Sheet on Health Instruction.* Teachers with the highest levels of fidelity of implementation devoted the greatest amount of time to teaching.

Finally, data concerning the main question under study, relationships between implementation and outcomes, were examined using two analytic procedures, analysis of variance and Spearman rank-order correlations. The first approach required that each teacher be categorized as a high implementer or a low implementer for each of the curriculum phases. Teachers were divided into two groups using the median percentage of activities implemented as planned in each phase. The teacher with the median value of activities implemented as planned was excluded from analysis for that phase. An analysis of variance was conducted with group classification as the independent variable and cognitive gain for the respective class as the dependent variable. No statistically significant differences were detected.

A second analytic approach was to calculate Spearman rank-order correlations between percentage of activities implemented as planned, modified, and not implemented in each of six curriculum phases and the corresponding cognitive gain score for the respective class. Resulting coefficients and probability levels are shown in Table 5. Although there was a moderately strong (>0.60) linear relationship between the extent to which the curriculum was implemented as planned and cognitive learning outcomes for five of the six phases, these relationships were not significant at the 0.05 level. In one phase of the curriculum (the introduction phase), the relationship was significant at the 0.10 level. To determine the extent to which selection differences may have accounted for these results, relationships between fidelity of implementation for each curriculum phase and pretest scores were examined. This examination suggested that results of the cognitive gain score analysis were not correlated with initial selection differences.

Conclusions and Discussion

This section begins by presenting seven conclusions and comments relevant to each of the research objectives. The intent of this article was to stimulate interest about conceptualizing and measuring implementation of health education programs in schools. This

section therefore also considers implications for practitioners related to developing implementation measures and the role of formative evaluation in health education practice.

First, the SHCP for the fifth-grade level as implemented had a positive effect on cognitive learning in students and thus may be considered an effective curriculum for achieving this specific type of outcome. This conclusion is consistent with previous research and supports using the SHCP for promoting cognitive learning in students. The practical significance of this finding is limited because cognitive achievement was measured using a norm-referenced test. The practical implications of SHCP students achieving a posttest cognitive mean score 8 points higher than the comparison group and unclear. The need for development of meaningful criterion-referenced tests should be emphasized.[63,64]

Second, implementation of the SHCP was not consistent from one classroom to another, and many of the SHCP teaching/learning activities were not implemented by some of the teachers who participated. It seems likely that this may be true of most or all other model curricula as well. This conclusion has important implications for researchers evaluating school health education curricula because it demonstrates the importance of measuring curricular implementation. Failure to do so may result in evaluation results being based on undefined programs that are likely to be very different from what was intended or assumed.

Third, the two types of activities that were most likely to be implemented, film/discussion and learning centers, may warrant special consideration from curriculum designers. Special efforts might be made to evaluate systematically the value of alternative audio–visual materials to ensure that these teaching tools are effective. Learning centers are physical areas (which may be a section of the classroom, such as floor, tables, a counter, or a small circle of desks, or an area outside of this classroom, perhaps a cafeteria or a hallway) in which particular concepts are explored through peer interaction and use of specialized resource materials. At times, a learning center will be larger and involve the entire class; however, smaller centers appear to be most common.[60] Heit[65] identified elimination of the "directive-authoritarian" role of teachers, which is facilitated by a great many activities (such as learning centers) calling for peer interaction, which emphasize developing group responses rather than right–wrong individual answers, as one of the most notable features of the SHCP.

Fourth, participating teachers' perceptions about health, health education, and the SHCP were very favorable. If programs are not acceptable to teachers, the chances for implementation will be minimized, and as demonstrated by the results of this study, even if a program is perceived favorably by teachers, this is not sufficient to ensure a high fidelity of implementation. Other factors that are necessaary but not sufficient to facilitate program implementation include adequate training, provision of materials, time to prepare and teach, and administrative support.

Fifth, principals in schools participating in the SHCP also perceived the project favorably. As a building administrator, the principal plays a key role in decision making about which programs are adopted, implemented, and maintained. Research involving principals to help us learn about barriers to implementing school health education and solutions to those barriers is needed.

Sixth, it may not be appropriate to assume that classes serving as a control group are pure controls. This may be the case even when randomization is used; thus, monitoring levels of program activity in control groups is recommended.[24] Several comparison-group teachers reported implementing informal health instruction under

other subject areas such as physical education, reading, science, and social studies. Many of the comparison-group teachers devoted class periods, on a monthly basis, to health instruction on diverse topics; some topic areas are covered in the comparison group to a greater extent than they were covered in the SHCP group. Teachers in the comparison group reported using diverse activities and materials for planned health instruction. At the elementary grade level, there is no departmentalization. Students stay in the same room except perhaps for art, music, and physical education. Thus, teachers have flexibility with time, and more integration and combining of subjects for instruction is possible.

