DOCUMENT RESUME

ED 116 917 SE 019 883

AUTHOR Shapiro, Bernard
TITLE The Notebook Problem. Report on Observations of Problem Solving Activity in USMES and Control Classrooms.
INSTITUTION Education Development Center, Inc., Newton, Mass.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE May 73
NOTE 26p.

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS *Decision Making; Educational Research; Elementary Education; *Elementary School Science; Instructional Materials; *Problem Solving; *Science Course Improvement Project; Science Education
IDENTIFIERS National Science Foundation; NSF; Research Reports; *Unified Science Mathematics for Elementary Schools; USMES

ABSTRACT

The aspect of the Unified Science and Mathematics for Elementary Schools (USMES) project described in this paper was undertaken in an effort to observe the problem solving behaviors of elementary school children. The Notebook Problem consisted of presenting a student with three notebooks arranged so as to differ from each other in terms of such dimensions as number of pages, number of lines per page, binding, price, etc.; the subject was asked to (1) select the most appropriate one for his class, and (2) indicate the reasons for his selection. Pretests and posttests were administered to randomly selected students from both control and USMES project groups. Scoring of responses was performed along the following lines: (1) whether any of the subject's reasons for selection were stated in measurable quantities, and (2) the highest level of warrant associated with the reasons stated. Representatives of the dimensions measurable were: (1) size-volume, (2) weight, (3) cost, etc.; while levels of warrant were determined by responses being: (1) personal opinion, (2) testable, or (3) had been tested. Chi-square analysis revealed significant improvement in pretest to posttest scores for the experimental group versus the control group. (Author/CP)

THE NOTEBOOK PROBLEM

Report on Observations of Problem Solving
Activity in USMES and Control Classrooms.

Prepared by

Bernard J. Shapiro
Boston University

for

Unified Science and Mathematics for Elementary Schools

of the

Education Development Center

under a grant from
the

National Science Foundations

Introduction

The evaluation of the Unified Science and Mathematics for the Elementary Schools (USMES) program in 1971-1972 encompassed a number of different strategic approaches in both developmental (first stage), implementation (second stage), and control (comparison) classes at various grade levels. Included in the range of evaluation techniques were teacher logs, classroom observations, and standardized tests. In addition to the data yielded by these approaches, however, there was a desire to observe the problem solving behavior of elementary school children in a situation which was standardized and structured but which provided the subjects an opportunity to consider and test hypotheses with concrete materials.

In order to accomplish this purpose, the Notebook Problem was devised. It consisted essentially of presenting the testee with three notebooks selected so as to differ from each other in terms of such dimensions as number of pages, number of lines per page, binding, price, space between lines, width of ruled margin, etc. etc., and asking the testee to (a) select the most appropriate one for his class, and (b) indicate the reasons for his selection. The test was designed to be administered individually, and the precise directions given by the tester to the testee were as follows:

> Suppose (insert Principal's name) decided that all the (insert testee's grade level) grade should have notebooks to keep their science and math work in. He writes a notebook company and asks for samples. They send him three (point to each notebook) notebooks. He (she) comes to you and says, "(Insert testee's name), I need your help." "Which of these books would be the best for the (insert testee's grade level) grade to keep their science and math work in?"

Procedure

The administration of the notebook problem test was limited to USMES implementation classes and their corresponding control groupings. The initial sample consisted of seventeen USMES and seventeen control schools with forty-three experimental and thirty-one control classes involved. In both categories classes in grades two through six were represented as were all current USMES implementation units. The test was to be administered at both the beginning (pretest) and end (posttest) of the school year, but given the necessity for individual administration, it was impracticable to administer to all members of each class unit. Therefore, testers (who were the classroom observers already being used for USMES activities) were asked to randomly select five pupils from each of the designated classrooms as testees. The pretest and posttest selections were made independently since it was felt that practice effects would be both large and uncontrolled in this kind of test setting.

Testees were taken from the classroom for the test administration. Each tester was asked to allow as much time as necessary for the testee to complete his or her work or the problem and to encourage as full a response as possible. Specific tools such as paper, pencil, and ruler were available but not pressed upon the testee.

