DOCUMENT RESUME

ED 095 198 TM 003 870

AUTHOR Wideen, Marvin F.; Butt, Richard L.
TITLE An Ex Post Facto Evaluation of the Implementation and Use of a Province Wide Junior High School Science Program.
PUB DATE [Apr 74]
NOTE 52p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974); For related document, see TM003832

EDRS PRICE MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS Instructional Materials; *Junior High Schools; Measurement Techniques; *Program Development; *Program Evaluation; *Science Programs; *Summative Evaluation; Teacher Characteristics; Teacher Improvement
IDENTIFIERS Saskatchewan

ABSTRACT

Research was conducted in 1972 to evaluate a program designed to improve the science teaching in grades seven, eight and nine in the Province of Saskatchewan. Since the program included a strategy for implementation and the use of an instructional package, a two-part evaluation was conducted. A general survey was used to evaluate the implementation phase and an in-depth study performed on a sample of classrooms to evaluate the effectiveness of the curriculum packages. The results have implications for curriculum change, in addition to assessing modified products from the curriculum reform movement. (Author)

BEST COPY AVAILABLE

AN *EX POST FACTO* EVALUATION OF THE IMPLEMENTATION AND USE OF A PROVINCE WIDE JUNIOR HIGH SCHOOL SCIENCE PROGRAM

Marvin F. Wideen

Simon Fraser University

Richard L. Butt

McGill University

A.E.R.A. - Chicago, April 1974

AN EX *POST FACTO* EVALUATION OF THE IMPLEMENTATION AND USE OF A PROVINCE WIDE JUNIOR HIGH SCHOOL SCIENCE PROGRAM*

Introduction

The purpose of this study was to conduct an evaluation of a program designed to improve the science teaching at the junior high school level in the Province of Saskatchewan. Since the program involved a strategy for implementation as well as the use of an instructional package, a two part evaluation was conducted.

The first part focused on the implementation phase of the program and involved the use of an instrument administered to all teachers in the province. The second part involved classroom visits and the use of student measures and observation schedules.

In this paper the background, procedures and results of the study are outlined briefly to provide a basis for the thoughts expressed in the concluding section of the paper concerning curriculum evaluation and curriculum change.

---

*The authors wish to acknowledge the support of the Saskatchewan School Trustees Association, The Department of Education and the University of Saskatchewan Regina Campus for financial assistance and to those members of the Faculty of the Regina Campus and the Division III Science Committee who made input at different stages of the research.

Background of the Study

By 1970 science education had undergone two decades of curriculum reform both in Britain and North America. Attention had focused on major curriculum projects designed to improve the teaching of science in the schools. The period had emphasized innovation, development and implementation. The developments associated with this wave of curriculum reform influenced the science programs organized for grades 7, 8 and 9 in Saskatchewan. By the early 1970's, a plateau had been reached in program development for these grades in the province which offered opportunity for conducting an evaluation of the programs in use.

It was evident in 1972 that new thinking related to science teaching was creating pressures for change. The emphasis on environmental education and the concerns being expressed for a more humanistic[1] approach to science teaching reflected a change in thinking from the previous emphasis on the disciplines of science. Further, the programs developed in the 60's were being criticized on several accounts.[2] Such

[1]For an example in the current literature, see Robert Samples, "Science - A Human Enterprise," The Science Teacher, 1972, 39, 26-29.

[2]In a talk delivered to the National Association of Biology Teachers on October 24, 1970, in Denver, Colorado, Paul de Hart Hurd proposed that the inquiry techniques developed in the 60's are not sufficient for problems facing students in the 70's.

thinking was surely to produce a reconstruction of science curriculum if not a new wave of curriculum reform in the 70's. To assure that these future developments be built on information rather than speculation, it was felt that an assessment of past attempts at curriculum change in the province was necessary.

The emphasis on curriculum evaluation and accountability so evident in the late 60's provided an additional backdrop to the study. The notion of accountability had reached the province by the early 70's as board members were beginning to raise the question of value for their dollar. Since the Division III science program involved considerable expense for implementation it seemed appropriate that some means of realting dollar value to implementation should be attempted.

Purpose of the Study

The study was designed to assess the degree of implementation of the program and the amount of change it had accomplished in the classroom. The specific objectives of the two part evaluation study were as follows:

(1) To assess the extent to which local school districts had adopted the instructional package and conducted the appropriate in-service education programs.

(2) To determine if factors such as cost, demographic variables of teachers and school organization were related to

program implementation and use.

(3) To construct a teacher profile of demographic characteristics and school profile of organizational factors typical of high and low implementation.

(4) To determine if the use of the instructional package could be related to the interests, attitudes, classroom perceptions and cognitive ability of the students and the instructional mode of the teacher.

The Division III Science Program

Development of the Program

The Department of Education established several committees in 1964 for the purpose of revising all courses of study at the junior high school level in the Province of Saskatchewan. The science area was to operate within a general philosophy especially established for Division III (grades 7, 8, and 9).

The science committee consisted of four teacher representatives, a principal, a superintendent, a representative from the Faculties of Education, and one from the Department of Education. Its charge was to revise or develop courses of study for grades 7, 8, and 9 science teaching.

The initial deliberations of the committee gave rise to two options for accomplishing their task. First, since the grades 7 and 8 courses had been revised only three or four years earlier perhaps only minor revisions were needed for these grades, allowing a major effort to be spent on a grade 9 program. Alternatively, the committee could start afresh without reference to previous programs.

The committee elected the latter alternative for several reasons. First, the feedback from teachers suggested that the grade 7 and 8 programs were not being well received. Secondly, science education was undergoing major changes--developments had occurred which had not reached fruition prior to the 1961 revision of the seventh and eighth grade programs.

The committee held regular meetings in 1964-65 and concentrated their efforts during a two week writing conference in the summer of 1965.

It soon became evident, during the summer writing conference of 1965 that the committee would not be able to go beyond setting up a structure for the science program and writing an introduction. The latter was to include a philosophy, objectives, and a list of skills.

A program which was being developed in the United States, the Earth Science Curriculum Project,[3] seemed to possess many of the characteristics the Committee intended

---

[3]Hereafter referred to as ESCP.

for the grade 8 earth science program. Schools in Saskatoon and Yorkton piloted the program during the 1966-67 school year.

