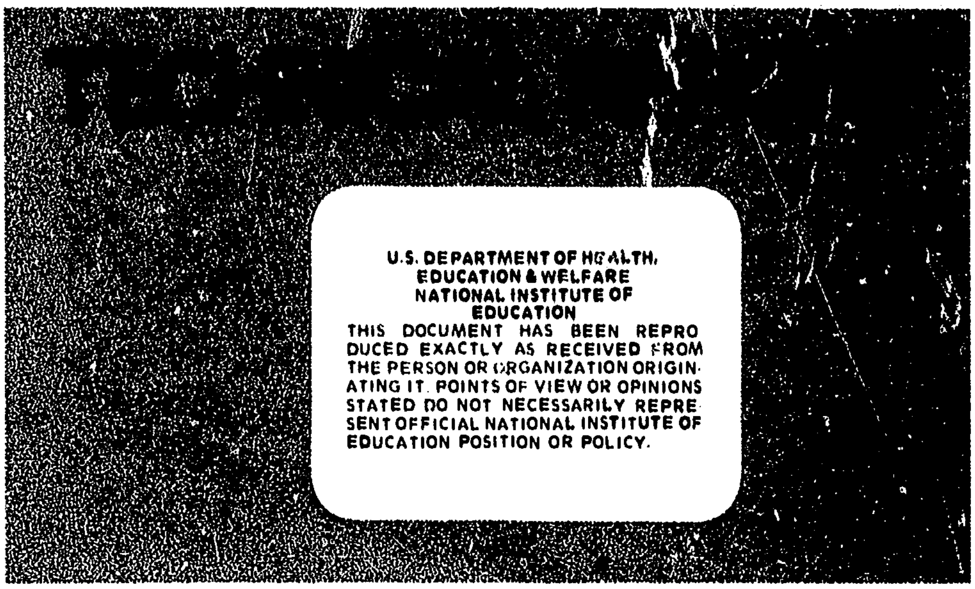
DOCUMENT RESUME

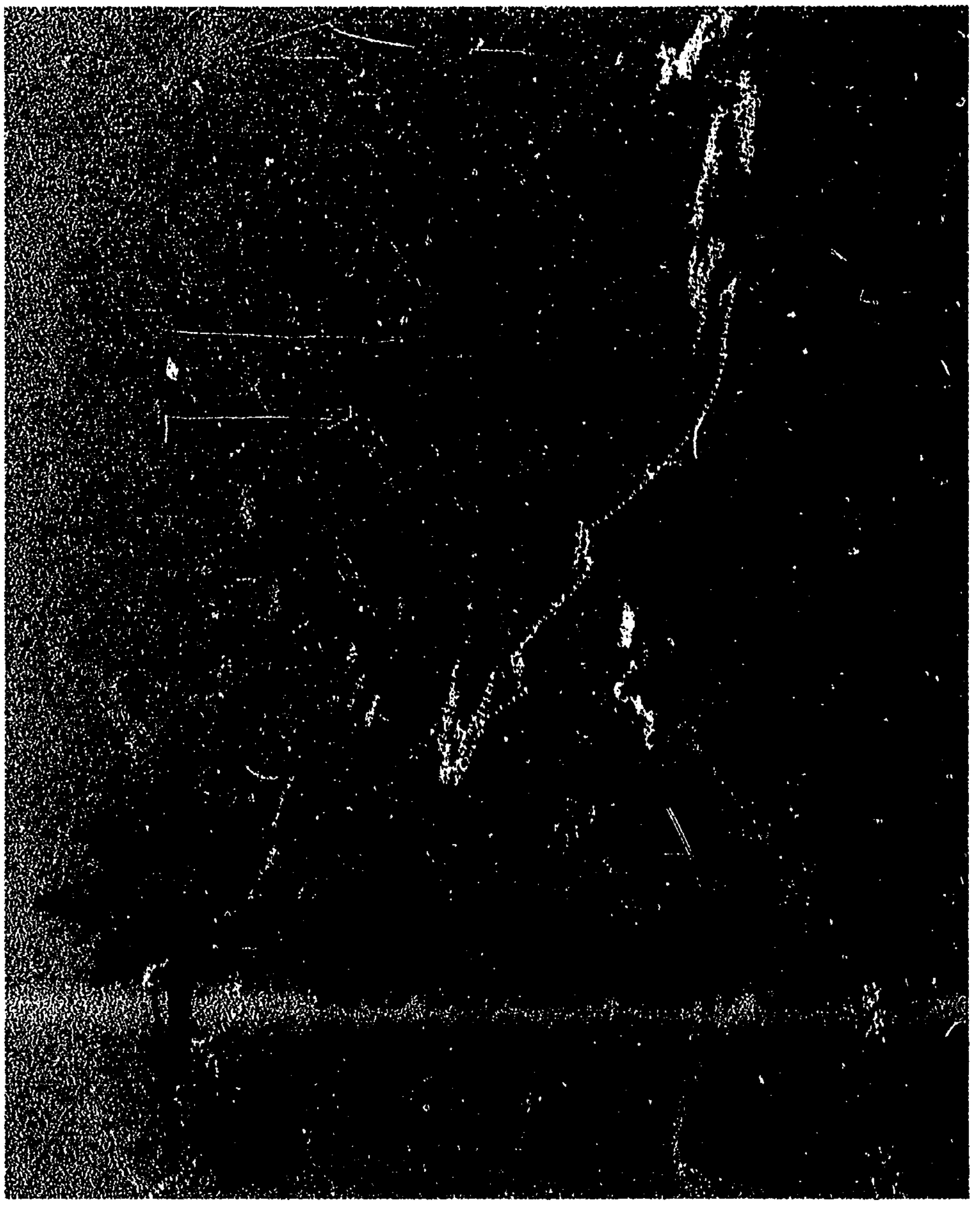
ED 100 719 SE 018 750

AUTHOR	Zalewski, Donald L.
TITLE	An Exploratory Study to Compare Two Performance Measures: An Interview-Coding Scheme of Mathematical Problem Solving and a Written Test. Part 2. Technical Report No. 306.
INSTITUTION	Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	WRDCCL-TR-306
PUB DATE	Aug 74
CONTRACT	NE-C-00-3-0065
NOTE	125p.; Report from the Project on Conditions of School Learning and Instructional Strategies. For Part 1, see SE 018 749
EDRS PRICE	MF-$0.75 HC-$5.40 PLUS POSTAGE
DESCRIPTORS	Cognitive Measurement; *Cognitive Objectives; Cognitive Tests; Data Analysis; Doctoral Theses; Grade 7; *Mathematics Education; *Problem Solving; *Research; Secondary School Mathematics; Tests; *Test Validity

ABSTRACT

In this study the researcher investigated the feasibility of constructing a valid paper-and-pencil measure of problem solving ability. (Rationale and design of the study are discussed in Part 1.) The principal feasibility criterion, correlation of at least .71 with scores on taped and coded individual "thinking aloud" problem-solving sessions, was not met; however, the obtained correlation (.68) for one test suggested to the researcher that more reliable tests might achieve the criterion. Rank ordering of subjects on the "thinking aloud" procedure and written tests were highly correlated. The use of the "thinking aloud" procedure to establish concurrent validity was evaluated and questions about the validity of this procedure with seventh-grade students were raised. Investigations of the functional differences between audiotaped and videotaped interviews revealed no differences in subject performance, but supported the superiority of videotaping as a research tool. Instruments used in the study and data displays are presented in appendices to this report. (SD)





U.S. Office of Education Center No. C-03 Contract OE 5-10-154

Technical Report No. 306 (Part 2 of 2 Parts)

AN EXPLORATORY STUDY TO COMPARE TWO PERFORMANCE MEASURES: AN INTERVIEW-CODING SCHEME OF MATHEMATICAL PROBLEM SOLVING AND A WRITTEN TEST

Report from the Project on Conditions of School Learning and Instructional Strategies

By Donald L. Zalewski

Thomas A Romberg and John G. Harvey
Principal Investigators

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

August 1974

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that agency should be inferred.

