DOCUMENT RESUME

ED 041 957 24 TM 000 069

AUTHOR Romberg, Thomas A.
TITLE Achievement Monitoring Via Item Sampling (Revised).
INSTITUTION Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Cooperative Research Program.
PUB DATE Mar 70
CONTRACT OEC-5-10-154
NOTE 19p.; From symposium "Some Methodological Consideration in Curriculum Evaluation," American Educational Research Association Meeting, Minneapolis, Minn., March 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$1.05
DESCRIPTORS Achievement Tests, *Arithmetic, Audiovisual Instruction, *Curriculum Evaluation, Elementary School Mathematics, Evaluation Criteria, *Evaluation Techniques, Grade 6, Item Analysis, *Item Sampling, *Program Development, Research Design

ABSTRACT

As a new program is developed, information must be collected to identify weaknesses, to guide the staff in the revision process, and to formulate decision-making procedures, i.e., formative evaluation techniques are essential. A set of criteria for the information needed in formative evaluation is set-up and a practical strategy for meeting these criteria is described. It is suggested that the ideas of achievement monitoring, time series experimental design, and item sampling be jointly applied. Such a procedure has been utilized for a formative evaluation of a new sixth grade arithmetic program. This evaluation, including suggestions for its improvement, is described in detail. An example Patterns in Arithmetic test with item profiles over time, is appended. (DG)

ACHIEVEMENT MONITORING VIA ITEM SAMPLING (REVISED)

A practical data gathering procedure for formative evaluation

Thomas A. Romberg
Associate Director

University of Wisconsin
Research and Development Center for Cognitive Learning

Paper read in the symposium:
SOME METHODOLOGICAL CONSIDERATION IN CURRICULUM EVALUATION
Presented at the
American Educational Research Association
March 2-6, 1970 Minneapolis, Minnesota

The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education and Welfare, under the provisions of the Cooperative Research Program (Center No. C-03, Contract OE 5-10-154).

ACHIEVEMENT MONITORING VIA ITEM SAMPLING:

A practical data gathering procedure for formative evaluation

Thomas A. Romberg
Associate Director

University of Wisconsin
Research and Development Center for Cognitive Learning

Introduction

Designing and developing new products is a difficult and expensive task. This is certainly true in education where during the past decade designing and developing new materials and programs has become a major enterprise. The initial problem faced by developers is to determine whether or not newly engineered components of an instructional system, such as new texts or new teaching strategies, are useful in reaching a set of specified goals. Making decisions about the utility of materials is an outcome of formative evaluations. The term "evaluation" includes both the techniques of gathering the data and the procedures for making the decisions. The term "formative" implies that the information would be collected during development to identify weaknesses and guide the staff in the revision of materials.

The purpose of this paper is to describe a procedure which has proven useful for gathering information needed for formative evaluations.

Background

Although vast funds, both government and private, have been and are continuing to be invested in new educational programs, well conceived, carefully designed and executed formative evaluations have not been conducted. The reasons for this lack of good formative studies are many. But, two facts seem clear: first, developers themselves have not nor are they likely to

create practical alternatives to the haphazard sujective methods in common use; and second, when they have asked for advice from others it has been inappropriate or impractical.

Although good formative studies have not been done, it is not because of a lack of interest on the part of developers. All admit to the value of information as to a product's use in order to revise and improve that product. In fact, it is unlikely that there exists a developer who has not gathered information from users. However, most of this information has been haphazardly collected subjective self-report data from teachers. The validity, reliability, and generalizability of which is highly suspect. But, it must be remembered developers are creative engineers not behavioral scientists. It is unrealistic to expect inventors to evaluate.

Unfortunately, well-intentioned but very naive research specialists have too often given in appropriate and impractical advice to developers. The lack of understanding of the dynamics of development and the sequencing of instruction has led to this failure. Too often evaluation has been viewed in terms of classical experimental designs. Constraints such as random assignment of students to treatments, or demanding behavioral objectives, or suggesting the use of comprehensive achievement tests have usually seemed quite foolish to most developers. Foolish in the sense that they have not or are not related to making the utility decisions necessary for revision of materials.

