DOCUMENT RESUME

ED 042 796 TM 000 054

AUTHOR Pinsky, Paul; Gorth, William P.
TITLE Descriptive Analysis of KA442: One-Semester, Eleventh & Twelfth Grade Trigonometry.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY Charles F. Kettering Foundation, Dayton, Ohio.
REPORT NO TM-22
PUB DATE Jul 69
NOTE 19p.

EDRS PRICE EDRS Price MF-$0.25 HC-$1.05
DESCRIPTORS Achievement Tests, Cognitive Tests, Course Objectives, Grade 11, Grade 12, Measurement, *Measurement Instruments, *Measurement Techniques, *Predictive Ability (Testing), Predictor Variables, *Test Construction, *Trigonometry
IDENTIFIERS *Comprehensive Achievement Monitoring& CAM

ABSTRACT

Similar results were obtained here to those reported in TM 000 052: Random and chronological item arrangements yielded equivalent scores; the cognitive tests were again limited predictors of posttest achievement; and there was no consistent pattern as to the learning curve providing the test fit for all students, even though these monitors contained five more items (14) than those used in TM 000 052. In contrast, however, the expected increase in scores did occur on this occasion. Tables of reliabilities are included. (DG)

Project C omprehensive
A chievement
M onitoring

Technical Memorandum No. TM-22 July 1969

DESCRIPTIVE ANALYSIS OF KA442: ONE-SEMESTER, ELEVENTH & TWELFTH GRADE TRIGONOMETRY

by

P. Pinsky and W. Gorth
Stanford University

The research and development reported herein was performed pursuant to a grant from the Charles F. Kettering Foundation to the Principal Investigator, Dr. Dwight W. Allen, Dean, School of Education, The University of Massachusetts. The Project CAM staff includes D. Evans, W. Gorth, P. Pinsky, N. Sims, L. Wightman, and G. Worle.

William P. Gorth
Project CAM
School of Education
Stanford University
Stanford, California 94305

or

David R. Evans
Project CAM
School of Education
The University of Massachusetts
Amherst, Massachusetts 01002

A DESCRIPTIVE ANALYSIS OF KA442
ONE-SEMESTER, ELEVENTH & TWELFTH GRADE TRIGONOMETRY

by
Paul Pinsky & William Gorth
Stanford University

This report contains an analysis of course KA442 which used CAM monitoring during the 1968 school year. The course is a one-semester eleventh and twelfth grade trigonometry course which is taught in the traditional teacher-paced method. While the report is basically of a descriptive nature, several hypotheses were tested. The data in this course behaved as one would expect under a CAM model. They are an excellent example of the basic principles of CAM.

The analysis indicated the following:

1. Random compared with chronological arrangements of items on the monitor forms yielded equivalent scores. The same phenomenon was noted in the course, HS420, but the lack of increase in achievement made this conclusion tentative (see TM-21). However, in this course there was an increase in achievement as expected. Therefore, the conclusion of no difference between random and chronological arrangement of items on the monitor forms has been more strongly substantiated. Thus, a chronological arrangement of items may be permissible for manual data processing.

2. As in HS420, it was very clear that individuals' total scores should not be compared in the CAM model as presently used.

3. As in the course, HS420, the cognitive ability test scores were not most useful in scheduling students to take various monitor forms; i.e., the scores were limited predictors of posttest achievement. Actually, the students were scheduled to take monitor forms based upon a grade point average in related courses in high school.

This scheduling procedure appears to be quite adequate for the CAM model.

4. A change in student scores throughout the semester behaved as expected; i.e., there was an increase in scores. Moreover, we were able to calculate an exact significance test for the change in scores from one period to another.

Scheduling Pattern

Nine sets of items, fourteen items per set, were used. In addition, there was a 32-item pretest and a 32-item posttest. Each of the nine sets of fourteen items were arranged two different ways to make a total of eighteen distinct test forms. One arrangement of items was random and the second was chronological relative to the presentation in the classroom of content they measured.