Center Contract No. NE-C-00-3-0065

STATEMENT OF FOCUS

Individually Guided Education (IGE) is a new comprehensive system of elementary education. The following components of the IGE system are in varying stages of development and implementation: a new organization for instruction and related administrative arrangements; a model of instructional programing for the individual student; and curriculum components in prereading, reading, mathematics, motivation, and environmental education. The development of other curriculum components, of a system for managing instruction by computer, and of instructional strategies is needed to complete the system. Continuing programmatic research is required to provide a sound knowledge base for the components under development and for improved second generation components. Finally, systematic implementation is essential so that the products will function properly in the IGE schools.

The Center plans and carries out the research, development, and implementation components of its IGE program in this sequence: (1) identify the needs and delimit the component problem area; (2) assess the possible constraints--financial resources and availability of staff; (3) formulate general plans and specific procedures for solving the problems; (4) secure and allocate human and material resources to carry out the plans; (5) provide for effective communication among personnel and efficient management of activities and resources; and (6) evaluate the effectiveness of each activity and its contribution to the total program and correct any difficulties through feedback mechanisms and appropriate management techniques.

A self-renewing system of elementary education is projected in each participating elementary school, i.e., one which is less dependent on external sources for direction and is more responsive to the needs of the children attending each particular school. In the IGE schools, Center-developed and other curriculum products compatible with the Center's instructional programing model will lead to higher morale and job satisfaction among educational personnel. Each developmental product makes its unique contribution to IGE as it is implemented in the schools. The various research components add to the knowledge of Center practitioners, developers, and theorists.

ACKNOWLEDGMENTS

Since this dissertation is the culmination of many years of formal learning, I am deeply indebted to the people who provided the opportunities for my education and to the people who helped me through the investigation. Although there is no way to repay them, I take this opportunity to say "Thank You!"

My first thanks go to the grade school, high school, and college teachers who made learning enjoyable and who encouraged me to continue my education. Thanks also to relatives and friends whose confidence in me often exceeded my own.

I owe special thanks to Professors Thomas A. Romberg, John G. Harvey, and J. Fred Weaver for the uncounted formal and informal occasions when I benefited from their expertise, opinions and advice. My major advisor, Dr. Romberg, was instrumental in formulating this study and was a constant source of encouragement and assistance. Professor Harvey gave generously of his time throughout the study and his editorial comments helped improve my writing style and the readibility of the thesis. Professor Weaver raised constructive questions during the study and served as the third member of the thesis committee.

I also wish to thank Professors Elizabeth H. Fennema and Larry J. Hubert for reading the thesis and contributing their opinions and suggestions.

The Wisconsin Research and Development Center for Cognitive Learning provided the equipment, facilities, and staff to conduct the study and complete this report. In particular, Mr. John McFee provided the technical expertise and manpower for the video taping, Misses Linda Junker and Pat Busk assisted in the data analyses, and Mrs. Ethel Koshalek typed the final draft. I am grateful for their assistance.

Professor John F. Lucas, presently teaching at the University of Wisconsin in Oshkosh, helped me to administer his coding scheme. Mr. Norman Loomer and Sister Ruth Meyer worked diligently as additional coders during the reliability tests. Their willing help made the task easier and I appreciate their efforts.

I am particularly grateful to the students who participated in the study and the teachers who allowed me to disrupt their schedules. Their cooperation made scheduling easy and provided the interesting data and observations recorded herein.

Finally, I thank the people who were most responsible for this dissertation: my parents, wife, and children. As my first teachers, my parents, instilled in me a patience and a faith which proved invaluable. My wife, Rita, encouraged me to continue the studies which resulted in this thesis and shared the moments of inspiration, perspiration, and desperation with me. Our children, Tami and Stephen, patiently understood their father's frequent absence and inattention during the research and the preparation of the manuscript. For their contributions and their confidence in me, I gratefully dedicate this thesis to my parents, wife, and children.

TABLE OF CONTENTS

BEST COPY AVAILABLE

Page

CHAPTER

Page

CHAPTER

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The investigation reported in this thesis is on assessment in mathematics education. Specifically, this study explored the feasibility of using a written test to predict seventh graders' mathematical problem solving achievement as assessed by an interview-coding procedure.

A search revealed that most available mathematical problem solving assessment procedures are commercial tests. The tests do not offer any definitions and their items are usually simple applications or algorithmic situations which do not satisfy the criteria established in this thesis for a mathematical problem.

The method for validly assessing subjects' mathematical problem solving achievement used in this study was a thinking aloud procedure. Interviews yielded audio and video taped protocols, and a coding system permitted classification, analysis, and scoring of the subjects' performances. Because of the complexity of the interview and coding scheme, a written instrument which hopefully had high concurrent validity was developed so that it could be used as a valid alternative to the interview and coding procedure.

Thirty-one seventh graders were asked to think aloud as they tried to solve six mathematical problems in individually taped interviews. The subjects' protocols were coded and scored to provide what

was assumed to be a valid assessment of their mathematical problem solving achievement. The 31 subjects also took two 20 item written tests which were scored by the number of correct responses. Three rankings were developed from the interview test and one ranking was developed from each written test.

The correlation coefficients between the written and interview test scores did not reach the .71 level established for feasibility. One coefficient reached .68 and the tests shared high rank order agreement. These results suggested that a more reliable test might attain the .71 correlation. Clustering and multidimensional scaling verified the structure imposed by the total score ranks.

Other findings indicated that present coding schemes can be applied reliably to describe subjects' problem solving behaviors and that the scoring system permits logical ranking of the subjects. However, serious questions were raised about the validity of the thinking aloud procedure. Video taping the interviews was advantageous because it captured silent indicators of problem solving behaviors and took less time to code.

Chapter VI

DATA AND ANALYSES

Introduction

This chapter presents the data, observations, and analyses from each of the three principal parts of the study. Scores, rankings, and statistics for the written tests are presented first. This is followed by the data of the interview test. The statistical analysis of the relationships of the ranks determined by the written tests and the IT and the results of exploratory statistical procedures conclude the chapter.

The Written Test (WT)

The purpose of the WT was to produce a ranking of the same subjects who were to be ranked by their mathematical problem solving achievement on the IT. The data and statistics for the WT and a subsequent WT2 are presented before feasibility factors are reported. The description of the development of the rankings from the written tests concludes this section.

Subject Response Data

Two classes totaling 63 seventh graders took the 20 item WT. The descriptive statistics for the WT are presented separately in Table 6.1 for the 32 subjects who had been rated below average in mathematics achievement (Group B) and the 31 students who had been rated average or above average (Group A) by their mathematics teachers.

Table 6.1

MEAN, STANDARD DEVIATION, AND RANGE FOR THE WT:
GROUP A, GROUP B, AND COMBINED

	Number of Subjects	Mean	Standard Deviation	Range (20 items)
Group A	31	7.4194	3.8796	2 to 14
Group B	32	3.7500	2.7238	1 to 12
Groups A and B Combined	63	5.5556	3.7963	1 to 14

According to Table 6.1, the results on the WT are consistent with the teachers ratings. Group A, the 31 subjects who also took the IT, averaged 7.4 correct responses to almost double the 38 mean of the lower rated Group B. Group A attained a higher number of correct responses as it ranged from a low of 2 to a high of 14 while Group B ranged from 1 to 12. Figure 6.1 illustrates the distribution of the number of correct responses for each group.