Perhaps achievement categories rather than detailed behavioral objectives would be more useful descriptions of goals in most development projects. In fact, unless behavioral objectives have been used to generate the instructional program, deriving them for formative evaluation could detract from the efforts of the staff. Also, advisors often have failed to see that most

programs have a variety of objectives. Some more important than others. Some possibly even at cross purposes with each other.

Developers fear that comprehensive tests are too gross to be useful. In fact, Stake believes "the standard achievement test is unlikely to encompass the scope or penetrate the depth of a particular curriculum being evaluated." (Stake, 1967, p. 6) What developers want is descriptive information as to how well students are performing related to certain important goals of the project.

Developers also recognize that instruction is dynamic not static. Thus, they intentionally design instruction to include spiral sequencing of topics. In contrast to what many people think, these instructional patterns are not clearly laid out during development. Most instructional programs have been developed around a set of goals, some interrelated and some independent with various tactics being used at various times of a year. Formative evaluation cannot be separated from development. The day-to-day engineering decisions have to be made when one is developing instructional materials. What is needed is descriptive information easily collected and easily handled which can be used to aid in making the decision. Based on this discussion, the following are proposed as practical conditions to be met for gathering information in formative evaluations:

1. Information collected should be logically related to the important objectives of the project.[1]

---

[1] This does not imply that unintended outcomes should not be looked for (see Messick, 1969). What it does imply is that the primary effort should concentrate on intended outcomes.

2. Information should be easy to collect and report.

3. Gathering information should take very little time away from implementing the program.

4. Costs should be minimal.

A Practical Procedure

The strategy described in this paper is practical in that it meets the above conditions. The procedure is called achievement monitoring. This refers to periodic achievement testing of the group being used to pilot the instructional materials. Periodic data gathering is necessary in formative evaluation since instruction on a particular topic may occur many times during the year. Simple pretest-posttest administration does not make much sense since only the cummulative effect is noticed and not the unique effect of a topic or set of lessons.

Periodic administration of an observation instrument is no more than the classic time series experimental procedure long used in the natural sciences. (See Campbell and Stanley, 1963) Here the usual design has been modified to include multiple intervention.

Item sampling is suggested as a means of data collection because it is both adequate for making good estimates of group performance as a result of instruction and it is efficient. The technique of item-sampling was first proposed by Lord (1962) and has been defined as follows: "In the item-sampling technique, a set of m items is randomly broken up into k subsets of items. The k subsets of items are then randomly assigned to p pupils or subjects. Each subject takes only a portion of the complete set of items." (Cahen, Romberg, Zwirner, 1970). For example, suppose there are 100 items about which one desires information. Instead of having each pupil

respond to all 100 items, one could construct five, 20-item tests and administer these randomly among pupils. Thus, it takes little class time to get this information.

It should be noted that achievement monitoring is not a new concept, neither are time series experimental designs, nor is item sampling. What is new is the joint application of these three ideas as a means of gathering data for formative evaluations. To carry out the strategy for gathering such data five steps must be followed.

1. The specification of the major terminal objectives of the program.
2. The collection of a pool of items to measure each objective.
3. The construction of a battery of tests via item-sampling.
4. Periodic administration of the battery.
5. Construction of item and objective profiles.

An Example--Patterns in Arithmetic

To elaborate on each of these steps, the formative evaluation for the Sixth Grade Patterns in Arithmetic program via TV (PIA-6) developed under the direction of Henry Van Engen through the auspices of the Wisconsin Research and Development Center will be used as an example. This formative evaluation was carried out by Mr. James Braswell (1970).

1. Specification of objectives

The method of determining major objectives of any program is up to the developer. For PIA-6 a typical content specification procedure was used. Prior to developing the program a proposed topic coverage was presented. (See Table 1). The sixteen subareas of that topic outline were used as the achievement categories around which instruction was engineered.