The class was divided into two groups. One took only the random arrangement and the other only the chronological. Each student took each set of items once throughout the semester. Each set of items was taken by the same number of students each testing period. The stratification of students was based on their mathematics grade point average in mathematics. Approximately nine students took each test form each test administration.

Scheduling Procedures

In the CAM project, student schedule groups are groups of students scheduled to take the same pattern monitor forms during the year and are selected by using stratified random sampling. The stratification of students was based on their grade point average in mathematics. There were a total of eighteen different schedule groups with approximately nine students per schedule group. The average of the students in each of the eighteen schedule groups on the posttest examination is presented, in Table TM-22.1, as indications of their post-course achievement and the adequacy of the stratification.

INSERT TABLE TM-22.1 ABOUT HERE

With the exception of schedule group 23, all schedule groups are about equal in ability. It appears that the stratification of students based on grade point averages, in similar courses, is useful.

Nine reference tests of cognitive ability (TM-18 and French, et al, 1963) were administered during the semester and their potential for usefully stratifying students was investigated. A correlation of the nine ability scores with the posttest score reveals the predictive power of these ability tests in this course. These correlations are presented in Table TM-22.2.

INSERT TABLE TM-22.2 ABOUT HERE

A step-wise regression was used to predict the final test score from the nine ability test scores. Eight of the nine covariates (the last one not being significant) yield a multiple correlation coefficient of .430. As a set the cognitive ability tests are moderate predictors of posttest achievement.

Monitor Form Characteristics

It was initially hoped that CAM would provide information about the progress of individual students. However, our analysis indicates that <u>total test scores</u> of individual students should not be used in their present form without <u>extreme caution</u>. Some analyses were performed on the characteristics of the total score on the different monitor forms.

The first analysis considered the difficulty level of test forms as measured by the total number of items answered correctly.

Table TM-22.1 Mean Posttest Scores of Students in Each Schedule Group

Schedule group (random forms)	Mean posttest score	Chronological forms	Mean posttest score
23	14.9	33	21.3
24	20.6	29	21.5
26	22.0	37	20.4
27	25.3	21	21.5
31	25.1	25	21.4
34	24.4	30	24.7
35	23.7	22	22.6
36	24.7	32	20.0
38	22.2	28	22.4
Grand mean	22.6	Grand mean	21.8

Note.-- Test score measured in number of questions answered correctly.

Schedule group designated by the first test form students took of their assigned sequence of tests for the semester. Schedule groups assigned tests containing the same items but arranged in either random or chronological order appear in the same row.

Table TM-22.2 Correlation of Posttest of Achievement with Cognitive Ability Test Scores

Number[a]	Test name[b]	Correlation
1	Wide Range Vocabulary	.360
2	Number Comparison	.008
3	Surface Development	.186
4	Cube Comparison	-.001
5	Letter Sets	.022
6	Word Arrangement	.038
7	Inference	.261
8	Maze Tracing	.074
9	Auditory Number Span	.001

Note.-- N = 125

[a] Tests listed in the order in which they were administered.

[b] Tests taken from French, et. al. (1963) and are described in TM-18.

All test forms were approximately the same difficulty. Table TM-22.3 presents the average number correct over all test administrations of each of the 18 test forms.

INSERT TABLE TM-22.3 ABOUT HERE

The second analysis considered a test form by test administration interaction; i.e., whether test forms change in difficulty over time. The model of a general linear hypothesis was used:

$$Y_{ijk} = N + f_i + P_j + (fP)_{ij} + \beta(X_{ij} - \bar{x}) + e_{ijk};$$

where

$i = 1, \ldots, 5;\ j = 1, \ldots, 9;\ k = ., \ldots, n_{ij};$

Y_{ijk} are number of correct responses on form f_i in period P_j for student k;

f_i are the monitor forms;

P_j are the monitoring periods;

$(P_1 = 2,\ P_2 = 3,\ \ldots,\ P_8 = 9,\ P_9 = 10);$

X_{ij} are average score on the posttest examination of all students who took form f_i in period P_j;

n_{ij} are number of students who took form f_i in period P_j;

and $(7 \leq n_{ij} \leq 10)$.