As detailed in Figure 6.1, everyone got at least one correct answer on the WT and no subject got exactly 10 correct responses. In addition, no Group B subject got exactly 8 or 11 items correct and only one subject attained the high of 12 right answers while six subjects achieved one correct response. Group A had three subjects attain the low of two correct while two subjects attained the high of 14 correct answers. The mode for Group A was nine as five subjects reached this score. Group B was bimodal as seven subjects answered two WT

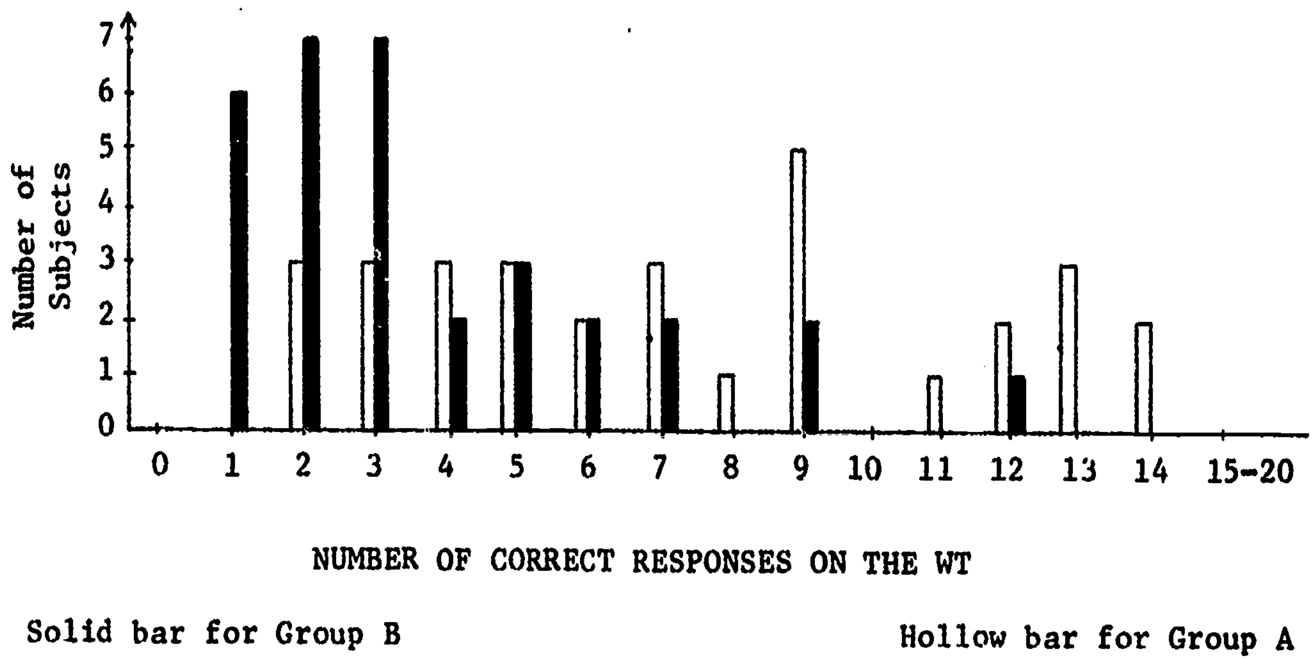


Figure 6.1 Distribution of the Numbers of Correct Responses on the WT

items correctly and seven other subjects gave three correct answers. Some of the differences between the numbers of correct responses between Groups A and B could be attributed to the number of omitted items: Group A subjects skipped an average of 2.7 items on the WT while Group B subjects omitted 4.1 items each.

The low averages and the number of items omitted by the WT subjects caused the investigator to question the representative mathematical ability of the seventh graders selected to participate in the study. In order to compare the subjects to other seventh graders, a second 20 item written test (WT2) was developed from the available pool of items. A random sampling procedure was followed with the restriction that any item which appeared on the WT could not be used on the WT2.

In May, 1974, 350 seventh graders including the original 63 were given the WT2. The investigator administered the WT2 to the 63 subjects (School 1) and a teacher from Des Moines, Iowa, had all three of his seventh grade classes (School 2) take it. One teacher from a middle school (School 3) in Madison, Wisconsin, gave the WT2 to all four of her classes and two teachers from another middle school (School 4) in Madison, administered it to 128 students. The descriptive statistics for the entire group, for each school separately and for the original groups A and B are presented in Table 6.2.

Table 6.2

DESCRIPTIVE STATISTICS FOR THE WT2:
BY SCHOOL, COMBINED, AND BY GROUPS A AND B

	Number of Subjects	Mean	Standard Deviation	Range (20 items)
School 1 (Group A and B)	63	6.1111	3.8189	0 to 16
School 2	66	3.9848	3.7107	0 to 17
School 3	93	4.3978	3.4108	0 to 13
School 4	128	7.9688	4.1544	0 to 19
Schools Combined	350	5.9343	4.1682	0 to 19
Group A	31	8.1290	3.7748	2 to 16
Group B	32	4.1562	2.7133	0 to 9

The results of the WT2 indicate that the subjects in this study (School 1) compared favorably to the other seventh graders who took the WT2. Their average of 6.1 was about half way between the 4.4 average of the Madison School from a low academic achievement area of the city and the 8.0 average of the Madison school from a high achievement area. Group A performed only slightly better than the highest mean of any school and all three Madison schools attained a higher average than the Des Moines school's mean of 4.0. According to the results on the WT2, it appeared that the subjects used in this study were not atypical seventh graders and that their low average on the WT was probably due to the general difficulty that students encountered with the test items.

WT Length and Reliability

The low averages achieved by the students did not affect the feasibility of the WT, but two factors, test length and reliability, were important. A test which took more than an hour to complete or which did not attain a reliability of .80 would not meet the expectations of the investigator. Hoyt's internal consistency measures of reliability for the WT and the WT2 are presented in Table 6.3.

Across the entire sample of students, satisfactory reliabilities of .82 on the WT and .84 on the WT2 were reached. Group A on both tests and School 3 on the WT2 had measure sufficiently close to .80 to be acceptable. Only Group B's reliabilities of .73 on the WT and .68 on the WT2 did not attain the desired minimum. However, since this group did not participate in the IT and the overall reliability for School 1 was adequate, the feasibility of the written test was not jeopardized.

Test length as measured by the time necessary for students to complete the test was a second factor by which feasibility was to be

Table 6.3

HOYT'S RELIABILITY COEFFICIENT FOR THE WT AND THE WT2

	Reliability	Standard Error
WT		
Group A	.7968	1.7045
Group B	.7270	1.3807
Group A & B Combined	.8179	1.5789
WT2		
School 1 (Group A & B)	.8023	1.6551
School 2	.8356	1.4665
School 3	.7897	1.5223
School 4	.8136	1.7484
Combined Schools	.8374	1.6380
Group A	.7737	1.7504
Group B	.6774	1.4921

determined. The investigator recorded the completion times for 59 of the 63 subjects during the WT2, forgetting to note two times. The two other missing subjects had not completed the test during the available class time (37 minutes) and worked during the next period with a second observer who did not record their completion times. Subjects in School 1 averaged 27 minutes to complete the WT2 with one student finishing in 16 minutes and three requiring 37 minutes. The two missing subjects who needed extra time would not appreciably alter the observations which were made. The 27 minute average indicated that School 1 subjects could respond to the 20 items on the WT2 without being rushed. Since School 1 was close to the combined school average in achievement, it was assumed that the completion time averages of other schools would not vary greatly from the 27 minutes. Furthermore, the time average was sufficiently low so that a large deviation such as ten minutes (the maximum recorded) would only produce an average of 37 minutes, a completion time which would be less than the one hour maximum for feasibility and which would permit the test to be administered to most students during a single class period of at least 45 minutes.