Table 1

PROPOSED TOPIC COVERAGE FOR PIA-6

I. Geometry (30%)
   A. Measurement
   B. Non-metric aspects of geometry

II. Fractions (30%)
   A. Interpretation
   B. Equal Fractions
   C. Operations (+, -, x)
   D. Decimals
   E. Operations (÷)

III. Counting Numbers (15%)
   A. Operations
   B. Factors
   C. Equivalence
   D. Functions

IV. Ratio (5%)
   A. Cross product
   B. Percent

V. Miscellaneous (20%)
   A. Probability
   B. Statistics
   C. Problem solving - extensions of program content

2. Developing an item pool

Measuring success in any school experience is a difficult task. Ideally, it would be useful to develop an algorithm which could be used as a means of generating a very large finite pool of items from which one could sample. (See Hively, et al., 1968). However, for most achievement categories it would be more useful to collect a pool of items which are judged to be appropriate to help decide whether or not instruction has been successful. For PIA-6 a large set of items were collected from various sources. After the staff reviewed the items 240 were selected as adequate and representative of the content areas above

3. Construction of a battery of tests

For PIA-6 the 240 items were distributed into twelve tests of twenty items each. However, the distribution was not strictly random as required by item-sampling. It was deemed necessary to control the assignment of items to tests. Each test was then assigned randomly at each testing period to a subset of pupils in the study group.[2]

4. Periodic administration of the battery

For PIA-6 the main instruction occurs twice weekly and is highlighted by a short TV arithmetic program. In all there are 64, 15 minute programs designed to be shown twice weekly. Teachers have definite responsibilities before, during, and after each TV program. Together, the pretelecast, telecast, and follow-up activities provide a concentrated treatment. For this study data were collected four (approximately equally spaced) times during the year. The entire PIA-6 arithmetic program can be viewed as 64 distinct programs or treatments. Figure 1 illustrates the design of the periodic testing procedure.

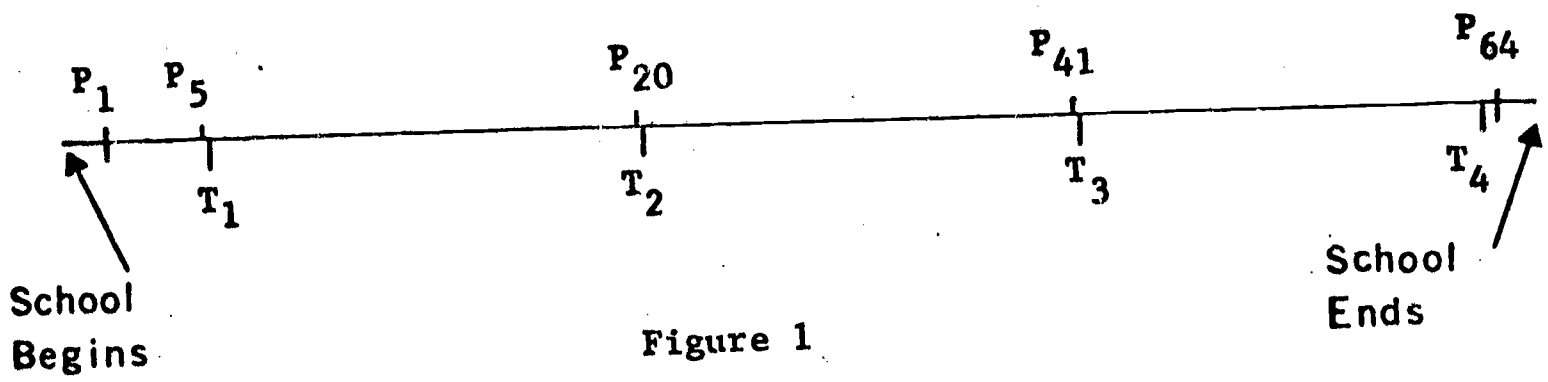


Figure 1

Testing Schema

---

[2] One of the tests is in the appendix.