During the next year 1966-67, three classrooms in Saskatoon made use of a variety of experiences on a trial basis with the expectation that these might be included in a grade 7 program. The experiences were drawn from a variety of sources and many were developed by the teachers involved. They were organized around conceptual schemes somewhat similar to those in BSCS.[4]

In the summer of 1967 a second writing conference was held in order to further develop the programs for the three grades. Separate teams worked on the life science program for grade 7, the earth science program for grade 8 and the physical space science materials for grade 9. Further work was done on the philosophy of the program and its conceptual structure. The seventh grade team wrote investigations based on the past year's trial. The earth science materials were revised somewhat to provide alternatives to the ESCP text which had been used the previous year. A group began work on the space science program attempting to draw together the diverse thinking about the area. The result of this work was a tentative course outline for the three grades.[5]

---

[4]BSCS is the acronym for Biological Science Curriculum Study.

[5]This document was entitled the <u>Tentative Course Outline Division III Science</u>, September 8, 1967.

The program was implemented in two stages. The first involved the spreading of pilot projects. In this phase in-service education was accomplished through teacher-interaction and co-operation. The second stage was more widespread and took place through half day workshops held throughout the province. The workshops were led by teachers trained by the Division III Science Committee during a two day workshop held in Saskatoon in August, 1968. Since this date in-service education has largely been the perogative of local school districts.

Structure and Objectives. The general aspects of the program were to include the following characteristics:

1. The program should have a unitary flavor for the grades and provide a three-year cycle. It was hoped that the grade 9 year would provide a capping effect upon the previous two years and that a student would touch upon the main content areas of life, earth and space science during his grade 7, 8 and 9 school career. The committee wished to avoid having three separate courses for the three grades.

2. The program should be built around conceptual and behavioral schemes in Science. Recent thinking in science education at that time seemed to suggest that a more fruitful way of organizing curricula was to utilize major ideas in science rather than subject

matter content. In addition, certain behavioral themes were identified which would serve to give coherence to student experience. These themes were to provide an overall structure to which the three content areas would contribute. Specific concepts within each subject area were identified for attainment at the three grade levels.

3. The Science program should emphasize student involvement with materials and the investigations should provide for an inquiry approach to learning. The implications of this facet of the program were that equipment and materials should be available. It was noted that the inquiry and investigation approach would require a major change in role for many teachers.

As it was noted in the previous section, the development of the program took place over a period of three years. At the point where the program was considered ready for adoption the seventh and eighth grade were more thoroughly developed than the ninth grade space program. Difficulties had arisen frequently on attempts to define an area of space science. This controversy found itself into implementation of the program as many schools adopted the Introductory Physical Science[6] program for grade nine rather than the

---

[6]Hereafter referred to as IPS.

space science program.

The objectives of the program were the following:

1. To develop in the student an appreciation for, and an interest in science.
2. To challenge the student to think and reason through scientific study.
3. To foster a spirit of inquiry.
4. To develop an understanding and appreciation of the methods by which scientists discover new knowledge.
5. To develop a deeper insight into the inter-relationships that exist in science.
6. To develop and broaden the student's understanding of the big ideas of science.
7. To help the student acquire some of the skills and processes of science.

The Evaluation Model

The difficulties associated with performing an evaluation after a program has been in operation for a number of years stem largely from the fact that little has been written concerning appropriate means for conducting *ex post facto* evaluations. The Campbell and Stanley[7] classification of research designs is largely inappropriate. The recent

[7] See D.T. Campbell and J.C. Stanley, *Experimental and Quasi-experimental Design for Research*. Rand McNally, 1963.

models developed for curriculum evaluation[8] provided a base to work from but were also somewhat inappropriate since they were designed to follow programs through the planning, development implementation and final evaluation stages. Hence only partial use could be made of them. The situation was further complicated because there was no baseline data to work from and no steps were taken as the program was introduced that might facilitate evaluation at a later date.

The model developed for this study utilized in part the notion of discrepancy between expectation and reality. It could be assumed that the program having been in effect for five years should have produced some practices which are observable. If such practices are occurring according to reasonable expectations, then it could be assumed that implementation had, in fact, been achieved. The degree to which a discrepancy exists between expectations and reality represent the degree to which implementation had not been achieved.

Such a notion requires the development of a set of expectations and a means of observing reality to see if those expectations were being met. The expectations for implementation were essentially a description of those conditions that

8 For a thorough discussion of curriculum evaluation models see Sanders and Worthen (in Press, Charles A. Jones, Publishing) or P.A. Taylor and D.M. Comley, Readings in Curriculum Evalaution, W.C. Brown, 1972.

ought to exist in classrooms where the program was in operation. These conditions would logically follow from a careful consideration of the program objectives. Hence the first stage of the evaluation model involved (a) clarification of program objectives and (b) generating of expectant conditions.

The second stage of the model involved the development of observational instruments. These fell into two broad categories--student outcomes and classroom transactions and conditions. The latter category gave rise to development of a general questionnaire designed to assess the degree of implementation as seen by teachers and a set of observational measures to be used by observers.

The third and fourth stages of the model obviously involved the collection and analysis of the data respectively. These stages were to reflect the intent of the evaluation. For example, the analysis of the data would focus in part on determining where and to what extent discrepancies exist. The model was also designed so that the evaluation would provide a descriptive picture science teaching in the province for the grades involved. Two other adjuncts to the evaluation model require comment at this point. One involved the use of an arbitrary implementation scale and the other the validation of measures.

In Figure I, a summary is provided of steps through which the evaluation was conducted.