Written Test Rankings

The rank of a subject on the WT was to be based solely on the number of correct responses, and only those subjects (Group A) who participated in the IT were ranked. Since two written tests, the WT and the WT2, were administered, rankings were determined for each instrument and are presented in Table 6.4.

As can be seen in Table 6.4, the rankings developed from the WT and the WT2 are similar. The rankings agree perfectly on subjects 8 (rank 6.5), 16 (rank 24), and 31 (rank 18.5), and agree closely on subjects 2 (WT rank 11 to WT2 rank 10.5), 10 (WT rank 4 to WT2 rank 2.5), and 27 (WT rank 1.5 to WT2 rank 2.5). Subject 19 was tied with subject 27 (WT rank 1.5) on the WT and ranked 6.5 on the WT2. The largest discrepancy in the rankings was for subject 20 as he had a WT rank of 24 and a WT2 rank of 9.

Since the two written tests were formed from the same item pool they should have been equivalent. However, the mean for Group A was slightly higher on the WT2 and the WT2 ranking of some subjects varied (from their WT ranking). The gamma statistic of Goodman and Kruskal was computed to check the degree of association and was found to be .55. This value indicated that given two subjects with untied ranks on the written tests, the probability that the ordering of their ranks is the same exceeds the probability that their ranks will have a different

Table 6.4

RANKINGS OF GROUP A BASED ON THE RESULTS OF THE WT AND THE WT2

Subject Number*	WT Number Correct	WT Rank**	WT2 Number Correct	WT2 Rank**
1	8	14	12	6.5
2	9	11	10	10.5
3	7	16	9	13
4	9	11	12	6.5
5	9	11	7	18.5
6	7	16	4	27
7	5	21	4	27
8	12	6.5	12	6.5
9	6	18.5	5	24
10	13	4	14	2.5
11	5	21	3	29
12	3	27	6	21.5
13	3	27	2	30.5
14	3	27	7	18.5
15	11	8	13	4
16	4	24	5	24
17	13	4	7	18.5
18	9	11	8	15.5
19	14	1.5	12	6.5
20	4	24	11	9
21	9	11	9	13
22	7	16	9	13
23	2	30	5	24
24	2	30	2	30.5
25	12	6.5	16	1
26	13	4	10	10.5
27	14	1.5	14	2.5
28	4	24	6	21.5
29	5	21	4	27
30	2	30	8	15.5
31	6	18.5	7	18.5

* The subject number represents the order of his/her appearance in the interviews. Subjects 1-16 were video taped and subjects 17-31 were audio taped.

** In case of ties on number correct, the ranks were averaged.

ordering. Despite the high ranking agreement, the investigator decided to compare both written test rankings to the IT ranking to see which test produced a stronger relationship.

The Interview Test (IT)

Group A, the students designated as at least average achievers in mathematics, also participated in an interview test (IT) where the thinking aloud procedure was followed. Their mathematical problem solving protocols were coded, scored, and ranked. The relevant data resulting from these procedures is reported in this section.

The Thinking Aloud Procedure

The first question posed in Chapter IV concerned the effectiveness of the thinking aloud procedure and related coding scheme for capturing and classifying the mathematical problem solving behaviors of seventh graders. Data and observations resulting from the interviews and coding were to provide empirical evidence for making judgments.

During the interviews, the investigator observed four indicators which could determine the effectiveness of the thinking aloud procedure. The signs included subjects' remarks concerning their ability to think aloud, periods of silence, the use of retrospection, and subject nervousness. Table 6.5 summarizes the occurrences of these indicators separately for the video taped and the audio taped interviews.

As seen in Table 6.5, two subjects from each taping made a direct comment about their ability to think aloud. For example, subject number five worked calmly but quietly, and after reading the fifth problem explained to the observer, "I'm gonna (sic) figure this out in my mind and tell you when I'm done--or else I can't get it". Audio taped subject 20 commented that she had to spend half of her time "concentrating

Table 6.5

INDICATORS OF THINKING ALOUD DIFFICULTIES

	During Video Taping	During Audio Taping
Number of Subjects Who Made Comments on Their Thinking Aloud Ability	2	2
Number of Subjects Who Explained by Retrospection	5	4
Number of Silent Pauses Which Occurred:		
of 30-60 seconds	20	25
over 60 seconds	19	21
Number of Subjects Who Were Judged to be Nervous	7	6

on thinking out loud". However, three subjects who made comments were rated (rating explained below) "Very Good" at thinking aloud and only one of these three indicated any nervousness.

Retrospection was indicated by the number of subjects who offered explanations after they had achieved an answer. Five video taped subjects used retrospection in a total of ten instances with one subject resorting to retrospection on all five problems which she solved. Four audio taped subjects accounted for eight instances of retrospection.

Silent pauses were periods of time during which subjects produced no codable behavior while they were attempting to solve the problems. Pauses less than 30 seconds were often used by subjects for assimilating information, organizing ideas, or silent recapitulation and were not considered indicators of thinking aloud difficulty. However, pauses longer than 30 seconds usually occurred in the protocols of subjects who generally had difficulties expressing their thoughts aloud. Since the number and duration of silent pauses seemed to be a strong indicator of the ease at which subjects could think aloud, all pauses over 30 seconds were recorded and dichotomized; pauses less than one minute and those lasting longer than one minute. As indicated in Table 6.5, the silent pauses occurred more frequently during audio taping. Twenty-five short and 21 long pauses were noted as compared to the 20 short and 19 long pauses which occurred during video taping. Six subjects made no pauses over 30 seconds and 13 used only one or two pauses of either length. At the other extreme were subjects 8 and 19, twin brothers who had much difficulty thinking aloud. Subject 8 paused eight separate times for a total of 570 seconds and his brother lapsed into silence 13 separate times for a total of 1,020 seconds. One of subject 13's silent intervals continued 270 seconds during which the observer used prodding questions four times without provoking a response which could be coded.

The third category in Table 6.5 was a result of the subjects' unspoken reactions to participating in the interview. Four video taped subjects and three audio taped subjects produced clear indications of nervousness. The most frequent and obvious signs included tapping a pencil, scratching parts of the body, or frequent shifting of body positions. Three other subjects from each taping procedure exhibited less obvious nervous behaviors. A subtle indicator of nervousness was the habit of subjects to read the problems rapidly or carelessly, sometimes slurring or mispronouncing words. Four subjects exhibited a noticeable physical habit and six read rapidly. At the beginning of one interview, a subject orally indicated some nervousness when she expressed concern about her ability to solve the problems, and a week after the interviews, another subject directly stated that she was nervous during the interviews.

The number of silent pauses noted earlier seemed to be a strong indicator of a subject's ability to think aloud. Thus, a categorizing scheme was created in order to rate subjects and judge the effectiveness of the thinking aloud procedure. The categories were "Very Good" (2 or less pauses), "Good" (3 or 4 pauses), "Fair" (5 or 6 pauses), and "Poor" (7 or more pauses). The results of applying the rating scheme is summarized in Table 6.6.

Table 6.6

THINKING ALOUD RATING OF SUBJECTS

Rating		Number of Video Taped Subjects	Number of Audio Taped Subjects
Very Good	(2 or less pauses)	11	8
Good	(3 or 4 pauses)	1	3
Fair	(5 or 6 pauses)	2	3
Poor	(More than 6 pauses)	2	1

As indicated in Table 6.6, the thinking aloud abilities of video taped and of audio taped subjects were comparable. Eleven video taped subjects rated "Very Good" while only eight audio taped subjects achieved that rating, but audio taping had three "Good" verbalizers to only one "Good" for video taping. Each type of taping had four subjects who were rated either "Fair" or "Poor" at thinking aloud. Over both taping procedures, 23 of the 31 subjects were able to think aloud without much silent hesitation and eight subjects had difficulty verbalizing their thoughts consistently.