The horizontal line represents the school year. Above this line P_1- P_{64} represent the 64 programs provided by PIA-6. Below the line T_1, T_2, T_3, and T_4 represent the four testing periods as they occurred during the year. T_1 followed program 5, T_2 followed Program 20, T_3 followed Program 41 and T_4 followed Program 63. In general, the frequency and spacing of assessments could be varied to meet the demands of the developer. Data was gathered on 1432 students from 57 classes in medium to small rural communities in Wisconsin.

Construction of Item and Objective Profiles

Since the same test items are used at each testing period a profile of item change across the year can be constructed. (Campbell and Stanley, 1963). If instruction is having an effect, one would expect a discontinuity in the measurements made at T_1, T_2, T_3, and T_4.

For PIA-6 profiles were plotted for each item. Figure 2 illustrates the item profile for item 14 on Test 11. The four testing periods are represented along the horizontal axis and the percent responding correctly (item difficulty) is plotted along the vertical axis. The triangular region(s) along the horizontal axis represents major program coverage at that point during the school year. The location of the triangle is a good approximation

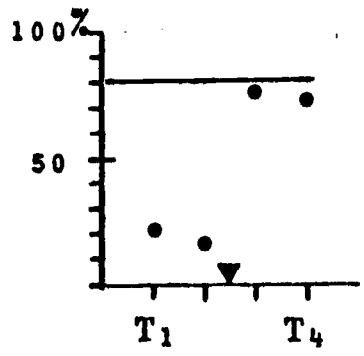


Figure 2

Growth Profile for
Item 14, Test 11

of where topic coverage occurred relative to the four testing periods. For this example, the item was very difficult at T_1 and T_2 since there was no attention given to the problem by the program. However, at T_3 there is growth as a result of extensive coverage following T_2. The horizontal line at 80% represents a criterion level associated with this item. This item was slightly below criterion T_3 and T_4.

For PIA-6 it seemed reasonable to set levels of expected performance on each item or set of items. While it was difficult to set a unique criterion level for each item, it was reasonable to classify items and to set a lower-bound criterion for each classification. Items used in the formative evaluation of PIA-6 were classified in one of five types (Romberg, in press).

(1) Mastery Level A Items in this category are expected to be be very easy by the end of the year. Most every pupil should have mastered the content of the item.

(2) Mastery Level B These items represent topics which receive major emphasis during the year. While items in this category test important objectives, a very high level of mastery is not expected (The majority of PIA-6 items fall in this category.)

(3) Mastery Level C Items in this category represent more complicated aspects of content covered. Story problems which lead to involved computations as well as problems which are conceptually difficult for the average pupil are C-Level items.

(4) Transfer Level X Transfer Level X items involve a minor extension of concepts. For example, the introduction of new notations or a problem which requires some insight belongs in the X category.

(5) Transfer Level Y These items are the more difficult ones used in the testing. Such items are usually conceptually difficult, and represent an extension of program content.

Before these classifications can be used to interpret results they must be quantified. Quantification of each classification consisted of placing a lower-bound criterion on the item difficulty (percent responding correctly). Table 2 indicates the lower-bound criterion for item difficulties for each classification level which reflect the aspirations of the PIA staff for performance at the end of Grade 6. Multiple-choice and free response items were considered separately since the former type involves an element of chance.

TABLE 2

LOWER-BOUND CRITERION FOR EACH CLASSIFICATION LEVEL

Level	Lower-Bound	
	Multiple Choice	Free Response
A	85	80
B	65	60
C	40	30
X	60	50
Y	35	10

Although the criterion levels of items is useful and informative, they are nevertheless arbitrary. Performance on an item may not reach that level

for several reasons:

1. The item was measuring a skill other than the one intended.
2. Poor instruction.
3. Coverage related to the item was not as originally planned. If items are chosen before detailed planning, it is possible that some topics will not get the intended treatment. Items related to such topics will naturally be more difficult than anticipated.
4. Criterion was set too high.
5. The estimate of item difficulty is subject to sampling error.