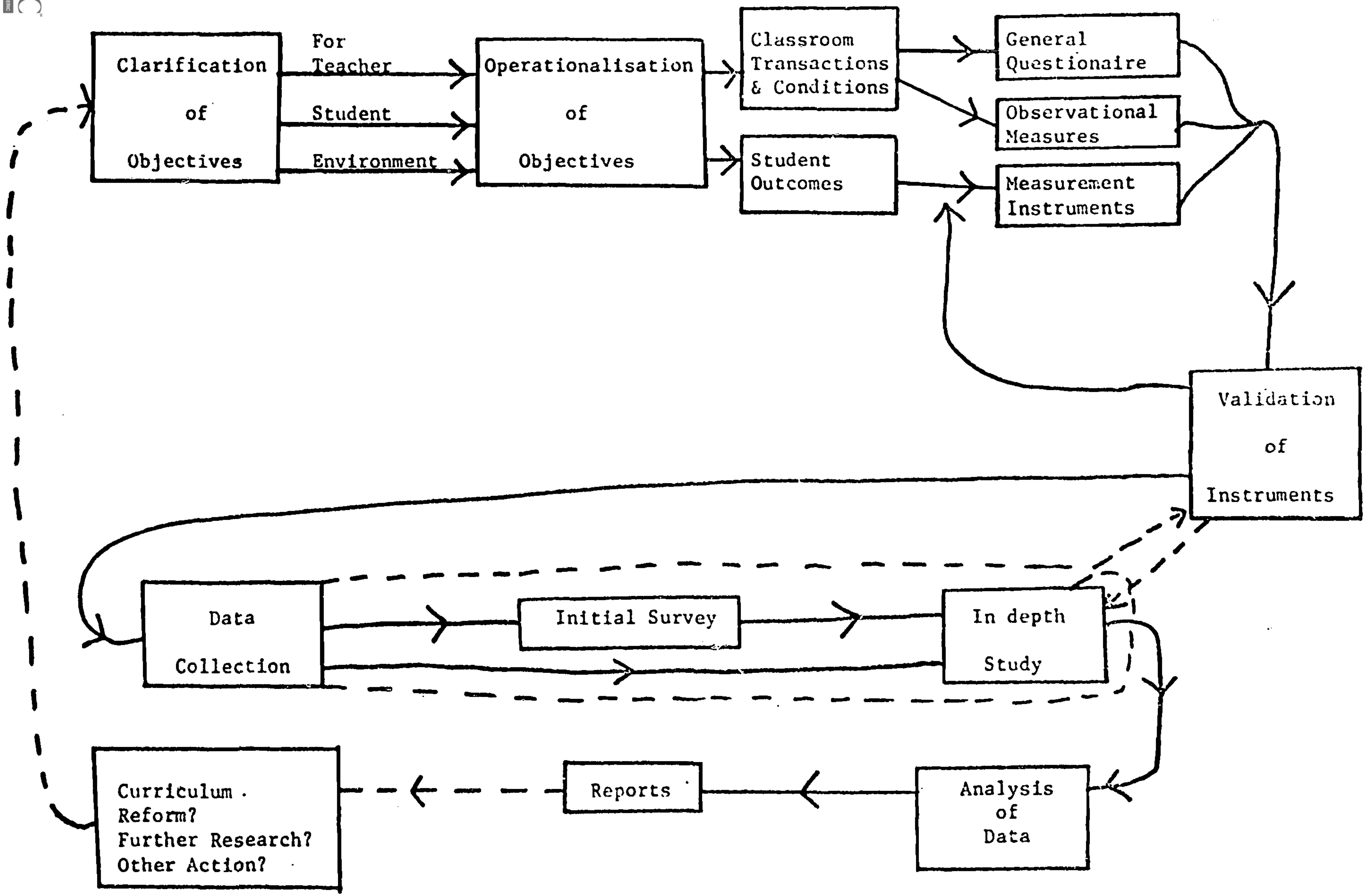

Figure I: Evaluation Model

The arbitrary implementation scale consisted of a set of 28 items built into the questionnaire. This scale was used in the analysis as a criterion for implementation against which to correlate costs, demographic variables and other factors of interest. It was also used to stratify classrooms in the in-depth portion of the study.

The other adjunct to the model, built in at every opportunity, involved the validation of measures and observational schedules. In some cases this was a "boot strap" operation while in other situations external criteria were sought and employed. In some cases a subsequent stage of the evaluation was used to validate a measure used at an earlier stage.

Procedure

Clarifying of Objectives

In late December, 1971, and early January, 1972, the research team examined the minutes, documents and curriculum guides produced by Division III Science Committee for Saskatchewan. Any material that was considered pertinent to the clarification of objectives was duplicated and distributed to members of the Division III Committee, together with samples of data gathering instruments for other sources.

The Division III Science Committee met, together with the research team, in early January of 1972. The purpose of this meeting was three-fold: firstly, to discuss the Division III Science Evaluation project research proposal; secondly, to clarify the objectives for the Division III Science program, and thirdly, to operatize these objectives in the form of a bank of items which could be used for the instruments. Through the use of "brainstorming" sessions a bank of over 200 items were generated. The process was exacting and difficult. It seemed much easier for the group to avoid considering objectives and move directly to the stating of conditions that ought to exist.

As a result of this procedure, a criterion emerged for program implementation. The items making up the criterion were essentially conditions that could be expected to exist if the Division III science program had been implemented successfully. The criteria can be summarized by the following:

1. Teachers should be using one of the components of the Division III instructional package.
2. Teachers should have access to adequate facilities and equipment.
3. Teachers should be devoting relatively more class time to student investigations, student discussion, field trips and activities permitting students to pursue their own interests and less time to lecturing, providing notes and demonstrating.

4. Teacher evaluation should reflect a greater reliance on check lists, inventories, interviews and student observation and less reliance on exams, tests and quizzes.

5. Teachers should have had in-service preparation and should perceive themselves as being capable of teaching the program.

6. Teachers should have a knowledge of the Division III science program and philosophy.

Instrumentation*

General Questionnaire

The items generated by the committee that were suitable for the initial questionnaire were subjected to examination for duplication and divided into categories. The research team prepared a first draft of the questionnaire which was revised on the basis of input from a small panel. A second revision of the questionnaire was given to both rural and city teachers, to science students and to University faculty, in order to validate items and gain feedback. As a result of this validation, which involved forty subjects, items were modified, subtracted, and added, in order to improve the instrument. The third version of the questionnaire was validated again using teachers, science students, and colleagues, as well as the members of the Division III Science Committee. When validating the

*Copies of the instruments are available upon request from Marvin F. Wideen, Faculty of Education, Simon Fraser University, Burnaby 2, B. C.

second and third versions of the questionnaire, the research team sought teachers who represented various degrees of implementation of the Division III Science program. Items and groups of items were studied to insure that the teacher's responses reflected their particular situation. In some cases the teachers discussed the questionnaire with a member of the research team as it was filled out. The resulting questionnaire included categories which involved biographical information, in-service training, teaching activities, teacher feelings, facilities and teacher opinion.