Results

(a) Sample

Administration of the notebook test proved to be somewhat more difficult than originally envisioned. A number of particular administrative problems arose (e.g., availability of testees for testing, time of testing, school

schedules, etc.), which made it impossible for the full quota of five pupils per classroom unit to be met. At the end of the school year pretest and post-test results were available for seven school districts (grades two through six), the total sample comprising of thirty-one USMES classes and twenty-two controls. The number of pupils tested in each class varied from two to six, and the final sample consisted of two hundred and twenty-seven pretests (one hundred and thirty-two USMES, ninety-five controls) and two hundred and forty-six posttests (one hundred and forty-four USMES, one hundred and six controls).

Seven of the USMES teachers were male and twenty-four were female. Their variation in terms of teaching experience was quite marked. The range was from one to twenty years with about twenty-five percent of the thirty-one USMES teachers having less than three years of classroom experience. The demographic variation of the seven school districts was also considerable. Three were at the center of large urban areas with largely lower and lower middle class pupils; three were in suburban areas with largely middle and upper class pupils, and one was a rural area with a student group of mixed socio-economic backgrounds.

(b) Test Data

(i) Scoring

All testers transcribed verbatim the testees' responses. These responses were then typed in preparation for scoring of the protocols. After examination of several pilot protocols, a rather elaborate set of response categories had been developed (cf. Appendix C). In examining the actual date, however, it appeared that there was not, in general, enough variability in the subject responses to make all of the sub-categories operative. In addition, there were some aspects of programmatic concern (e.g., whether or not subjects resorted to

direct measurement in their problem solving) which were not being addressed. Therefore, a simpler but more direct schema was developed. In this new approach each protocol was to be assessed in terms of:

(a) whether or not any of the subject's reasons for selection were stated along dimensions that were measurable within the test situation, and

(b) the highest level of warrant associated with the given reasons for selection.

The dimensions measurable within the test situation were (i) size-volume (e.g., "bigger sheets than the small one," etc.), (ii) weight (e.g., "heavier" etc.), (iii) quantity (e.g., "more sheets than the large one," "more lines per page," etc.), and (iv) cost (e.g., "doesn't cost as much as the big one," "costs less for number of sheets," etc.). Three categories were developed for level of warrant. These were (i) reasons given was expressed simply as a personal opinion, (ii) a test was suggested to assess the reason given, and (iii) a test was actually performed to test the reason given. These levels were considered a hierarchy in increasing order of appropriateness, and each protocol was assigned to the highest level present among the several that an individual subject may have used.

Four graduate students in mathematics and science education were trained for a full day in the use of the scoring categories with the help of sample protocols. Inter-judge reliability was assessed through the intra-class correlation. The coefficients yielded at the start of training were +.67 and +.71 for the measurement dimension and the level of warrant respectively. At the end of a day of training, the corresponding coefficients were +.87 and +.89.

In the final scoring of the protocols, the pretests and postests were intermixed, and the protocol pool was then randomly divided among the four raters.

(ii) Data Analysis

In terms of whether or not any of the subject's reasons for selection of a particular notebook were stated along dimensions that were measurable within the test situation, the summary data by District areas are presented in Tables 1 and 2 for the USMES and Control classes respectively. The pattern in these tables is quite clear. At the beginning of the school year, the pretest data indicate that in all districts and in both USMES and Control classes, only a small minority of the pupils state reasons for their notebook selection in terms of dimensions that are directly measurable within the test situation. At the end of the year, however, the posttest data indicate that for the USMES group there has been a considerable change. At this later juncture, the USMES pupils are virtually all responding in terms of measurable dimensions with only eight of one hundred and forty-four USMES students having protocols whose response rationales all fall into the non-measurable category. As indicated in Table 1, Chi Square contingency tests indicate a statistically significant

Table 1 about here

relationship between test-time (i.e., pretest vis a vis posttest) and category response (i.e., measurable vis a vis non-measurable reason for notebook selection) in all USMES districts. As described above, in each case, this shift is from the use of non-measurable dimensions to the use of measurable ones.

The posttest situation with the control classes, however, presents a startling contrast. The great majority of these subjects (cf. Table 2) continued as in the pretest to offer rationales stated in terms of non-measurable dimensions. Thus, in all eight District/School areas for which complete control

Table 1

District/School Data on Reasons for Selection
Measurable vs. Non-Measurable
USMES Group

District/School	Reasons Given				Chi Square*
	Pretest N		Posttest N		
	Measurable	Non-Measurable	Measurable	Non-Measurable	
1	4	13	16	3	11.04**
2	2	6	8	0	6.67**
3	4	12	13	1	18.86**
4	7	15	24	0	21.28**
5	12	33	44	1	64.57**
6	4	12	18	2	13.19**
7	2	6	7	1	4.06**

*df = 1
**statistically significant at the five percent level

group data were available, the Chi Square contingency tests yielded no

Table 2 about here

statistically significant relationship between test-time and response category.