Some of the same reasons may also explain why an item reachs criterion.

Item profiles are obviously useful in evaluating the effectiveness of the program with respect to the item considered. The baseline data provided by testing at T_1 can be used to determine the effects of the intervening treatments. Moreover, if the treatment related to a given time occurs between T_2 and T_3, then data from T_1 and T_2 provide baseline data to compare the results from T_3. The item profiles were used by the PIA-6 staff in the following manner. Typically, a meeting was held a few days after a testing period and the results of the testing were reviewed and interpreted. Items related to topics covered by intervening programs were the focus of discussion. Skills and concepts which had been covered earlier in the year were also examined to see how well they were being retained.

Results of having this data undoubtedly functioned in subtle ways that were not always observable. However, in some cases minor revisions were made after considering the data. Braswell (1969) reports many such examples.

Even though for PIA-6 item profiles proved to be useful, for most formative evaluations objective profiles would be preferred. Clearly decisions

made on the basis of information from a single item could be unreliable.

The following example illustrates what could be done with data on a topic or an objective. For the topic titled "number line" five items were administered. The percentage correct for each item at each administration is shown in Table 3.

Table 3

PIA-6 RESULTS BY CONTENT AREA: NUMBER LINE

Location*	Content	T_1		T_2		T_3	T_4
(1,4)	Name 1 3/4 on number line	.78	r**	.81	r	.86	.89
(5,2)	Betweenness on number line	.64	x	.75		.76	.70
(7,5)	Name point B? 11/4	.43	x	.68	r	.83	.63
(11,12)	Which number nearest zero? 1/16	.74	r	.78	c	.89	.88
(12,4)	Name a point on number line. 5/8	.23	r	.23	c	.26	.34
	Estimated means	.56		.65		.72	.69

* Location (a,b), a = test number, b = item number

** Coverage between test periods, x = extensive, c = some, r = review, blank = none

Extensive coverage of the content related to the number line showed substantial increase in performance on only two items between T_1 and T_2. Further coverage and review improved scores on four of the five items.

From the PIA-6 experience the use of this data gathering procedure, achievement monitoring via item-sampling, provided useful information for planning and revision. Using this procedure it was possible to monitor simutaneously many dimensions of the curriculum. Moreover, the item profile provided a "history" of the item across the year. Effective revisions were using that information.

Comments on the Procedure

Using this procedure with PIA-6 provided the staff with valuable information about PIA-6 and about the procedure itself. If it were to be used again,[3] a number of things must be more carefully considered.

1) Decision rules. Although evaluation involves both data collection and decision rules, only a procedure for data collection has been discussed. Another study done at the University of Wisconsin R & D Center has addressed this problem (see Kriewall, 1969).

2) Item validity. Better procedures for collecting items and validating them should be followed (see Cronbach, 1969).

3) Periodic Administration. To be most useful tests sould be administered more often than was done in PIA-6. Also, data from the previous school year and the following school year would be helpful.

4) Profiles. Profiles (based on several items) could be generated for behavioral objectives. If so, then tests of significance could be used to determine effects of instruction. (See Campbell and Stanley, 1963, p. 42-43).

---

[3] It should be noted that formative evaluations are unreplicatable. Instruction programs are only developed once.

BIBLIOGRAPHY

Braswell, James. The Formative Evaluation of Patterns in Arithmetic Grade 6 Using Item Sampling. Technical Report 113. Madison, Wisconsin: University of Wisconsin, Research and Development Center, 1970.

Cahen, L. S., Romberg, T. A., and Zwirner, W. The estimation of mean achievement scores for schools by the item sampling technique. Educational and Psychological Measurement, Spring, 1970, pp. 41-60.

Campbell, Donald T. and Stanley, Julian C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1963.

Cronbach, Lee J. Validation of educational measures. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, revision in press.