The calculations were made by the computer program, BMD05V (Dixon, 1968). The program allows a maximum of 60 independent and dependent variables in the model. Eighteen test forms times 9 test administrations exceeds the program maximum for independent variables. Therefore, the pair of test forms which contained the same items arranged differently were considered the same, reducing the number of variables by a factor of two. Further, two overlapping sets of test forms, five in each group, were analyzed. In Set A, f_1 = test forms 26 and 37; f_2 = test forms 35 and 22; f_3 test forms 27 and 21; f_4 = test forms 36 and 32; and f_5 = test forms 38 and 20. In Set B, f_1 =

Table TM-22.3 Mean and Variability of Scores of All Students who took Each Test Form over All Test Administrations

Test form number		Mean score[a]		Measure of variability[b]	
Random	Chrono-logical	Random	Chrono-logical	Random	Chrono-logical
23	33	4.8	5.0	5.21	3.74
24	29	5.3	5.2	4.84	6.47
26	37	5.6	5.3	5.59	4.96
27	21	5.1	5.0	4.41	2.92
31	25	5.2	5.0	5.75	5.12
34	30	4.4	4.5	3.91	5.25
35	22	5.0	5.5	3.13	3.88
36	32	4.8	4.4	6.61	5.68
38	28	4.8	4.5	6.65	5.27
ALL FORMS		5.0	4.9		

Note.--Approximately 8 observations per form per period yields about 72 observations per form for the semester.

[a] Mean score of all students who took the form over all periods.

[b] N_{jk} = # of students who took form k in period j.

X_{ijk} = score of student i during period j on form k.

$\frac{1}{N_{jk}} \sum^{n_{jk}} X_{ijk} = \bar{X}_{.jk}$ = average score of students on form k in period j.

$V_{jk} = \frac{1}{n_{jk}-1} \sum_{i=1}^{n_{jk}} (X_{ijk} - \bar{X}_{.jk})^2$ = variability of student's scores on form k in period j.

$\bar{V}_{.k} = \frac{1}{n_{.k}-1} \sum_{j} \sum_{i} (X_{ijk} - \bar{X}_{.jk})^2$ = variability of form k over all periods and students.

test forms 24 and 29; f_2 = test forms 31 and 25; f_3 = test forms 27 and 21; f_4 = test forms 23 and 33; and f_5 = test forms 34 and 30. The posttest score of the student was used as the covariate in the model. The hypotheses tested were:

hypothesis 0 is the full model;

hypothesis 1 is $(fP)_{ij} = 0$;

hypothesis 2 is $(fP)_{ij} = 0$ and $f_i = 0$;

hypothesis 3 is $(fP)_{ij} = 0$ and $\beta = 0$; and

hypothesis 4 is $(fP)_{ij} = 0, \beta = 0$, and $f_1 = 0$.

The results are displayed in Table TM-22.4.

INSERT TABLE TM-22.4 ABOUT HERE

The interaction effect is significant in Set B and not significant in Set A at the 99% level. The posttest score is highly significant in predicting how well the student did throughout the year, suggesting a consistency in their performance.

Test-retest reliability is calculated for each pair of test administrations. The reliability coefficients were calculated assuming that all the monitor forms contained the same items.

INSERT TABLE TM-22.5 ABOUT HERE

The increase over time in standard deviation and standard error of measurement reflects the increase in students' scores from zero to slightly above 50 per cent.

The experimental design for this course was to ascertain the effect of different arrangements of items on the test forms. An analysis did not identify any effect. Similar results were found in the course HS420 (TM-21). Table TM-22.3 presents the average number correct for each test form over the whole year.