The data and observations resulting from the problem solving interviews did not produce any clear indications of the effectiveness of the thinking aloud procedure. However, it was obvious that some seventh graders found it very difficult to think aloud, as evidenced by their silent pauses and retrospection. The implications of the subjects' inability to verbalize are discussed in Chapter VII. Data resulting from the application of the coding system to the subjects' interview protocols is introduced next.

The Coding Systems

After all the interviews were conducted, the resulting taped protocols were coded according to the revised coding system found in Appendix G and were scored by Lucas' point system which was described in Chapter IV and is summarized in Appendix F. The solution and coding times data, coder reliability measures, and observations about the coding system are presented in this section.

During the pilot study, the investigator used Lucas' coding system for the protocols and was fortunate enough to receive his assistance as a second coder. Using a direct ratio of the frequency of agreements to the total frequency of agreements and disagreements, an agreement measure was computed for the process-sequence coding (.72), the checklist (.67), and the scoring system on "Approach" (.93), "Plan" (.86), and "Result" (.86). The agreement measure for each area was acceptable and the sources of disagreement on the checklist and process-sequence codings were examined in order to improve the investigator's interpretation and application of Lucas' system.

The modifications of Lucas' system for this study necessitated additional agreement measures and three coders including the investigator (Coder 1) were used to establish them. Coder 2 was Norman Loomer, a mathematics instructor at Ripon College in Ripon, Wisconsin. He was also conducting a study which utilized Lucas' coding and scoring system, thus little additional training and few practice comparisons were necessary for him to apply the investigator's system. In addition, Loomer made coding suggestions and helped in coder agreement decisions. Coder 3 was Ruth Meyer, a mathematics education graduate student at the University of Wisconsin-Madison. She is also an experienced teacher who has taught mathematics at all levels from elementary school through college. After Meyer practiced using Lucas' system, the coded protocols were compared and recoded until close agreement was reached with the investigator.

After the training and practice periods, the investigator randomly selected one video taped protocol and one audio taped protocol from each problem of the IT. These 12 protocols and four randomly selected protocols from Loomer's study formed the sample for establishing coder agreement. No protocols which had been used for practice were included in the sample and the three coders all coded the same 16 protocols.

Since good agreement had been established with Lucas during the pilot study, the new variables and modifications which the investigator introduced were the central concern of the second interjudge agreement measure. However, in order to assist Loomer in establishing an intercoder agreement for his study, a large subset of behavioral variables which represented both Lucas' and this investigator's coding system was selected. The new variables Rr, DX, TR, and TS and key variables S, Mf, Me, Alg, DS, DA, and C represented processes. The variables Rs (restates the problem in his own words), An (reasoning by analogy), Vs (varies the process), and Vm (varies the problem) were omitted because the behaviors appeared infrequently during the tapings. The variable R (reads the problem) was omitted because each subject was directed to read the problem aloud before he began to solve it and any later reading was coded as Rr (rereading). The N (not classifiable) was not considered an important process and was omitted.

Lucas' five outcome variables and his punctuation marks were sufficiently well defined so that not much practice disagreement occurred on these variables. Furthermore, some disagreement on these variables could be tolerated without affecting the evaluation of a subject's achievement. Thus these variables were omitted from the agreement comparisons.

The error variables "se" (structural error in process) and "ee" (executive error in process) were included because a new checklist category had been established for structural error. However, the error correction variables were omitted because high agreement was noticed during practice comparisons.

Eight variables from the checklist were included in the subsystem for determining coder agreement. Variables X_{20} (misinterprets data), X_{21} (misinterprets question), X_{16} (algebraic manipulations), and X_{17} (arithmetic computational error) were included to check the clarity of the new error categories. The variable for using an appropriate representative diagram (X_6) was included because subjects in Loomer's study used drawings frequently. The only other checklist variables of common interest to Loomer and this investigator were the performance measures involving scores: X_{26} (Approach), X_{27} (Plan), and X_{28} (Result). A fourth scoring measure, X_{29} (Total), was dependent upon the others and thus not included. The remaining checklist variables were omitted from the study because they did not depend heavily upon individual judgment (i.e., rereads entire problem) or they appeared too infrequently (i.e., recalls related formula) to get a meaningful and reliable agreement measure.

After the 16 protocols were coded, comparisons were made between two coders at a time. The frequencies of agreement, of disagreement, and of positive observations were recorded. A positive observation was an instance in which either coder alone or both coders simultaneously identified the occurrence of the behavior. The frequency of agreement included the number of protocols in which both coders agreed that the behavior did not occur. After the three frequencies were obtained, agreement measures were computed.

Two agreement measures were computed for each variable. A direct ratio of the frequency of agreements to the sum of the frequencies of agreements and disagreements produced a simple agreement measure based only on positive observations. However, coders can disagree consistently (Coder A regularly codes the behavior at least as many times as Coder B does) or inconsistently (Coder A codes the behavior more frequently than Coder B does for some subjects, but Coder B codes the behavior more often for other subjects) and the type of disagreement was important, especially for Loomer's study of heuristic training effects. Thus, indices of reliability which included coder biases were also computed. Kruskal's gamma statistic (cf. Hays, 1963, p. 655) is reported for the dichotomous variables M_f, X_6, and X_{26}, and a product-moment correlation coefficient is reported for the remainder of the variables. Appendix J contains the frequencies and agreement measures for each pair of coders and Table 6.7 presents the averages computed from the three pairings.

According to the agreement ratios in Table 6.7, Me (model by equation or relation) produced the lowest value of .61 and the remainder of the variables were agreed upon by the coders at least 70 percent of the time. Since Me was not a new or important variable, the value of .61 was accepted. Furthermore, the reliability index indicated that the disagreements on Me formed a highly consistent pattern and that coder bias was not a critical factor.

For the three variables, S (separates and summarizes the data), DX (deduction through exploratory work), se (structural error), and X_{20} (se in data) which had low reliability indices of .48, .56, .58,

Table 6.7

AGREEMENT MEASURE AVERAGES OVER CODERS 1, 2, AND 3

Variable	Index of Reli-ability	Agree-ments	Dis-agree-ments	Positive Observa-tions	Agree-ment Ratio
Rr (Rereading)	.91	23.3	7.0	25.0	.77
S (Sep. Data)	.48	13.7	3.0	5.3	.82
DS (Deduction/Syn.)	.94	26.3	8.0	28.0	.77
DX (Deduction?Exp.)	.56	14.7	1.7	4.3	.90
DA (Deduction/Anal.)	.85	19.0	7.7	15.7	.71
TS (Syst. Trials)	.92	16.7	1.3	6.0	.93
TR (Rand. Trials)	.73	14.7	0.7	3.3	.95
Me (Model)	.88	33.7	21.7	47.7	.61
ee (Exec. Error)	.96	18.0	5.0	16.3	.78
se (Struc. Error)	.58	13.0	3.7	7.3	.78
M_f (Diagram) d	1.00	16.7	0.0	3.3	1.00
Alg (Algorithm)	.88	44.0	18.0	60.0	.71
C (Check)	.81	16.3	5.7	12.0	.74
X_6 (Rep.Dia./Yes) d	1.00	16.0	0.0	5.0	1.00
X_{16} (Albeg./ee	.83	18.7	2.3	8.0	.84
X_{17} (Arith./ee)	.82	16.7	0.3	4.3	.98
X_{20} (Data/se)	.34	14.0	2.0	3.7	.88
X_{21} (Question/se)	.79	15.3	0.7	1.7	.96
X_{26} (Plan Score) d	.96	14.0	2.0	16.0	.88
X_{27} (App. Score)	.67	11.3	4.7	16.0	.71
X_{28} (Res. Score)	.92	13.0	3.0	16.0	.81

d = Dichotomous variable

and .34 respectively, high agreements were computed. This inconsistency was a result of the distribution of the disagreements, the small number of positive observations, and the high ratio of the number of disagreements to the number of positive observations. For example, X_{20} (misinterprets data) was found only four time in comparing the coding of Coders 1 and 2. The low reliability index resulted because Coder 1 identified the behavior when Coder 2 did not note it in one instance and, in two instances, Coder 1 failed to identify the behavior when Coder 2 had noted it. A high agreement ration (.84) resulted because the coders agreed once when the behavior did occur and they agreed that the behavior was not present in 12 observations, this producing a ratio of 13 agreements to 16 (13 agreements and 3 disagreements) positive observations.