Seventh, no statistically significant relationships between fidelity of implementation of the SHCP and cognitive outcomes were observed. There did appear to be a consistent trend suggesting that fidelity of implementation may be positively associated with cognitive gain scores. Since none of these relationships, however, was statistically significant, the meaning of these relationships in terms of educational significance is uncertain (here the term educational significance differs from statistical significance in that the former relates to practical implications of results, whereas the latter refers to the probability that the results observed occurred by chance). Although no definitive educationally significant relationships were detected, this does not confirm that these do not exist. Inability to detect such differences may have resulted from the small number of classes in the study or the insensitivity of the measures used.

A key question to be addressed by those conducting outcome-oriented evaluations is what program is being evaluated.[13] Or, how is program implementation defined and measured? In a recent interdisciplinary comprehensive methodological review, Scheirer and Rezmovic[51] utilized five measurement criteria to judge the quality of measures of "degree of implementation": (1) use of multiple measures; (2) presence of an operational definition; (3) examination of reliability; (4) assessment of validity; and (5) use of sampling. These criteria are worthy of consideration by those constructing implementation measures. We suggest a sixth measurement criterion: acceptability. Instruments and data collection procedures must be acceptable to those from whom data will be gathered, and they must also be acceptable to those audiences that have power or influence to act on results.

Construction of implementation measures should begin with a clear description of what the innovation is, who will be affected by the implementation, what new actions staff are expected to perform, and what organizational changes are needed to support implementation. Researchers constructing measures should be familiar with procedures used previously to measure curricular implementation.[24,51,66] Instrument design must be guided by the function the data are intended to serve. Quantitative and qualitative approaches will be more or less appropriate to investigate the questions under study. The construct of program implementation, like other complex behavioral patterns, is often multidimensional, and thus qualitative and quantitative methods used in combination often may be most appropriate. If the goals of measurement are to verify that specific program components have been used or that there is a dose/response relationship between implementation and outcomes, then quantitative methods alone may be appropriate. However, if the objectives of research are to determine barriers and generate "grass roots" solutions, to adapt programs to local needs, to determine questions related to why programs are or are not implemented and how program components are causally related to components, then qualitative methods will be necessary to provide the most relevant information. Qualitative measures, such as in-depth inter-

views, videotapes, participant observations, and focus groups, can provide a richness in data that is not possible through quantitative instruments.

The importance of the role of formative evaluation in health education practice must be reinstated. Answers to formative evaluation questions can contribute significantly to the science and art of education and learning about health. For example, questions that seem highly relevant to health education practice but have not received sufficient attention in the literature include: How can programs be adapted to be more compatible with local circumstances? What are the reasons administrative personnel and policy makers support or do not support health instruction in the curriculum? Is the intended target population being reached? Is the program being implemented as planned? Which program activities and materials are most and least likely to be implemented as planned, modified, or omitted? What is the nature of program personnel with which the program is most likely to be successful or unsuccessful? What is the nature of the settings in which the program is most likely to be successful or unsuccessful? What are the factors that are most significant relevant to facilitating and hindering program implementation? What are the most appropriate strategies for promoting program implementation? How might programs be modified to promote implementation? What are the key elements of the program? Emphasis on outcome-oriented research has delayed progress in improving programs and reformulating theories on which they are based. Program planners should recognize that building an implementation-monitoring system into interventions may be one of the most effective techniques available for improving fidelity of implementation. It should also be recognized, however, that implementation monitoring is not sufficient to ensure fidelity and that ensuring fidelity may not always be an appropriate goal.

Investigators should be cautious not to let methodological requirements result in restructuring the operation and administrative processes of the program under evaluation.[67] Program evaluators should also be cognizant of the rigor-versus-significance dilemma described by Green (p. 155)[5]:

> We end up sometimes with rigorously defined but trivial interventions, and other times with significant interventions that are too vaguely defined to be replicated.
>
> What we know, then, is that health education works if it is sufficiently adapted to the problem, the population, and the circumstances in which it is implemented. What we do not know is how to describe those crucial adaptations.

In summary, to avoid Type III errors in evaluations of program effectiveness, implementation of the program being evaluated should be assessed. Monitoring implementation in classes participating in programs, as well as assessing the level of activity in classes not participating in the specific program being evaluated, can enhance the validity of summative evaluations and help to ensure that the efficacy of programs is not evaluated until they have been implemented. Evaluators can assist decision makers within schools to realize that for programs to be implemented will require sufficient attention to professional preparation of teachers, time for teachers to prepare and teach, availability of curriculum materials, administrative support, and curricula that are compatible and adaptable to local circumstances. Implementation monitoring can help to reinstate the importance of formative evaluation in health education practice and thus contribute to improving the science and art of education and learning about health.