Arbitrary Implementation Scale.[9] This scale consisted of 28 items that were judged to be key indicators of implementation, and were built into the general questionnaire. The scale is included as Appendix E with the weighting and direction for scoring each item. They were selected under five categories reflecting the most persuasive elements associated with implementation that had grown out of the discussions of the Division III Science Committee. These categories included:

1. The opportunity and extent of in-service education.

---

[9]Throughout the remainder of this report the instruments will frequently be referred to by their abbreviated acronyms. These include: II = Interest Inventory; COG = Cognitive Measure; SPSC = Student Perceived Science Classroom; ATT = Attitude Scale; SCOF Science Classroom Observation Form; AIS = Arbitrary Implementation Scale.

2. The acceptance, knowledge and agreement with the Division III philosophy.
3. The self perception of teaching ability for the program.
4. The extent to which certain factors had hindered or helped in implementation.
5. Specific practices in teaching and evaluation.

The selection of items was made prior to the collection of data.

Student Measures

The instruments administered to the students were contained in a question booklet which was reused with the students responding on an answer sheet.

Interest Inventory (II) This measure was developed to assess the degree of student preference for science oriented tasks or activities. The 34 items included 11 which were non-science oriented and 23 which were science oriented.

The introductory words for the statements were verbs such as learn, collect, write, discuss, be involved, explore and plan. They were selected to give some balance to the activities and where possible each was used on four occasions, three times for science oriented tasks and once for a non-science oriented task.

The measure was developed by generating a large pool of items and administering them to classes of grade 7, 8, and 9 students. From this the best items were selected on the basis of how well each seemed to provide variance in response and was related to the students' interest.

The measure was treated as a Lickert scale and a total score generated for each subject.

Cognitive Measure (COG) The 20 items for this test were generated from the life, earth and space science areas in grades 7, 8, and 9 respectively. Every attempt was made to construct items so that a correct response required thought processes at the comprehension and application levels or above. The items were administered to selected classes of seventh, eighth and ninth grade students prior to their use in the study and revised after items analysis. In addition, the items were critiqued by the Division III Science Committee.

Student's Perceived Science Classroom (SPSC) This instrument was developed to provide a measure of the student's perception of his classroom. The items made reference to activities which typified a classroom where students were able to utilize materials in conducting investigations, discuss results with teacher and students, plan some of their activities and offer their own ideas. The items were reversed to provide a high score where students agreed with the statements. It was assumed that in classrooms where students agreed with the statements, the activities described did in fact take place.

Attitude Scale (ATT) This instrument was developed by Moore and Sutman.[10] It had a reported reliability of .93 and was designed to assess both intellectual and emotional attitudes toward science. It is a Lickert type scale and provides a single measure of attitude.

Classroom Observation Schedules

Observers made use of the Science Classroom Observation Form and a set of questions to guide the interviews that were conducted with each teacher.

Science Classroom Observation Form (SCOF) The basis of the Division III science program required a classroom which was student centered, inquiry oriented and provided a setting for manipulative activities. The instrument, SCOF, was developed in order to provide a rating of salient factors operating within this type of classroom. Two sources were used in developing the instrument. The first involved notions and theoretical constructs prevalent in the literature. The second included the specific behaviors listed in the Division III Curriculum guide.[11]

---

[10] For additional information on this test see R.W. Moore and R.Y. Sutman. "The Development, Field Test and Validation of an Inventory of Scientific Attitudes," Journal of Research in Science Teaching, 1970. 7: 85-94.

[11] See the Tentative Course Outline Division III Science, September 8, 1967.

Ideally it might be expected that a teacher who had implemented the program should act as a guide to learning, ask divergent questions and treat science as a tentative rather than absolute discipline. Emphasis should be given to the use of materials as students observe, interpret and draw conclusions. The teacher should stimulate and challenge the students in a psychological atmosphere of freedom and openness where they interact with materials and each other.

Implicit within such a situation is that materials, apparatus, and other learning aids should be available for each lesson enabling the student to develop the "big ideas" and see science as a way of knowing. The student needs to touch, feel, observe, and manipulate materials to provide the external referents for internal processes.

Using these notions and theoretical constructs items were generated for the instrument representing observable manifestations of the theoretical constructs.

This first draft of SCOF was circulated to the Division III Science Committee, colleagues, and students at the Faculty of Education, Regina Campus for comments and suggestions. The revised form was examined, discussed, and reviewed by the assembled Division III Science Committee and the research team in Regina. This provided for agreement of stipulated semantic interpretations of statements and behavior. This was followed by visits to two science classrooms where seven observers completed a SCOF. The responses were compared for each item and each observer providing some rationale for his responses. This enabled further training of observers in consistency of rating and in common interpretation of items and behavior. Several items were rewritten at this time.

The scale was arbitrarily divided into three sections: environment, students' behavior and teachers' behavior. Since the majority of the items were considered interactive between the three, the subscales were considered tentative.

The steps taken in developing the instrument were designed to insure face and content validity. The fact that the instrument distinguished between classrooms at opposite ends of a continuum in terms of implementation during the trials provided a measure of criterion-oriented validity.

SCOF was scored by <u>at least</u> two observers at each observation period and each teacher was observed on at least <u>three</u> occasions. The measure was treated as a Lickert scale with a total score generated for each subject as well as a student, teacher and environment subscale score.

<u>Interview</u>. At some time during each visitation each teacher was interviewed by the team or a member of the team. Although this interview was intended to be open-ended, a list of 11 questions was used as a guide.

Collection of Data

The general questionnaire was distributed to all grade seven, eight and nine science teachers in the Province of Saskatchewan in February, 1972 via superintendents and principals. Of approximately 960 teachers who received the questionnaire, 612 returned a completed version representing a 64% return. The data collected from this portion of the study was in response to objectives one, two and three listed on pages three and four of this paper.