Since their agreement ratios were uniformly high, the low indices of reliability for S, DX, se, and X_{20} were considered spurious. However, each variable was examined further to check its effect in this study. The S variable was not influential in determining a subject's ranking.

For DX, Coders 1 and 2 had an agreement ratio of .88 and a reliability of .68. Since these two coders were the implementers of the coding scheme, the consistency was judged adequate. The disagreements on DX were negated when coders used DS accompanied by se to indicate that the subject was combining the data indiscriminately or that the subject misinterpreted the question. Both codings resulted in a lower score for the subject's attack (Plan) or for his understanding (Approach) of the problem, and a subject who exhibited the errant behaviors usually attained a poor score for his answer (Result). The acceptable agreement ratios and reliability indices for X_{26}, X_{27}, and X_{28} supported judgement that the disagreements on DX did not seriously affect the subject's rankings.

Structural errors (se) were an important factor in applying Lucas' scoring system and a reliability index of .58 appeared low. An inspection of the sources of disagreements discounted possible IT ranking inconsistencies. The investigator, Coder 1, share reliabilities of .71 and .66 with Coders 2 and 3 respectively. These values indicated that the structural errors were applied with acceptable consistency by the investigator. Inconsistency arose when coders used DS accompanied by se instead of DX. Other disagreements occurred when a coder classified an error as ee instead of se. Uncorrected errors of either type or poorly planned process irregardless of the label also resulted in a lower subject score and ranking. Thus, the inconsistencies of se labeling did not adversely affect the scoring and ranking system.

The variable X_{20} was dependent upon the identification of se, thus its effect upon the scoring and ranking system was also discounted. The type of disagreements which accounted for the low reliability index of se were chiefly responsible for the low index of X_{20}.

After agreement ratios and reliability measures were computed, examined, and evaluated, the coded protocols and scores were used to search for ranking schemes. The IT ranking procedures are described next.

The IT Ranking Schemes

The second major question posed in Chapter IV was, "Is it possible to assess, separate, and rank seventh graders according to their coded mathematical problem solving protocols?". Lucas' scoring system was used for assessing problem solving achievement and determining rankings.

After the application of Lucas' scoring system, four measures were available for each problem: Approach (0 or 1), Plan (0, 1, or 2), Result (0, 1, or 2), and Problem Total (0-5). (Appendix K) The first ranking scheme (Ranking A) was developed by summing problem totals for each subject across the six problems and assigning the rank of 1 to the highest sum. Tied ranks were averaged. The sums represented the combined evaluation of a student's understanding of the problem, the quality of his plans, and the accuracy of his results. The totals and ranks for A are presented in Table 6.8.

According to ranking A, subject 15 had the highest total (24 points) and was ranked first, while subjects 24 and 29 scored no points and shared the average of ranks 30 and 31. Other ties occurred at scores of 18, 10, 9, 8, 5, 4, and 3 points. Five subjects were tied at 9 to share rank 14 (average of 12-16) and five other students were tied at 8 to share rank 19 (average of 17-21). Except for three subjects tied at 18 points, the remaining ties occurred in pairs.

Table 6.8

INTERVIEW TEST SCORES AND RANKINGS A, B, AND C

Subject	Approach Sub-Total A_i	Plan Sub-Total P_i	Result Sub-Total R_i	Total Interview Test Score	Ranking A	Ranking B	Ranking C
1	5	5	4	14	8	6	9
2	2	3	4	9	14*	19.5*	19.5*
3	2	3	3	8	19*	21	21
4	5	7	6	18	5*	4.5*	4.5*
5	3	4	1	8	19*	15	12
6	1	1	1	3	28.5*	29	29
7	2	2	1	5	24.5*	24.5*	24.5*
8	2	3	4	9	14*	19.5*	19.5*
9	6	7	6	19	3	2	3
10	2	2	3	7	22	22	22
11	4	3	1	8	19*	11	16
12	5	3	1	9	14*	7	14
13	2	2	2	6	23	23	23
14	2	1	1	4	26.5*	26	27
15	6	10	8	24	1	1	1
16	1	2	1	4	26.5	28	26
17	3	4	3	10	10.5*	13.5*	10.5*
18	2	2	1	5	24.5	24.5*	24.5*
19	4	7	7	18	5*	8	6
20	3	3	2	8	19*	17	18

Table 6.8
(cont'd)

Subject	Approach Sub-Total A_i	Plan Sub-Total P_i	Result Sub-Total R_i	Total Interview Test Score	Ranking A	Ranking B	Ranking C
21	3	6	4	13	9	12	8
22	3	4	3	10	10.5*	13.5*	10.5*
23	3	3	3	9	14*	16	17
24	0	0	0	0	30.5*	30.5*	30.5*
25	4	6	7	17	7	9	7
26	5	8	7	20	2	3	2
27	5	7	6	18	5*	4.5*	4.5*
28	2	1	0	3	28.5*	27	28
29	0	0	0	0	30.5*	30.5*	30.5*
30	2	4	2	8	19*	18	13
31	4	3	2	9	14*	10	15

* Ties occurred

Note: Subtotals were a subject's partial scores summed across the six interview problems.

The large number of ties in Ranking A did not separate subjects well and was likely to produce a low association with written test ranks. Thus, two additional schemes (Rankings B and C) which better differentiated between subjects were developed. Seeing that subjects with tied scores earned their points in different phases of the problem solving process, the investigator attempted to categorize subjects by their subtotals for Approach (A), Plan (P), and Result (R): A_i was equal to the sum of the Approach scores for subject i across the six problems; P_i was equal to the sum of the Plan scores; and R_i which was equal to the sum of the Result scores. Thus, subject j who achieved scores of (1, 1, 0), (1, 2, 2), (0, 0, 0), (1, 2, 1), (1, 1, 1) and (1, 1, 2) for his Approach, Plan, and Results respectively, attained subscores of $A_j = 5$, $P_j = 7$, and $R_j = 6$.

Ranking B was based on A_i, P_i, and R_i, but gave priority to subjects who demonstrated an understanding of the most problems. By this system, the highest A_i score was ranked first. In case of ties, the subject with the highest P_i scores received the next rank. If subjects were tied after comparing the A_i's and P_i's, then the R_i's were compared with the higher value receiving the next rank. If ties existed for all three scores, the ranks were averaged.

Ranking C was similar to Ranking B, but it emphasized the subject's plans and processes. The P_i scores of subjects were the first determiner of ranks and the A_i and R_i scores were compared in that order if ties occurred. Table 6.8 presents the A_i, P_i, and R_i scores with the total scores, and Rankings A, B, and C.

As can be seen in Table 6.8, Rankings A, B, and C agree on the ranks assigned to subjects 7, 10, 13, 15, 18, 24 and 29 and are similar in the other ranks. Since four pairs of subjects had identical subscores, Rankings B and C each produced four pairs of ties and any other ranking system based on ordering A_i, P_i, and R_i would have had similar results. The rank of subject 11 varied the most as it was 19 on Ranking A and 11 on Ranking B.