References

1. Cook TD, Campbell DT: *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago, Rand McNally, 1979.
2. Dobson D, Cook TJ: Avoiding type III errors in program evaluation: Results from a field experiment. *Eval Prog Plann* 3:269–276, 1980.
3. Rezmovic EL: Program implementation and evaluation results: A reexamination of Type III error in a field experiment. *Eval Prog Plann* 5:111–118, 1982.
4. Scanlon JW, Horst P, Nay JN, et al: Evaluability assessment: Avoiding type III and IV errors, in Gilbert GR, Conklin PJ (eds): *Evaluation Management: A Source Book of Readings*. Charlottesville, VA, US Civil Service Commission, 1977.
5. Green LW: Evaluation and measurement: Some dilemmas for health education. *Am J Publ Health* 67:151–156, 1977.
6. Kreuter MW, Green LW: Evaluation of school health education: Identifying purpose, keeping perspective. *J Sch Health* 48:228–235, 1978.
7. Kolbe LJ: Evaluating effectiveness: The problems of behavioral criteria. *Health Educ* 10:12–13, 1979.
8. Green LW, Heit P, Iverson DC, et al: The school health curriculum project: Its theory, practice, and measurement experience. *Health Educ Q* 7:14–34, 1980.
9. Baric L: Evaluation: Obstacles and potentialities. *Int J Health Educ* 23:142–149, 1980.
10. Hochbaum GM: Certain problems in evaluating health educaton. *Health Values* 6:14–20, 1982.
11. Levine S, Kolbe LJ: Making a difference: Formative evaluation of health education. *Focal Points*, September, 1983.
12. Campbell DT, Stanley JC: *Experimental and Quasi-Experimental Designs for Research*. Chicago, Rand McNally, 1963.
13. Guttentag M, Struening EL: The handbook: Its purpose and organization, in Guttentag M, Struening EL (eds): *Handbook of Evaluation Research* (vol 2). Beverly Hills, Sage, 1975.
14. Boruch R, Gomez H: Sensitivity, bias, and theory in impact evaluation. *Prof Psychol* 8:411–434, 1977.
15. Morris L, Fitz-Gibbon C: *How to Measure Program Implementation*. Beverly Hills, Sage, 1978.
16. Patton MQ: Evaluation of program implementation, in Sechrest L, West SG, Phillips M, (eds): *Evaluation Studies: Review Annual* (vol 4). Beverly Hills, Sage, 1979.
17. Williams W, Elmore RF: *Social Program Implementation*. New York, Academic Press, 1976.
18. Elmore RF: Backward mapping: Implementation research and policy decisions, in Williams W (ed): *Studying Implementation: Methodological and Administrative Issues*. Chatham, Chatham House, 1982.
19. Charters WW, Jones JE: On the risks of appraising non-events in program evaluation. *Educ Res* 2:5–7, 1973.
20. Hall GE, Loucks SF: A developmental model for determining whether treatment is actually implemented. *Am Educ Res J* 3:263–276, 1977.
21. Patterson JL, Czajowski TJ: Implementation: Neglected phase in curriculum change. *Educ Lead* 34:204–206, 1979.
22. Hall GE, Loucks SF: Bridging the gap: Policy research rooted in practice, in Lieberman A, McLaughlin MW (eds): *Policy Making in Education: Eighty-First Yearbook of the Society for the Study of Education*. Chicago, Society for the Study of Education, 1982.
23. Joint Commission on Standards for Educational Evaluation: *Standards for Evaluation of Educational Programs, Projects, and Materials*. New York, McGraw Hill, 1981.
24. Basch CE: Research on disseminating and implementing school health education. *J Sch Health* 54:57–66, 1984.
25. Leithwood KA, Montgomery DJ: Evaluating program implementation. *Eval Rev* 4:193–214, 1980.
26. Williams W: Implementation analysis and assessment, in Cook TD, Delrosario ML, Hennigan KM, et al (eds): *Evaluation Studies: Review Annuals* (vol 3). Beverly Hills, Sage, 1978.
27. Fullen M, Pomfret A: Research on curriculum and instruction implementation. *Rev Educ Res* 47:335–397, 1977.
28. Goodlad JI: *Curriculum Inquiry: The Study of Curriculum Practice*. New York, McGraw Hill, 1981.
29. Freeman HE: The present status of evaluation research, in Guttentag M (ed): *Evaluation Studies: Review Annual* (vol 2). Beverly Hills, Sage, 1977.
30. Patton MQ: *Utilization Focused Evaluation*. Beverly Hills, Sage, 1978.
31. Rossi PH: Issues in evaluation of human services delivery. *Eval Q* 2:573–599, 1978.
32. Sechrest L, West SG, Phillips M, et al: Some neglected aspects in evaluation research, in Sechrest L, West SG, Phillips M, et al (eds): *Evaluation Studies: Review Annual* (vol 4). Beverly Hills, Sage, 1979.
33. Boruch RF, Cordray DS, Pion G, et al: A mandated appraisal of evaluation practices: Digest of recommendations to Congress and to the Department of Education. *Educ Res* 10:10–13, 1981.
34. James RK, Hall G: A study of the concerns of science teachers regarding an implementation of ISCS. *J Res Sci Teach* 18:479–487, 1981.