Following the return of the general questionnaire classrooms were selected on a partially random basis to insure a sample exhibiting various degrees of implementation. The specific steps in the procedure were as follows:

1. Each questionnaire was numbered on the basis of its return. From the 612 questionnaires 30 were drawn using a table of random numbers.
2. Each of these questionnaires were examined by three people and rated on a one to five scale in terms of how well the respondent had implemented the Division III science program. Although the rating was done independently by the three observers, agreement was very close in all cases.
3. The ten cases ranked highest in terms of implementation and the ten ranked lowest were then located geographically in the province. This location

involved one classroom within a school. To cut down costs of driving and to utilize the observers' time most economically, all grade 7, 8, and 9 classrooms within that school were observed. In rural areas the closest small town was picked for a similar observation of all Division III classrooms. This step was utilized in order that the observers' time could be used most efficiently.

4. The original two sets of classrooms were then ordered at random from 1-10. Beginning with the first number the superintendent and principal responsible for the school in which the classroom was located was contacted and arrangements made for the visitation.

This procedure continued until approximately 40 classrooms had been obtained.

The procedures used were designed to strike a compromise between the selection of a stratified random sample and the constraints of limited resources. The data obtained from the in depth portion of the study was used relative to the fourth objective listed on page four.

Data Analysis

The results of the questionnaire were considered at three levels in keeping with the intent of the evaluation model. The first level was descriptive involving frequency counts on the items and determining means and standard deviations. A second level of analysis involved the use of chi-square where sets of items were against other items in the questionnaire. These two levels of analysis provided a means by which the general picture of Division III science teaching could be considered. The relationships detected through chi-square were intended to provide some type of composite picture of program implementation relative to the criteria that had been generated. The descriptive analysis would permit an assessment of the specific areas where discrepancies existed between the intended and the real situation. It was through this data that a profile of a teacher who had successfully implemented the program was developed. The third level of analysis involved the scoring of the arbitrary implementation scale and correalting the results with demographic variables on the questionnaire.

The data from the in depth study analysed at through the use of correlational techniques, analysis of variance and an attempt to conceptualize the responses to the interviews with teachers.

Student scores on the four criterion measures, their response to the activities, grade and sex were intercorrelated. This was done without consideration for the degree of program implementation. In order to obtain a picture of the possible interactions among sex, grade and subject preference a 2 x 2 x 3 factorial design was run for each of the criterion measures. In this analysis, some subjects had to be randomly discarded to reduce the cell size to 99 due to limits of the BMD 05V program.

In order to assess the relationship between the scores on SCOF and AIS and the student outcomes, the student results were averaged for each classroom and considered along with scores on the SCOF measure and AIS scores for each teacher. Correlation coefficients were calculated among these variables as well as a 2 x 2 factorial analysis of variance on the four criterion variables.

Validation of Measures

The time devoted to the selection and development of the measures was substantial as items were critiqued and reviewed at every opportunity and rewritten or reviewed where appropriate. This was done to insure a high degree of face and content validity.

Since time did not permit an extensive pilot study it was considered that for the student measures, additional evidence for validity would be provided if certain a-priori conditions were met by the instruments from the results of the study.

1. The student measures were selected because they

were conceptually different. Therefore, high correlations among them were not expected.

2. Previous experience with the ATT measure and research on attitudes would suggest a low positive correlation with the scores on the COG measure.

3. It was expected that the COG scores should increase with successive grades.

4. Students selecting science as a favorite study should score higher on the II than those who selected other subjects.

The criteria used to support the criterion-oriented validity of the AIS included the following:

1. The AIS should correlate with the teachers' perception of how well the program has been implemented.

2. The AIS scores should show significant differences between samples of classrooms using the Division III program and the former program.

3. The AIS scores should correlate highly with the ratings of the research team done on a random sample of classrooms.

Results

Validity of the Instruments

The *a-priori* conditions listed on pages 25 and 26, were designed as expectations that the measures should meet in order that some claim for their criterion-oriented validity could be made. These conditions are discussed relative to the data collected in the study.

The intercorrelations shown in Table 1 are based on all subjects in the sample and provide data relative to the first two conditions listed on page 26. They are low ranging from .06 to .39 indicating that the measures are conceptually different. The correlation of .39 between scores on the COG and the ATT measures was within the expected range of a low positive, significant correlation. The positive, significant relationships that exist between the II scores and SPSC scores as well as the ATT scores and SPSC scores were not anticipated. While the degree of relationship, .34 and .32 respectively, is low, it is nevertheless significant and deserves to be noted. It is conceivable that those students who showed a high interest in science or had a positive attitude toward science would also tend to see their classrooms as places where they could use materials in conducting investigations, discuss results with teacher and student, plan some of their activities and offer their own ideas as the two relationships suggest.

It was expected that the scores from the COG measure would increase with each successive grade. This was found to be the case as indicated in Table 9 where the scores from the COG measure show a significant correlation of .39 with grade. It should be noted, however, that this correlation is based on the classroom means rather than individual student scores. The value of the correlation indicates that the means of the eighth and ninth grade classrooms were successively higher than the seventh grade classrooms.

Relative to the fourth expectation the sample of students was split into those who indicated that science was their first or second choice of subjects and those who selected other subjects. When compared on the basis of their means on the II the groups differed widely. Those favoring science had a mean of 117.25 and the other group 105.66. This difference which was significant at the .01 level was approximately one standard deviation. While this item of information seems to support the obvious, it does nevertheless add something to the validity of the II measure.

Since the use of AIS represented a crucial stage of the evaluation model additional steps were taken to gather evidence oncerning its validity in keeping with the criteria listed on page 26.

Data from the general questionnaire provided information relative to the first and second criteria. The responses to items 38, 39, and 40 which asked teachers in the three grades to state their opinion on the degree to which the Division III program had been implemented, were correlated with scores on the AIS. The correlations were significant

Table 1

Intercorrelations Among Student Measures, Sex and Grade with Reliabilities+ for Student Measures Shown in the Diagonal

Variable	Student Measures			
	II	SPSC	ATT	COG
Student Measures				
II	.73			
SPSC	.34*	.87		
ATT	.27	.32	.71	.56
COG	.13	.06	.39*	
Sex	-.34*	-.07	.01	-.07

+ The reliabilities reported represent internal consistency except in the case of the COG which is KR-20 reliability.