Lucas' scoring system made it possible to develop three rankings of the subjects and his measures were also used in the exploratory ranking procedures of Part III. The association of Rankings A, B, and C to the written test rankings is reported after other data resulting from the interview and coding procedures is presented.

Audio Versus Video Taping

The incorporation of video taping into the study prompted questions about tape type differences in recorded information, in subjects' performances, and in coding time. Data and observations are presented to identify the differences between audio and video taping.

The physical differences in audio and video taping are immediately apparent. Instead of a single tape recorder which the observer can operate alone, video taping requires at least one camera, special lighting, and a technical assistant. More than one pre-focused camera or a single camera which can be regularly refocused is necessary to effectively capture a subject's actions and writing. Compared to audio taping, the array of equipment and technical assistance necessary for video taping is more costly to the investigator and perhaps more distracting to the subject.

The disadvantages of video taping were offset by the information which would not have been captured on an audio tape. Interesting physical actions such as a subject's smile, frown, or grimace, and his nervous habits of scratching parts of his body or shifting his postition were recorded. Unspoken problem solving procedures were the most important observations noted on video tape. For example, subjects reread the problem or parts of it silently, but clearly indicated their behavior by following the sentences with their eyes or pencil, by moving their lips, or by asking a question immediately after staring at the problem. Ninety-five occurrences of these re-reading behaviors which would not have been recorded on audio tape were noted for the 16 video tape subjects. Furthermore, a comparison of the observer's notes to the coded protocols revealed that 49 silent rereadings were not recorded by the audio tape.

Another problem solving strategy which was not readily discernable on audio tape occurred whenever subjects drew or modified a diagram without orally indicating their exact actions. Problem 4 on the IT was solved by five subjects through the sketch of a ladder, but the coder used the completed diagrams and the subjects' verbalizations to speculate on the sequence of modifications during all five protocols. Routine computations were also subject to coder guessing if the student did not adequately verbalize his actions. For example, one subject performed seven written multiplications silently as she attempted to divide 100 by 8.

The advantages of video tape for recording subject behaviors in interview situations were clear without any need for statistical comparisons. However, the questions about possible performance differences due to video taping were answered by significance tests. The total process sequence scores and the total solution times of subjects were used as measures of performance differences.

From the pilot study results, the investigator suspected that the presence of novel and distracting video taping equipment caused the subjects to behave differently than if they were audio taped. It was felt that video taped subjects spent less time solving the interview test problems and that the haste of the video taped subjects would result in lower scores. These suspicions were checked statistically when two hypotheses were tested:

H1: The mean of video taped subjects' total interview test scores equals the mean of audio taped subjects' total interview test scores.

H2: The mean of video taped subjects' total solution times on the interview test equals the mean of audio taped subjects' total solution times on the interview test.

The individual total scores are presented in Table 6.8 and the total solution times are presented in Appendix I. The analysis of variance statistics for hypotheses H1 and H2 are reported in Tables 6.9 and 6.10 respectively.

Table 6.9

ANALYSIS OF VARIANCE FOR TOTAL INTERVIEW TEST SCORES

Source	df	MS	F	p<
Treatments	1	.24	.006	1.00
Error	29	38.31		

As Table 6.9 indicates, the null hypothesis H1 cannot be rejected. The very low F ratio of .006 was an indirect result of the close similarity of the video and audio taped subjects' scores. The video taped subjects averaged 9.7 points with a standard deviation of 5.8 while audio taped subjects achieved a mean of 9.9 with a standard deviation of 6.2.

Table 6.10

ANALYSIS OF VARIANCE FOR SUBJECTS' TOTAL SOLUTION TIMES ON THE INTERVIEW TEST

Source	df*	MS	F	p<
Treatments	1	101.00	3.97	.10
Error	27	25.44		

*Due to erasure of tape, two subjects' protocols could not be timed.

As seen in Table 6.10, the significance level of .05 was not reached and the null hypothesis H2 is not rejected. However, the F ratio of 3.97 was significant below the .10 level and the analysis suggested that there were some treatment differences. The video taped subjects' solution time mean of 16.7 minutes compared to the audio taped subjects' mean of 13.0 minutes made it apparent that video taped subjects took about the same amount of solution time as did the audio taped subjects.

Lucas suggested that coding video taped protocols took less time than coding audio taped protocols. His observation was tested with hypothesis H3:

H3: The mean of the coding times for video taped subjects' protocols equals the mean of the coding times for audio taped subjects' protocols.

The coding time for each subjects' protocol is presented in Appendix I and the analysis of variance statistics is reported in Table 6.11.

Table 6.11

ANALYSIS OF VARIANCE FOR CODING TIMES

Source	df*	MS	F	p<
Treatments	1	.68	.002	1.00
Error	27	292.09		

*Due to erasure of tape, two coding times could not be measured.

As reported in Table 6.11, the extremely low F ration of .002 did not reach the .10 significance level. Thus, the null hypothesis H3 is not rejected and it appears that audio tapes and video tapes require similar coding times. The sample means of 42.3 (VT) and 42.6 (AT) and sample variances of 17.3 (VT) and 15.8 (AT) indicate that the coding time distributions were nearly identical.

The difference in the means of audio taped and of video taped subjects' solution times prompted a further analysis of coding times. Direct observation of the data suggested that solution times were not commensurate with coding times. Thus, solution time totals and coding time totals across subjects were found for audio taping and for video taping. The ratios of coding time to solution time were computed for each tape type and the difference between the ratios was found. The results are presented in Table 6.12.

Table 6.12

COMPARISON OF CODING TIME RATIOS

	Total Solution Time	Total Coding Time	Coding Time / Solution Time
Video Tape (15 subjects)*	251 minutes	635 minutes	2.53
Audio Tape (14 subjects)*	182 minutes	597 minutes	3.28

Savings: 3.28 - 2.53 = .75 minutes per one minute of tape

* Due to erasure of tape, one coding time for each tape type could not be measured.

As indicated in Table 6.12, the video taped protocols lasted 251 minutes and took 635 minutes to code while 182 minutes of audio taped protocols took 597 minutes to code. Thus, one minute of audio tape took 3.28 minutes to code and one minute of video tape took only 2.53 minutes to code. The .75 minutes difference represents a savings of approximately 22 percent of the audio coding time on a minute of tape.

The data and observations resulting from the interviews and coding procedures were used to seek answers to principal and secondary questions of the study. However, the central concern of the study depended upon the correlation of the rankings identified earlier in this chapter.

The correlations and exploratory statistics are reported next.

Statistical Analyses of Rankings

The feasibility of using a written instrument as a substitute for the complex interview and coding procedure depended upon the relationships resulting from the written test and the interview tests. Two written tests, the WT and the WT2, were administered and three rankings, A, B, and C, were developed from the IT. The exploratory procedures which were used to seek additional rankings are explained after the initial statistics are reported.

Relationships of the Written and Interview Tests

Two comparisons were possible after the written and interview tests were scored and their rankings were developed. A product-moment correlation coefficient r_{XY} was computed between the raw scores (number correct) on the written tests and the interview test total and subtotal scores used for developing each ranking. Thus, the correlations involving Ranking A were based on the total IT scores while correlations involving Ranking B used the IT subtotals for Approach and correlations involving Ranking C used the subtotals for Plan. For each correlation coefficient, a hypothesis that the population statistic ρ_{xy} equals zero was tested by a t test with N-2 degrees of freedom.

In addition to the correlation between scores, the relationship between the rankings developed from the tests was also measured. Kendalls tau (Hays, 1963) with ties was computed for the association between the rankings and the significance level of tau was found by computing z values. Because of ties within rankings, Kruskal's gamma statistic was computed to provide a simpler interpretation of Kendall's tau. The correlations and rankings statistics for the pairs WT and Ranking A, WT and Ranking B, WT and Ranking C, WT2 and Ranking A, WT2 and Ranking B, WT2 and Ranking C, and (WT + WT2) and Ranking A are presented in Table 6.13.