Table TM-22.4 Analysis of Variance by Test Form and Test Administration

Set	Hypothesis	ss	df	F
A	0	1884	537	
	1	1997	569	1.00
	2	2074	573	1.50
	3	2731	570	7.31*
	4	2817	574	7.19*
B	0	1954	530	
	1	2181	562	1.93
	2	2208	566	1.92*
	3	2784	563	6.83*
	4	2818	567	6.34*

* $p < .01$

Table TM-22.5 Characteristics of Tests Across All Forms for Each Test Administration

Test administration	Test-retest reliability	Standard deviation	Standard error of measurement
2	.31	1.52	1.26
3	.55	1.45	.97
4	.42	1.85	1.41
5	.54	1.88	1.28
6	.50	2.16	1.53
7	.49	2.10	1.50
8	.57	2.37	1.56
9	.40	2.41	1.87
10			

Note.-- Test-retest reliability, its standard deviation, and standard error of measurement are calculated from test administration n to n+1 and are recorded in the row for test administration n.

Graphs of the total test score for the nine test administrations of the random and chronological arrangements of items were compared. There were no systematic differences in total score between the random and chronological arrangements of items on test forms. Also, split-halves internal reliability coefficients were calculated at each test administration across all test forms with either the random or chronological arrangement of items, Table TM-22.6.

INSERT TABLE TM-22.6 ABOUT HERE

Table TM-22.6 shows that the split-halves reliability coefficients of the chronological tests are consistently higher than the random tests. One would expect this phenomenon under the CAM Model because on the chronological arrangement of items, students identify items which they can answer and those they cannot. The answerable items are always at the beginning of the test and the unanswerable ones at the end. Therefore, the split-halves reliability coefficients are high. The random arrangement of items distributes answerable items throughout the tests, thus lowering the coefficients until the end of the year when all the items are answerable.

Achievement profiles were calculated for the students who took only randomly arranged tests or chronologically arranged tests. Eight achievement profiles were calculated for each group; i.e., all questions separately, and questions in each unit; i.e., one through seven. Table TM-22.7 presents the achievement profile for unit 5. They are essentially identical.

INSERT TABLE TM-22.7 ABOUT HERE

An analysis of the pretest and posttest was made. All students took the same 32-item test, although the pretest was different from the posttest. The split-halves reliability coefficients were .49 for the

Table TM-22.6 Split-halves Reliability Coefficients Across All Test Forms for Each Test Administration

Test administration	Item arrangement	
	Random	Chronological
2	.36	.44
3	.04	.32
4	.05	.43
5	-.21	.49
6	.29	.44
7	.11	.48
8	.18	.69
9	.28	.42
10	.58	.48

Note.-- All test forms contain fourteen items.

Table TM-22.7 Achievement in Unit 5 for Students Taking Tests with Random and Chronological Arrangements of Items at Each Test Administration

Test administration	Arrangement of Items Taken by each student group[a]	
	Random	Chronological
2	2	0
3	2	0
4	2	2
5	4	2
6	6	4
7	10	14
8	26	24
9	42	48
10	48	42

[a] Achievement measured as per cent items answered correctly.

pretest and .86 for the posttest. The average number of correct responses on the pretest was 2.01; the average number on the posttest was 22.16. The standard error of measurement was 1.51 for the pretest; 2.56 for the posttest.

Positional Effects

Project CAM has attempted to determine whether student fatigue or warmup effects were affecting results. The objective was to try to determine an optimal length of the monitor forms in the CAM system. The analysis considered the forms in which the items were randomly arranged and summed the total number of correct responses across these forms. It was performed and indicated no consistent pattern.

Individual Differences

It was hoped that measures of individual student performance could be obtained from the CAM system. However, having already seen the standard error of measurement, one should be cautious. Nevertheless, we attempted fitting various learning curves to the data. The computer program, BMD05R (Dixon, 1968) was used to fit a first, second, and third degree curve to the total number of correct responses for each of the students across all the test administrations. A sample of ten students who had completed all tests in the course was taken and three different types of learning curves were fit. First, a learning curve was fit to the raw data; that is, the total number correct. Then a learning curve was fit to the data modified to reflect overall test form difficulty. And finally the data were modified to reflect average test form difficulty on a period by period basis. Neither of the two modifying procedures appeared to improve quality of learning curve. As in the course HS420 (TM-21), it was subjectively observed that there was no consistent pattern as to whether a linear, quadratic, or cubic curve was the best fit for all the students.