* Significant at the .05 level.

with values of .53, .47 and .34 respectively for seventh, eighth and ninth grade teachers.

The second criteria stated that the AIS should show differences between different groups in the sample. Means for different groups classified according to grade and programs taught are provided in Table 2. It can be seen that differences do exist. For example, "Life science 7" group are approximately half a standard deviation above the "Earth 8" group. This was expected since evidence pointed to a better implementation picture in grade seven than the eighth and ninth grade. Teachers of the "IPS" group, a highly laboratory oriented program, exceeded all others by nearly one standard deviation. Of most significance is the mean of the "Science Activities 7 and 8" which is well below all other groups.

The data provided in Table 3 is related to the third and fourth criteria. As indicated earlier the three member research team independently ranked a random sample of completed questionnaires according to degree of implementation. This was done on the basis of a subjective study of each questionnaire without prior consultation as to prior criteria for ranking, or specific reference to the AIS items. The AIS scores from these questionnaires were later determined for the upper and lower ten. The mean difference between these groups as shown in Table 3 is

Table 2

Mean Values of AIS Scores for Different Groups in the Study

Group	N	Mean	SD
Grade 7	126	109.58	15.76
Grade 8	79	105.05	14.43
Grade 9	124	108.69	13.95
Grade 7 & 8	145	108.10	15.77
Grade 8 & 9	36	109.19	17.32
Grades 7, 8 & 9	28	110.54	16.56
Life science 7	152	112.28	15.19
Life & earth 7 & 8	143	110.07	15.28
Earth 8	54	103.69	14.60
Space 9	74	104.36	12.02
Modern Earth 8 & 9	19	102.80	14.73
Science Activities 7 & 8	17	97.33	12.63
IPS	27	121.15	12.89

Table 3

T-Tests Between Means of Selected Groups for Two Comparisons Representing Implementation and Non-Implementation

Comparison	Group	Mean	SD				
Rating of Questionnaires by Research Team	Upper Ten (X_1)	126.40	7.94	X_1-X_2	df	T	Prob
	Lower Ten (X_2)	91.73	11.19	34.67	18	7.72	.001
Supervisors Rating of Classrooms	Recommended Program (X_3)	118.38	11.39	X_3-X_4	df	T	Prob
	Science Activities (X_4)	103.88	14.25	14.50	19	2.45	.01

highly significant.

The data for the second comparison shown in Table 5 was obtained by selecting a school system where variation in program implementation existed. The superintendent was then asked to indicate the programs used by the teachers. Those teachers using the Science Activities were compared with those using Division III programs on the basis of the mean scores from AIS. It can be seen that the difference between the two groups is significant.

The reliability of the measures are presented in Table 1 in the case of the student measures and Table 7 for the AIS and SCOF. With the possible exception of the COG measure, all others would appear adequate for purposes of this study in which group comparisons were being made.

On the basis of the steps taken during the development of the measures and the results reported in this section, the investigators feel confident in stating that the measures have distinct conceptual underpinnings as indicated by low intercorrelations, possess a reasonable degree of validity (measure what they purport to measure) and are sufficiently reliable (measure consistently) for purposes of this study.

Implementation

The first objective of the study has been to assess the extent to which local school districts had adopted the instructional package and conducted the appropriate in-service education program. Adoption of the instructional package was intended, in an ideal sense, to satisfy the criteria for program implementation (summarized on page 14-15). The results considered relative to this objective included responses on the questionnaire, teacher comments, and the relationships detected through the use of chi-square. The criteria are repeated here to facilitate discussion.

<u>criterion</u>: Teachers should be using one of the components of the Division III instructional package.

Between 85% and 90% of the teachers were using one or more components of the suggested package.

<u>criterion</u>: Teachers should have access to adequate facilities and equipment.

Fifty-five per cent of the classrooms were operating with minimum facilities for science teaching. The average teacher reported having access to approximately $200.00 of equipment for the teaching program.

<u>criterion</u>: Teachers should be devoting relatively more class time to student investigation 3, student discussion, field trips and activities permitting students to pursue their own interests and less time to lecturing, providing note and demonstrating.

<u>criterion</u>: Teacher evaluation should reflect a greater reliance on check lists, inventories, interviews and student observation and less reliance on examins, tests and quizzes.

The amount of class time devoted to teaching activities and the emphasis given different aspects of evaluation ~~are~~ were ranked ~~in Table 4~~ according to their mean values. Student investigations rank high in the class activities along with lecturing and demonstrating and discussions. Field trip activities and pursuing interests rank lowest. Quizzes and tests are the most predominent form of evaluation while interviews rank lowest in this regard.

When these rankings were compared across different components of the instructional package the order remained approximately the same but the mean values differed markedly. The emphasis placed on different evaluation activities showed a more varied pattern for the different instructional components than was the case for instructional activities.

<u>criterion</u>: Teachers should have had in-service preparation and should perceive themselves as being capable of teaching the program.

Results indicated that only 51% of the respondents had had an opportunity for in-service education. Forty-eight per cent reported that science was not the subject that they were most capable of teaching.

criterion: Teachers should have a knowledge of the Division III service program and some understanding of its philosophy.

Data relative to this criterion were gathered from items on the questionnaire, comments made by teachers as well as the interviews. While teachers seemed generally aware that a Division III program existed there were large numbers who exibited substantial ignorance as to its nature and philosophy.

Factors Critical in Implementation

The second objective of the study was to relate implementation to such factors as the costs for equipment, the science background of teachers, the amount of in-service and facilities. This was done by correlating responses on these variables to the scores on an arbitrary implementation scale. Table 4 summarizes these results.

It can be seen that costs for equipment, science teaching capability and the availability of in-service show the highest relationship with scores from the arbitrary implementation scale. Age and teaching show negligible correlations. The amount of academic science and availability of facilities are also positively correlated with implementation.