The individual classroom data for both pretests and posttests along the measurement/non-measurement dimension are given in Appendix A to this report. Examination of these data for both the USMES classes (Appendix A, Table 1) and the Control classes (Appendix A, Table 2) consistently substantiate by classroom unit the results described above for the District areas. Although the small N's per individual class made statistical tests inappropriate, it is apparent that as with the summary District areas, there was in the USMES groups a shift over the treatment period from the use of non-measurable dimensions to the use of measurable ones while over the same period, there was no comparable shift in the Control groups. The extreme consistency of this pattern appeared to make special tests for grade level, USMES unit groups, teacher experience, etc., unnecessary.

In terms of the Level of Warrant associated with the subject's reasons for selection of a particular notebook, the summary data by District areas are presented in Tables 2 and 4 for the USMES and Control classes respectively. As with the previous data, the pattern in these tables is quite clear. At the start of the school year, the pretests indicate that in all districts and in both USMES and Control classes, the great majority of students rationalized their notebook selections solely in terms of personal opinion. Some few did suggest a test of their hypotheses, but virtually none actually tested their rationales in the test situation. At the end of the school year, however, the

Table 2

District/School Data on Reasons for Selection
Measurable vs. Non-Measurable
Control Group

District/School	Reasons Given				Chi Square*
	Pretest N		Posttest N		
	Measurable	Non-Measurable	Measurable	Non-Measurable	
1	1	3	1	4	.39
2	2	11	5	13	.67
3	1	3	2	3	.06
4	7	17	7	17	.10
5	10	20	6	24	.77
6	3	9	2	14	.13
7	2	6	3	5	.30

*df = 1

posttests show that for the USMES group there was a marked change in test behavior. At the end of the school year - after having worked with the various USMES materials - the USMES classes are using higher levels of warrant. Almost all of the USMES pupils are either suggesting tests that would assess the validity of their notebook selection or actually performing the test within the problem solving situation. Only twelve of the one hundred and forty-four USMES students tested at posttest time offered personal opinion as a warranty of their response. As indicated in Table 3, Chi Square contingency tests indicate a statistically significant relationship between test-time (i.e.,

Table 3 about here

pretest vis a vis posttest) and category response (i.e., personal opinion, suggested test, actual test) in all USMES districts. As described above, in each case this relationship indicates a shift from personal opinion to suggested and actual testing as warrants for recommended action.

In the control classes, however, there appeared no striking shift from the use of personal opinion as a warrant for notebook selection. As indicated in Table 4, the great majority of the control subjects continued as in the

Table 4 about here

pretest situation to rely on personal opinion. Only sixteen of one hundred and six pupils suggested a test to validate their selection and none actually performed a test in the problem solving setting. Thus, in all eight District areas for which complete control group data were available, the Chi Square contingency tests yielded no statistically significant relationship between test-time and frequency of use of the various levels of warrant.

The specific classroom data for both pretests and posttests along the level of warrant dimension are given in Appendix B to this report. Examination

Table 3

District/School Data on Level of Warrant
Opinion, Suggested Test, Actual Test
USMES Group

District/School	Level of Warrant						Chi Square*
	Pretest N			Posttest N			
	Opinion	Suggested Test	Actual Test	Opinion	Suggested Test	Actual Test	
1	16	1	0	3	11	5	22.19**
2	7	1	0	0	6	2	12.58**
3	14	2	0	2	12	6	21.98**
4	19	3	0	2	12	10	29.13**
5	32	12	1	2	17	26	50.48**
6	14	2	0	3	6	11	19.92**
7	8	0	0	0	4	14	16.00**

*df = 2
**statistically significant at the five percent level

Table 4

District/School Data on Level of Warrant
Opinion, Suggested Test, Actual Test
Control Group

District/School	Level of Warrant						Chi Square*
	Pretest N			Posttest N			
	Opinion	Suggested Test	Actual Test	Opinion	Suggested Test	Actual Test	
1	4	0	0	4	1	0	.01
2	13	0	0	15	3	0	1.38
3	4	0	0	4	1	0	2.39
4	21	3	0	22	2	0	.00
5	28	2	0	25	5	0	.65
6	12	0	0	13	3	0	.94
7	7	1	0	7	1	0	.57

*df = 2

of these data for both the USMES classes (Appendix B, Table 1) and the Control classes (Appendix B, Table 2) consistently substantiate by individual classroom unit the results as described above for the summary areas. Although as with the measurable/non-measurable dimension, the small N's per individual class made statistical tests inappropriate, it is obvious that as with the larger areas, there was in the USMES groups a change from the use of personal opinion as a warrant to the use of suggested and actual tests. Over the same period, however, there was no comparable shift in the control groups. The basic consistency of this pattern made further comparisons by grade level, USMES unit, etc., inconsequential.