35. Tornatzky LG, Johnson EC: Research on implementation: Implications for evaluation practice and evaluation policy. *Eval Prog Plann* 5:193–198, 1982.
36. Gross N, Giacquinto JB, Bernstein M: *Implementing Organizational Innovations: A Sociological Analysis of Planned Social Change*. New York, Basic Books, 1971.
37. Shipman MD: The impact of a curriculum project. *J Curric Stud*. 5:47–54, 1973.
38. Berman P, McLaughlin MW: Implementation of educational innovation. *Educ Forum* 40:347–370, 1976.
39. Leinhardt G: Program evaluation: An empirical study of individualized instruction. *Am Educ Res J* 14:277–293, 1977.
40. Leinhardt G: Applying a classroom process model to instructional evaluation. *Curr Inquiry* 8:155–176, 1978.
41. Gibney T, Karns E: Mathematics education—1955–1975: A summary of findings. *Educ Lead* 36:357–359, 1979.
42. Yager RE, Stodghill R: Social science in an age of reason. *Educ Lead* 36:439–445, 1979.
43. Ponder G: The more things change. . . : The status of social studies. *Educ Lead* 6:515–518, 1979.
44. Leinhardt G: Modeling and measuring educational treatment in evaluation. *Rev Educ Res* 50:393–420, 1980.
45. Cooley WW, Leinhardt G: The instructional dimensions study. *Educ Eval Pol Anal* 2:7–25, 1980.
46. Yin RK: Changing urban bureaucracies: How new practices becomes routinized. Lexington, Lexington Books, 1979.
47. St Pierre RG, Cook TD, Shaw R: *An Evaluation of the Nutrition Education Training Program: Nebraska's Experience Nutrition Curriculum*. Cambridge, ABT Associates, 1981.
48. Gersten RM, Carnine DW, Williams PB: Measuring implementation of a structured educational model in an urban school district: An observational approach. *Educ Eval Pol Anal* 4:67–80, 1982.
49. Basch CE: *Evaluation of a School Health Educational Curriculum: Implementation Analysis and Outcome Assessment*. Unpublished doctoral dissertation, Southern Illinois University, Carbondale, IL, 1982.
50. Conrad KJ, Eash MJ: Measuring implementation and multiple outcomes in a child parent center compensatory education center program. *Am Educ Res J* 20:221–236, 1983.
51. Scheirer MA, Rezmovic EL: Measuring the degree of implementation. *Eval Rev* 7:599–634, 1983.
52. Kreuter MW, Christenson GM, Davis R: School health education research: Future issues and challenges. *J Sch Health* 54:27–32, 1984.
53. Basch CE, Sliepcevich EM: Innovators, innovations and implementation: A framework for curricular research in school health education. *Health Educ* 14:20–24, 1983.
54. Foster RL: The fifth graders are excited about how their lungs work. *J Natl Tuberc Respir Dis Assoc* 12:4–9, 1971.
55. Foster RL: A search for inclusion. *Health Educ Monogr* 36:56–60, 1973.
56. Davis RL: Making health education relevant and exciting. *Healh Ser Rep* 88:100–105, 1973.
56. Edson L: Schools attack the smoking problem. *Am Educ* 9:10–14, 1973.
58. Albino J, Davis RL: A health education program that works. *Phi Delta Kappa* 54:256–259, 1975.
59. Bruess CE: School health education: General gains from specific tasks. *J Sch Health* 49:93–95, 1979.
60. School Health Curriculum Project. Atlanta: Bureau of Health Education, DHHS Pub. No. (CDC) 80-8359, 1977.
61. Evans NL, Middleton KH: Two programs exemplify nationwide involvement in school health education. *Health Educ* 10:42–45, 1979.
62. Olsen LK, Redican KJ, Krus PH: The school health curriculum project: A review of research studies. *Health Educ* 11:16–21, 1980.
63. Popham WJ: Appropriate measuring instruments for health education. *Health Educ* 13:23–26, 1982.
64. Iverson DC: Issues related to designing and conducting school health education research. *J Sch Health* 54:50–56, 1984.
65. Heit P: The Berkeley model. *Healh Educ* 8:2–3, 1977.
66. Yin RD: Studying the implementation of public programs, in Williams W (ed): Studying implementation: Methodological and administrative issues. Chatham, NC, Chatham House, 1982.
67. Mullen PD, Iverson DC: Qualitative methods for evaluation research in health education program. *Health Educ* 13:11–18, 1982.