As a further analysis to attempt to attribute some meaning to this curve fitting, a correlation analysis was run between the following variables which are calculated for each student: pretest score, post-test score, 0 to 60 day criterion score (TM-6), -200 to 10 day criterion score, 60 to 200 day criterion score, the average number of items correct over all the periods, the slope of the best fit linear line for the student data, the standard error of this slope, and the change in difficulty of each test form. These correlations are presented in Table TM-22.8.

INSERT TABLE TM-22.8 ABOUT HERE

The conclusion is that even using smoothing techniques, such as fitting curves to the modified data of individual students, virtually no meaningful information can be gained about these individual student learning curves when comprehensive monitors containing only 14 items are used.

Group Performance

In contrast to the data in HS420, the group performance parameters behaved as would be expected in the CAM model. A summary of the class performance on the seven units or chapters of the course for each of the nine test administrations is given in Table TM-22.9.

INSERT TABLE TM-22.9 ABOUT HERE

A question of interest when examining this table is when does the change in percentage of correct responses for a given unit from one test administration to the next test administration reflect the true change in the group parameter? Utilizing the theory of item sampling, one can develop a t-statistic to test the significance of the change of these parameters (Husek and Sirotnik, 1968). The t-values were calculated for the data presented in Table TM-22.9. The values which represent a significant change at the 95% level are presented in the table by an asterisk.

Table TM-22.8 Correlations of Various Measures of Student Performance

No.	Source	2	3	4	5	6	7	8	9
1	Criterion score: -200 to -10 days	.31	.28	.26	.28	.58	.49	.14	.11
2	Criterion score: 0 to 60 days		.80	.05	.76	.88	.23	.57	.03
3	Criterion score: 60 to 200 days			.03	.90	.76	.10	.57	.06
4	Pretest				.09	.01	.08	-.06	.25
5	Posttest					.64	.09	.48	.06
6	Mean[a]						.37	.56	.03
7	Intercept[b]							-.53	-.08
8	Slope[b]								-.04
9	Correlation coefficient[c]								

Note.-- N = 126.

[a] The mean is the average number of correct responses per period over all 9 periods.

[b] The intercept and slope are the values of the least squares line fit through the number of correct responses per period.

[c] The correlation coefficient is a measure of the variability of the data about the least squares line.

Table IM-22.9 Percentage of Correct Responses by Unit and Test Administration

Test administration	Unit 1	2	3	4	5	6	7
Pretest	8	10	4	6	2	3	6
2	59	28	18	6	2	5	4
3	*68	*47	16	6	1	3	6
4	65	*58	*38	7	3	4	7
5	68	61	*63	*16	4	7	6
6	63	55	59	*36	6	5	8
7	*76	63	59	*52	*13	9	11
8	77	69	*73	55	*27	12	15
9	73	71	74	57	*46	15	13
10	79	71	65	54	45	*48	*33
Posttest	81	75	72	66	59	47	47

Note.-- Each cell in the table contains more than 220 observations.

* = significant change at the 95% level.

References

Dixon, W.J. (Ed.) Biomedical computer program: BMD. Berkeley: University of California Press, 1968.

French, J.W., Ekstrom, R.N., & Price, L.A. Manual for a kit of reference tests for cognitive factors. Princeton: Educational Testing Service, 1963.

Gorth, W. & Pinsky, P. Technical memorandum no. TM-18: Demographic, aptitude, & attitude surveys of the students, teachers, and schools in Project CAM. Stanford, Cal.: Project CAM, School of Education, 1968.

Husek, T.R. & Sirotnik, K. Item sampling in educational research. Los Angeles: Center for the Study of Evaluation, 1967.

Lindeman, R., Gorth, W., & Allen, D. Technical memorandum no. TM-6: The evaluation of item performance in an item sampling case. Stanford, Cal.: Project CAM, School of Education, 1968.

Pinsky, P. & Gorth, W. Technical memorandum no. TM-21: Descriptive analysis of HS420, eleventh grade algebra, first semester. Stanford, Cal.: Project CAM, School of Education, 1969.