Hively, Wells; Patterson, H. L.; and Page, S. H. A "universit defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, pp. 275-290.

Kriewall, Thomas E. Applications of Information Theory and Acceptance Sampling Principles to the Management of Mathematics Instruction. Technical Report 103. Madison, Wisconsin: University of Wisconsin, Research and Development Center, 1969.

Lord, Frederic M. Estimating norms by item sampling. Educational and Psychological Measurement, 22, 1962, pp. 259-267.

Messick, Samuel. The criterion problem in the evaluation of instruction: assessing possible, not just intended outcomes. Symposium presented at the Wisconsin Research and Development Center, Madison, November 1969.

Romberg, T. A. Evaluating School Mathematics. Charles Merrill Company, Columbus, Ohio, in press.

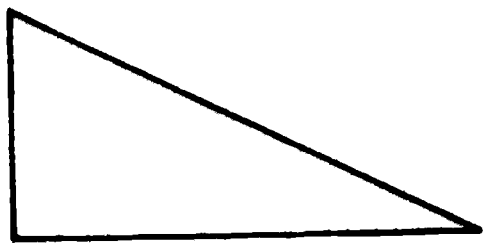
Stake, Robert E., Toward a technology for the evaluation of educational programs. AERA Monograph series on Curriculum Evaluation, Perspectives of Curriculum Evaluation, 1, 1967, pp. 1-12.

PATTERNS IN ARITHMETIC
Grade 6
9

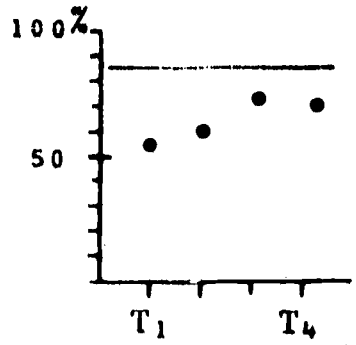
QUESTION SHEETS

Instructions: You may write anywhere on the question sheets. Questions 1-13 are multiple choice. You should decide which choice is correct and <u>circle</u> your choice on the <u>answer sheet</u> provided.

You should have enough time to work on every question. Do <u>not</u> spend too much time on any problem.

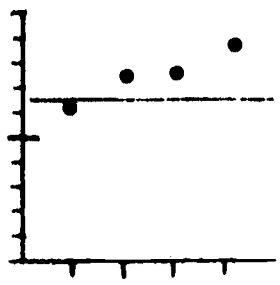


1. The above figure is a
   a) circle
   b) rectangle
   c) square
   d) triangle
   e) parallelogram



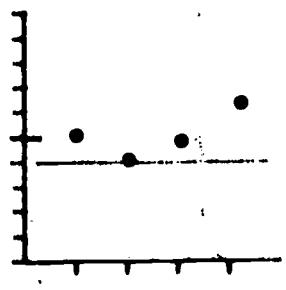
****

2. Which choice below is equal to $\frac{15}{8}$?
   a) $\frac{1}{8} + \frac{5}{8}$
   b) $1\frac{5}{8}$
   c) $7 + \frac{1}{8}$
   d) $1\frac{7}{8}$



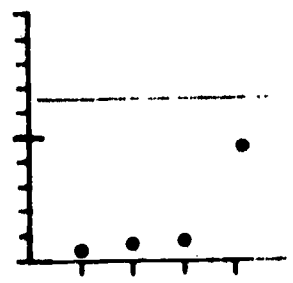
****

3. If the scale length of $4\frac{1}{2}$ inches represents an actual distance of 72 miles, how many miles does the scale length of 7 inches represent?
   a) 2
   b) 56
   c) $74\frac{1}{2}$
   d) 112
   e) 504