The third objective of the study was to obtain a

Table 4

Correlations Between Arbitrary Implementation Scale and Selected Variables

Variable	Correlation with Arbitrary Implementation Scale
Sex	.06
Age	.01
University Education	.22*
Academic Science	.31**
Teaching Experience	.06
Science Teaching Capability+	.50**
In-service Education (Opportunity to attend)	.47**
Number of Science Classes Taught	.24*
Facilities	.37**
Costs for Equipment - Grade 7	.57**
- Grade 8	.49**
- Grade 9	.40**

* < .05

** < .10

\+ Response to an item asking if science was the subject they were most capable of teaching.

profile of characteristics, both demographical and organizational, typical of ghigh implementation. Several sources of data were used in arriving at this profile including the correlational analysis discussed previously, teacher comments, interviews and the relationships detected among variables through the chi-square analysis.

University education and academic hours of science appeared to be strong factors in implementation, sharing many parallel relationships. Those with more university education and academic hours of science reported positive perceptions with regard to their capabilities, pre-service and in-service education, philosophy and knowledge of the program. They reported using a variety and mixture of learning and evaluative activities that indicated implementation. Those with less university education exhibited opposite trends, notably using fewer student investigations and having a pattern of activities which indicated low implementation of the program.

Teachers in urban areas reported a pattern of responses suggesting a higher level of implementation than their rural counterparts. The exception to this was that the retention of older programs was higher in large cities than rural areas. Smaller centres reported a better implementation of the grade 7 life science program.

Older teachers tended to be slightly more "formal" with their activities, utilizing more written reports and notes taking, besides feeling they had to stick closely to the subject matter, despite attending more in-service sessions than the younger teachers.

Those who were teaching several classes in science perceived themselves more capable of teaching the subject, had attended more in-service sessions and generally might be considered as having more success with the program than those teaching one class.

On the general questionnaire teachers were asked to rate on a 5 point scale eight factors in terms of their usefulness in helping to implement the program and how much they had hindered implementation. These factors are ranked according to their means in Table 5. It can be seen that administrative and school board support, favorable student reaction and belief in philosophy rank high in terms of usefulness in implementation. However, these four are the lowest factors in terms of hindering implementation. Lack of equipment is ranked highest in this regard.

From these results it was possible to construct a profile of factors reported by teachers associated with a relatively high level of implementation. They were more likely working in elementary or junior high schools, teaching the seventh grade life science program, and teaching science to more than one class. They typically had more

Table 5

Ranking of Factors which Help and Hindered Program Implementation

Factors Useful in Implementation	Mean	Factors Hindering Implementation	Mean
Administrative support	3.73	Lack of equipment	3.03
Favorable student reaction	3.65	Poor space facilities	3.01
Belief in Philosophy	3.53	Inadequate Pre-service	2.97
School board support	3.43	Inadequate Reference Library resources	2.97
Availability of Equipment	3.33	Adverse student reaction	2.08
Adequate reference and library	3.09	Lack of Belief in philosophy	1.96
Good space facilities	2.92	Lack of school board support	1.90
Adequate Pre-service	2.67	Administrative constraints	1.76

years of university and academic science with 3 to 10 years of teaching experience. Such teachers were more likely to have attended in-service sessions, to be working in class-rooms with good space facilities and equipment, and perceiving administrative support as being useful. They had a high understanding of philosophy and objectives, a high knowledge of content and a strong agreement with the philosophy of the program.

The teachers who revealed the lowest level of implementation exhibited a profile opposite to that described above. In particular, they taught in a rural area, had fewer hours of academic science, taught science to fewer classes and perceived administrative support as a hindrance.

Program Use and Student Outcomes

The fourth objective of the study was to compare the student outcomes for those classrooms that had implemented the program with those that had not in order to relate student outcomes to implementation. The scores on the AIS and SCOF were used as measures of classroom implementation.

The data in Tables 6 and 7 provide pertinent information. It should be noted that the correlations in Table 5 were based on all 43 classrooms involved in the study with pairwise deletion for missing data. The data in Table 7 is based on 30 classrooms because those cases having missing data were excluded from the analyses.

Table 6

Intercorrelations among the Classroom Averages of Student, SCOF and AIS Variables for Forty-Three Classrooms

Variable	Student Measures				Grade	SCOF			
	II	SPSC	ATT	COG		ENV	Pupil	Teacher	Total
Activity Preferences									
-discussing									
-experimenting									
-reading									
Student Measures II									
SPSC									
ATT									
COG									
Grade	-.33*	-.46**	.07	.39**					
SCOF Scores									
-environment	.05	.20	.01	-.22	[illegible]	[illegible]+			
-pupil	.30*	.39**	.10	-.18	[illegible]	[illegible]**	.79+		
-teacher	.13	.31*	-.05	-.04	[illegible]	[illegible]	.36**	[illegible]+	
-total	.17	.39**	.10	-.15	[illegible]	[illegible]	[illegible]	[illegible]	.74+
AIS	.01	.37*	.22	-.03	[illegible]	[illegible]	.23	-.35*	.24

* significant at .05

** significant at .01

\+ Reliability coefficients

Table 7

Analysis of Variance for SCOF Total and AIS Scores on Classroom Averages For II, SPSC, ATT and COG Measures

Source of Variance	df	II ms	II f	SPSC ms	SPSC f	ATT ms	ATT f	COG ms	COG f
AIS	1	4.50	.02	140.69	4.03*	55.00	1.09	.01	.00
SCOF	1	23.50	1.52	87.25	2.47	3.00	.06	3.46	1.83
AIS x SCOF	1	20.06	1.33	34.44	.70	1.00	.02	3.22	1.75
Error (df)		15.46 (26)		34.91 (26)		50.24 (26)		1.8[illegible] (26)	

* Significant at the .10 level

From Table 6 it can be seen that only in the case of the SPSC does a pattern of significant relationships emerge with SCOF and AIS. Here correlations of .31, and .39 exist between the SPSC scores and the pupil, teacher and total scales of SCOF respectively. Also a significant relation exists between the AIS scores and the scores on SPSC. The II scores bear a relationship to only the pupil subscale. The ATT and COG scores are unrelated to the SCOF and AIS scores.

The analysis of variance using SCOF and AIS as main effects and the student measures as dependent variables is shown in Table 7. The analysis is complementary to the previous correlation analysis but it provided here to determine if there were interactive effects between the AIS and SCOF main effects.