Conclusion

The present aspect of the 1972-73 USMES evaluation was undertaken in an effort to observe the problem solving behavior of elementary school children (both USMES and controls) in a situation which was standardized and structured but which provided the subjects an opportunity to consider and test hypotheses with concrete materials. The dependent variables of concern were (a) the use of rationales stated in terms of dimensions that were measurable within the problem solving situation, and (b) the level of warrant associated with solutions in the same problem solving situation.

The problem solving behavior of the children was observed at both the beginning and the end of the school year, the treatment being the use of one or other of the USMES units between these two occasions. Data analyses indicated that both USMES and control groups began the school year (a) by using non-measurement dimensions in their problem solving and (b) by relying on personal opinion as the warrant for the validity of their solutions. At the end of

the school year, however, the USMES pupils were relying primarily on measurement dimensions and using suggested and actual tests to validate their work. The control pupils on the other hand continued to exhibit the pattern of behavior of the pretest situation. Thus, it would appear that in terms of the two dependent variables studied, the USMES experience had, irrespective of units and teachers involved, a marked and positive effect on the students' problem solving behavior.

Caveat

The present effort was intended to be an exploratory "first step" in the development of a series of problem solving tasks appropriate to the evaluation of USMES-type programs. It suffered from some of the difficulties of first stage developments. The directions given to the testers were not adequately specific so that there was considerable variation in the style of administration. For example, it was unclear to the testers just how much time was "adequate" or how to judge when a testee had finished. Further, there is some difficulty over the partially "ex post facto" nature of the rating categories, and these would certainly need to be cross validated in future work. Finally, it must be pointed tha the control classes cannot be considered comparable to the USMES groups in the strict sense of the term in that very relaxed standards were used in their selection although efforts were made to chosse groups in the same school and at the same grade level.

The redeeming aspect to these difficulties lies in the clarity and consistency of the actual results. Nevertheless, future work in this area should take steps to eliminate or at least reduce the confounding effects of these variations.

APPENDIX A

Classroom Data on Reasons for Selection

Measurable vs. Non-Measurable

Table 1

Classroom Data on Reasons for Selection
Measurable vs. Non-Measurable
USMES Classes

District	Class	Grade	Reasons Given			
			Pretest N		Posttest N	
			Measurable	Non-Measurable	Measurable	Non-Measurable
1	a	6	2	3	4	1
	b	6	1	4	4	1
	c	5	1	3	4	1
	d	3	0	3	4	0
2	a	4	1	4	4	0
	c	4	1	2	4	0
3	a	4	1	3	5	0
	b	4	1	3	5	0
	c	2	1	3	4	1
	e	3	1	3	5	0
4	g	6	2	2	4	0
	h	5	1	3	4	0
	i	5	1	3	4	0
	j	4	2	2	4	0
	k	3	0	2	4	0
	l	3	1	3	4	0
5	a	4	1	4	5	0
	b	4	2	3	5	0
	c	4	0	5	5	0
	d	5	1	4	5	0
	e	5	1	4	5	0
	f	5	2	3	4	1
	g	6	2	3	5	0

Table 1 (Cont'd.)

District	Class	Grade	Reasons Given			
			Pretest N		Posttest N	
			Measurable	Non-Measurable	Measurable	Non-Measurable
5	h	6	1	4	5	0
	i	6	2	3	5	0
6	a	3	0	4	4	1
	c	5	1	3	4	1
	e	4	2	2	5	0
	f	4	1	3	5	0
7	a	5	1	3	3	1
	b	4	1	3	4	0

Table 2

Classroom Data on Reasons for Selection
Measurable vs. Non-Measurable
Control Classes