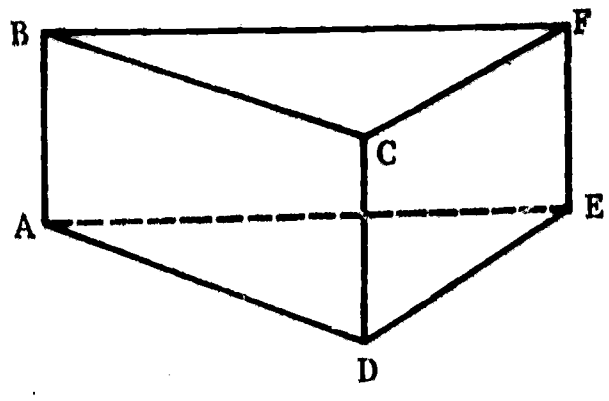


****

4. Which number is the greatest?
   a) 0.03
   b) 0.29
   c) 0.293
   d) 0.2093



****



. In the above prism, the back face is named by

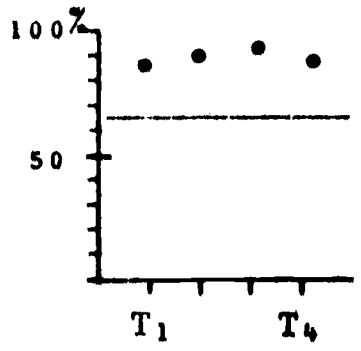
a) BCF

b) ABFE

c) CDEF

d) ADE

****



. Which is the least common denominator for $\frac{1}{6}$ and $\frac{1}{8}$?

a) 6

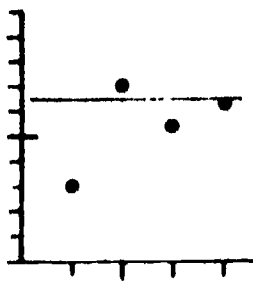
b) 8

c) 14

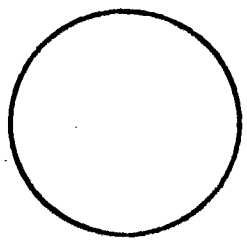
d) 24

e) 48

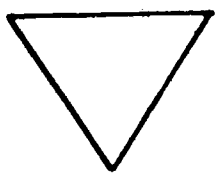
****



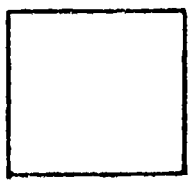
I.



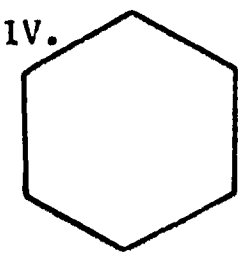
II.



III.



IV.



7. Which of the above figures has more than 6 lines of symmetry?

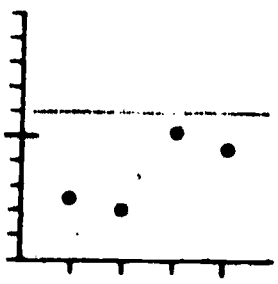
a) I

b) II

c) III

d) IV

****



8. In a football game the Red Team was penalized 15 yards and on the next play passed for a gain of 11 yards. Which sentence tells what happened to the Red Team on the two plays?

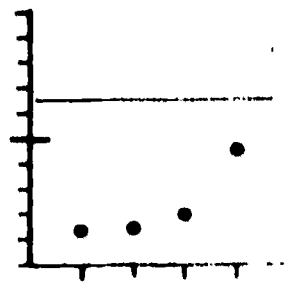
a) 15 + 11 = n

b) 15 - 11 = n

c) -15 + 11 = n

d) 11 - 15 = n

****



9. 3 is what percent of 6?

a) 0.5

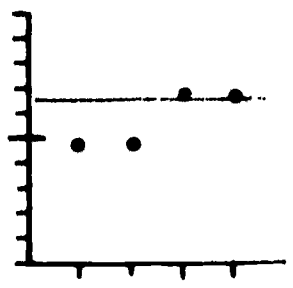
b) 2

c) 3

d) 50

e) 200

****



10. Jane is going to make cookies. She rolls out the dough and starts to cut out the cookies. Jane has 3 shapes of cookie cutters - a circle, a square, and a star. If each cutter has the same area, which cookie cutter would probably give the most cookies after the dough is rolled out once?