It can be seen that in only one case is a F value significant. This was where the SPSC scores were dependent variables and AIS the main effect. An examination of the cell means indicated an evaluation of scores on the SPSC for the top AIS group. No other F values were significant indicating no effects for the other measures and no interactive effects.

In order to control for grade three other analysis were conducted. The AIS scores and grade level were used as main effects in one analyses and SCOF scores and grade level in another. Two analyses were run because empty cells resulted when a 2 x 2 x 3 analysis was conducted. A third analysis was run using the 2 x 2 design used above with grade level as a co-variate. No differences were revealed in these analyses that had not been revealed previously.

Discussion

It was found that while over 85% of the schools had adopted the curriculum package, classroom practice and evaluation techniques had not changed substantially from what might have been expected prior to the introduction of the program. Also, it was found that the implementation phase of the program had broken down most significantly in the area of in-service education among teachers. These two results are linked and suggest the obvious conclusion that substantial educational changes can not be brought about through legislation, fiat, or widespread adoption that is not coupled with intense and prolonged in-service education.

Another result had implications related to this conclusion. It was found that teachers with six to ten years experience were having most success with the program. These were teachers who were entering or had just entered the teaching force when the program was being introduced into the schools. This corresponds with a period of fairly intensive in-service training when workshops were being held concerning the program and when teachers were being asked to try new programs. This could be taken as evidence that in-service education does have an effect upon the long range practices of teachers. In addition, it underlines the need for in-service activities to be carried out over an extended period following program implementation.

A main purpose of the study was to compare student outcomes for classrooms in which the program had been implemented with those where it had not. The results indicate that in only one case was a significant difference found. This difference occurred in the area of classroom perception where it was found that in those classrooms where a higher degree of implementation had occurred the students perceived themselves better able to utilize materials in conducting investigations, to plan their own activities and to offer their own ideas. None of the other criterion variables were found to be significantly different for the two groups.

These results might be interpreted as another example of the "no significant differences" pattern so typical in education. However, the interpretation would seem to go further to suggest that (a) fundamental and widespread changes in classroom teaching are unlikely through this type of program development and implementation, (b) levels at which change in education can be detected would seem to exist.

The first of these interpretations requires little elaboration except to reiterate that the study did not reveal substantial changes in teacher behavior and practices of evaluation, and then one should not expect differences in student outcomes.

In the case of the second interpretation, there would appear to be implications for future research in curriculum change. The differences in student outcomes revealed in this study occurred at the level of a student's perception of his or her classroom. A subsequent level might be the interests that students develop and further level their attitude and cognitive ability. Thus a hierarchy might well exist, the lowest rung of which, is the student perception of his classroom. Hence when attempting to identify the effects of a program this level might well provide an initial indicator.

The results of the study both directly from interviews and implied from other data pointed to several constraints to effective teaching and program change.

It would appear that the constraints on teachers and their perception of how curriculum reform should come about are diametrically opposed. None of the teachers interviewed indicated a desire to have curriculum change emanating from an outside group. Yet the constraints they outlined on their effective teaching and the lack of professional interchange seemed to leave little opportunity for curriculum development. In short, if a teacher is too busy to do an effective job of teaching, how can he be expected to work on an improved program of his students.

The sex differences for interest were very clear in this study and are consistent with other findings.[12] The implications remain obvious for those interested in curriculum development and for those concerned with a science program designed for the general population of students. The context of the courses and the predominance of male teachers must surely represent the most likely causes of this type of finding. However, it would seem that further research is needed to pin down the precise reasons for these differences and that future curriculum planning should take this result into account.

The profile of factors associated with a relatively high level of implementation was as follows: working as an elementary or junior-high teacher; teaching science to more than one class; having had more years of university education and more hours of academic science; having three to ten years of teaching experience; having attended

---

12 See M.F. Wideen,"Comparison of Student Outcomes For Science - A Process Approach and Traditional Science Teaching For Third, Fourth, Fifth and Sixth Grade Classes: A Product Evaluation" (In Press, Journal of Research in Science Teaching)

in-service sessions and found them useful; indicated having a high understanding of philosophy and objectives; indicated a high knowledge of content, materials, and equipment; having a high agreement with philosophy, working in classrooms with good space facilities and equipment, perceiving administrative support as being useful.

The teachers showing the lowest level of implementation exhibited a profile opposite to that described above. The strongest of these included the following: teaching in a rural area; having fewer hours of academic science; teaching in a rural area; having fewer hours of academic science; teaching science to fewer classes; and perceiving administrative support as a hindrance.

The two most successful programs in the package appeared to be the grade 7 life science program and the IPS program in grade 9. The two bear little resemblance except that they are both highly laboratory oriented. They differ in context, the amount of structure involved and in the manner in which they were developed. The IPS program was developed as a curriculum package in the United States and in a highly organized program centering around a single conceptual scheme. The life science program was developed within the province, and lends itself to diversity.

The success of the two may be due in part to the committment of the teachers using them. The life science program is taught by teachers who for the first time are involved in a laboratory teaching situation, who have some equipment to work with and are facing enthusiastic students. These factors would surely combine to produce a highly committed and motivated teacher. The IPS teachers are using a program not recommended by the Department of Education. Therefore to use it, it might be assumed that they have a strong commitment to make it work.

The concept of commitment may also be a cause for the rather poor level of implementation and success of the space science program and to a lesser degree the earth science program. In having had the program introduced throughout the province many teachers may not be committed to its value.

The notion of commitment to a program and its success would appear to be a fruitful area for further study. If it can be shown that the level of commitment is a crucial factor in determining a program's success, then numerous implications exist for program development.

In conducting the evaluation, the investigators were constantly aware of the constraints imposed by the absence of context and input which should have been available from earlier stages of program development. Such data would have provided a basis with which to compare the results but more importantly, a context from which to raise more pertinent questions. It would have also aided in the selection of programs. Thus the process of proceeding with program development prior to beginning the process of evaluation and data gathering has little to recommend it.

A further constraint upon the evaluators was the lack of guidelines and curriculum models from which they could draw. The theoretical models were largely inappropriate since they remained theoretical and little published data were not available on their effectiveness or use. The suitable strategies to employ in a post hoc situation faced in this study, were not available through the literature. Thus the approach or model used in this evaluation which made use of an implication scale is offered a means others may wish to consider.