District	Class	Grade	Reasons Given			
			Pretest N		Posttest N	
			Measurable	Non-Measurable	Measurable	Non-Measurable
1	e	5	1	3	1	4
2	b	4	0	2	1	3
	d	4	0	3	1	3
	f	5	0	4	1	4
	g	3	2	2	2	3
3	d	4	1	3	2	3
4	a	3	1	3	1	3
	b	3	1	3	1	3
	c	4	2	2	1	3
	d	5	1	3	2	2
	e	5	0	4	0	4
	f	6	2	2	2	2
5	j	4	1	4	1	4
	k	4	2	3	1	4
	l	6	2	3	2	3
	m	6	1	4	1	4
	n	5	2	3	1	4
	o	5	2	3	0	5
6	b	3	0	4	0	6
	d	4	1	3	1	4
	g	4	2	2	1	4
7	c	4	0	4	1	3
	d	5	2	2	2	2

APPENDIX B

Classroom Data on Level of Warrant

(i) Opinion
(ii) Suggested Test
(iii) Actual Test

Table 1

Classroom Data on Level of Warrant
Opinion, Suggested Test, Actual Test
USMES Classes

District	Class	Grade	Level of Warrant					
			Pretest N			Posttest N		
			Opinion	Suggested Test	Actual Test	Opinion	Suggested Test	Actual Test
1	a	6	4	1	0	1	3	1
	b	6	5	0	0	1	3	1
	c	5	4	0	0	1	2	2
	d	3	3	0	0	0	3	1
2	a	4	4	1	0	0	3	1
	c	4	3	0	0	0	3	1
3	a	4	3	1	0	0	4	1
	b	4	4	0	0	1	3	1
	c	4	4	0	0	0	3	2
	e	3	3	1	0	1	2	2
4	g	6	4	0	0	1	2	1
	h	5	3	1	0	0	2	2
	i	5	4	0	0	1	2	1
	j	4	3	1	0	0	3	1
	k	3	2	0	0	0	2	2
	l	3	3	1	0	0	1	3
5	a	4	4	1	0	0	3	2
	b	4	4	1	0	0	3	1
	c	4	4	1	0	0	2	3
	d	5	3	2	0	0	1	4
	e	5	2	2	1	0	2	3
	f	5	3	2	0	0	2	3
	g	6	3	2	0	0	1	4
	h	6	4	1	0	1	2	2
	i	6	5	0	0	0	1	4

Table 1 (cont'd.)

District	Class	Grade	Level of Warrant					
			Pretest N			Posttest N		
			Opinion	Suggested Test	Actual Test	Opinion	Suggested Test	Actual Test
6	a	3	4	0	0	1	2	2
	c	5	3	1	0	1	1	3
	e	4	4	0	0	1	2	2
	f	4	3	1	0	0	1	4
7	a	5	4	0	0	0	1	3
	b	4	4	0	0	0	3	1

Table 2

Classroom Data on Level of Warrant Opinion, Suggested Test, Actual Test Control Classes

District	Class	Grade	Level of Warrant					
			Pretest N			Posttest N		
			Opinion	Suggested Test	Actual Test	Opinion	Suggested Test	Actual Test
1	e	5	4	0	0	4	1	0
2	b	4	2	0	0	3	1	0
	d	4	3	0	0	4	0	0
	f	5	4	0	0	4	1	0
	g	3	4	0	0	4	1	0
3	d	4	4	0	0	4	1	0
4	a	3	4	0	0	4	0	0
	b	3	4	0	0	3	1	0
	c	4	3	1	0	4	0	0
	d	5	3	1	0	3	1	0
	e	5	4	0	0	4	0	0
	f	6	3	1	0	4	0	0
5	j	4	5	0	0	4	1	0
	k	4	5	0	0	5	0	0
	l	6	4	1	0	3	2	0
	m	6	5	0	0	4	1	0
	n	5	4	1	0	5	0	0
6	b	3	4	0	0	5	1	0
	d	5	4	0	0	4	1	0
	g	4	4	0	0	4	1	0
7	c	4	4	0	0	4	0	0
	d	4	3	1	0	3	1	0

APPENDIX C

Original Response Categories

I. MEASURABLE (during test situation)

a. size-volume
b. weight
c. quantity
d. cost

II. TESTABLE (could be studied, examined in future)

a. versatility
b. construction (durability, lack of defects)
c. manageability
d. health
e. specific utility

III. QUALITATIVE (general statements of opinion)

a. attractiveness
b. appeal to:

tradition
authority
peers
personal preference
nationalism (made in USA)

c. prestige
d. uniqueness