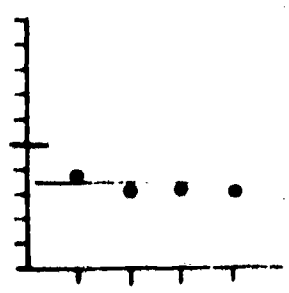
a) the circle

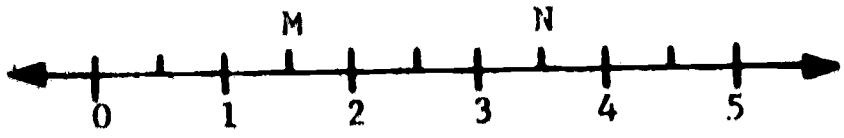
b) the square

c) the star

d) all the same

****





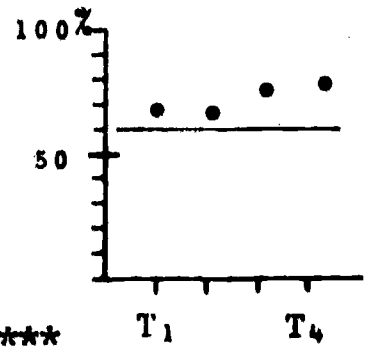
11. What is the distance from M to N on the number line?

a) 2

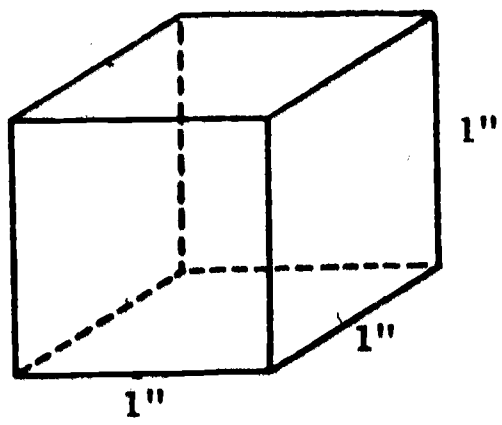
b) $1\frac{1}{2}$

c) 3

d) $3\frac{1}{2}$



****



12. If E = number of EDGES
F = number of FACES
V = VOLUME

What is (E,F,V) for the cube above?

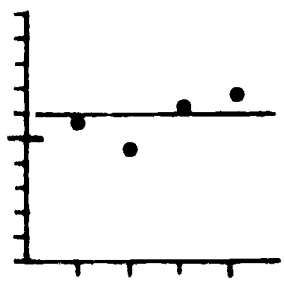
a) (9,4,3)

b) (12,6,1)

c) (9,6,1)

d) (12,4,3)

****



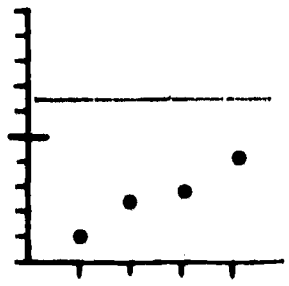
13. What is the area of a square $\frac{1}{2}$ inch on a side?

a) $\frac{1}{2}$ square inch

b) 1 square inch

c) $\frac{1}{4}$ square inch

d) 4 square inches



****

GO TO NEXT PAGE

Work the remaining problems on scratch paper or beside the problem and place your answer on the answer sheet in the space provided.

14. 0.6 x 30 = ?

15. $7\frac{1}{2} + 5 = ?$

16. 19 x 2010 = ?

17. $\frac{35}{36} + \frac{1}{36} = ?$

18. Oranges cost 79¢ a dozen. To the nearest penny, how much would one orange cost?

19. If the ratio of $\frac{\text{grams}}{\text{pounds}}$ is $\frac{453}{1}$ then 3 pounds is how many grams?

20. A mouse has two walls to go through and several holes. If he starts at point A, how many routes can he take to get to point B? The dotted line shows one possible way.