Table 6.13

CORRELATION AND RANKING STATISTICS FOR THE INTERVIEW TEST AND THE WRITTEN TESTS

	r_{xy}	tau	p(tau)	gamma
WT & Ranking A	.61*	.44	.001	.48
WT & Ranking B	.40**	.33	.007	.34
WT & Ranking C	.59*	.39	.002	.41
WT2 & Ranking A	.64*	.49	.001	.52
WT2 & Ranking B	.48**	.38	.002	.40
WT2 & Ranking C	.61*	.45	.001	.46
(WT + WT2 & Ranking A)	.68*	.50	.001	.52

* Significant at the .001 level in two tailed t test of $H_o: \rho_{xy} = 0$

** Significant at the .05 level in two tailed t test of $H_o: \rho_{xy} = 0$.

As reported in Table 6.13, none of the correlation coefficients between the seven pairs of written and interview test scores attained the desired minimum of .71 although the combined scores of the WT and the WT2 produced an encouraging correlation coefficient of .68 with the total IT score. The Plan subscore used for Ranking B produced the lowest correlations: the correlations the WT score and the WT2 score were .40 and .48 respectively. Two pairs of scores, WT & Ranking A and WT2 & Ranking C, each resulted in a correlation of .61. Statistically, all seven correlation coefficients resulted in t test values which were significant at the .05 level. Thus, the hypothesis that no correlation exists between written and interview test scores was rejected.

The associations between the rankings reported in Table 6.13 resulted in values which appeared to be low but which were statistically significant. Kendall's tau values ranged from a low of .33 for WT & Ranking B to a high of .50 for (WT + WT2) & Ranking A. However, the probabilities for all seven tau values were below .01 and four probabilities fell below the .001 chance level. Kruskal's gamma statistic ranged from .34 for WT & Ranking B to .52 for two pairs of rankings, WT & Ranking A and (WT + WT2) & Ranking A. The gamma values indicated that if two subjects had untied rankings, the probability was favorable that their ranks would have the same ordering on the written and on the interview tests.

Exploratory Procedures

As indicated in Chapter IV, exploratory statistical analyses, namely latent partitioning and clustering, were to be used to search for underlying patterns among subjects and to possibly produce other ranking schemes. Because the computer program for latent partitioning was not available, another pattern seeking program called multidimensional scaling was substituted. A similarity measure D (Figure 6.2) based on subscores for Approach, Plan, and Result was computed for each pair of subjects and was used in both analyses. The matrix

D = Distance Measure $\qquad Z_{Aj}$ = Aj Normalized
Aj = Total Approach Score of Subject J $\qquad Z_{Pj}$ = Pj Normalized
Pj = Total Plan Score of Subject j $\qquad Z_{Rj}$ = Rj Normalized
Rj = Total Result score of Subject j

$$D(Si, Sj) = (Z_{Aj} - Z_{Aj})^2 + (Z_{Pi} - Z_{Pj})^2 + (Z_{Ri} - Z_{Rj})^2$$

Notes: 1. D (Si, SJ) = 0
2. D (Si, Si) = 0
3. D (Si, Sj) = D (Sj, Si)

Figure 6.2. Similarity Measure Formula

of resulting values was organized by incorporating the multidimensional scaling data and is presented in Appendix L.

Guttman-Lingoes multidimensional scaling program (Lingoes, 1973) searches for underlying patterns or structures among the similarity measures. The program then represents the structure in a spatial model by assigning coordinates to the objects (subjects) and computes stress values to measure the agreement between the order of the spatial distances and the order of the similarity measures. Higher agreement is indicated by low stress values. A second measure, the coefficient of alientation, deals with the type of monotonicity criterion for the relationship between distance and similarity measures. The coordinates, stress values, and coefficients of alientation for one, two, three, and four dimensions were produced by the Guttman-Lingoes program.

Coordinates and accompanying values for two through four dimensions are listed in Appendix M. The one dimension results closely paralleled earlier rankings and are discussed here. Table 6.14 presents the one dimension scaling coordinates in an order which permitted a ranking to be imposed.

As can be seen in Table 6.14, the multidimensional scaling program assigned subject 15 one extreme coordinate of - 100.000 and assigned subject 29 a coordinate of 100.000. The parallel to Ranking A was immediately obvious and by assigning Rank 1 to subject 15, Rank 2 to subject 26, and continuing until rank 31 was assigned to subject 29, a ranking very similar to Ranking A was obtained. Kendall's tau of .96 and Kruskal's gamma statistics of .99 verified that the agreement between the two rankings was almost perfect and that little information was lost by basing Ranking A on total scores. Conversely, not much information was gained by using the subscores. Kruskal's stress measure of .11557 indicated that there was fairly strong agreement between the rank orders of the spatial distances and of the similarity measures. A perfect coefficient of alientation (.00000) resulted from weak monotonicity (distance from coordinate i to coordinate j = distance from coordinate k to coordinate ℓ whenever the similarity of subjects i and $j \leq$ the similarity of subjects k and ℓ) requirements.

Johnson's (1967) max clustering algorithm was the second exploratory procedure used to group subjects according to some structure underlying the similarity measures. The program defines a sequence of partitions of a set of objects and uses the similarity values to

Table 6.14

ONE DIMENSIONAL SCALING COORDINATES
AND A RESULTING RANKING

Kruskal-Guttman-Lingoes-Roskam Smallest Space Coordinates for M=1 (Weak Monotonicity)

Variable (Subject)	Coordinate	Rank	Variable (Subject)	Coordinate	Rank
15	-100.000	1	11	28.709	17
26	-66.349	2	20	81.710	18
9	-62.808	3	5	32.155	19
27	-51.554	4	3	36.062	20
19	-51.124	5	30	37.456	21
4	-50.628	6	10	42.423	22
25	-43.609	7	13	48.544	23
1	-19.661	8	7,18	55.265	24.5
21	- 8.553	9	14	61.291	26
12	5.851	10	16	65.620	27
17,22	17.527	11.5	28	68.045	28
31	19.234	13	6	71.542	29
23	23.108	14	24	98.692*	30
2,8	25.569	15.5	29	100.000	31

*Error: Subjects 24 and 29 had identical subscores. Therefore, they should both have coordinates of 100.000 and ranks of 30.5.

Kruskal's stress = .11557 in 6 iterations

Guttman-Lingoes' coefficient of alienation = .00000 determine "diameters" of the subset. The max procedure attempts to construct hierarchical partitions which contain subsets of minimum diameter and assigns a partition rank to each pair of objects. Goodman and Kruskal's (1954) gamma is computed to measure the agreement between the rank order of object pairs obtained from the partition hierarchy and the rank order of the pair's similarity value. Figure 6.3 presents the iterative steps of the clustering algorithm and illustrates the partitions of subjects who were homogeneous in some way. Appendix N contains the gamma values which correspond to each iteration.

As seen in Figure 6.3, the clustering algorithm started with each subject as a distinct group and at each iterative step, joined two groups which were most similar. Thus, iteration 1 joined subjects 24 and 29, iteration 2 joined subjects 4 and 27, and iteration 3 joined subjects 17 and 22. The iterations continued through iteration 30 which produced one group composed of all 31 individuals. Of particular interest is the partition formed by iterations 28 and 29. At this level, the entire group of subjects is divided into two disjoint subsets: The subset under iteration 28 contains subjects 1, 21, 4, 27, 26, 9, 19, 25 and 15 while the subset under iteration 29 contains the remaining subjects. Further observation of Figure 6.3 indicates that iteration 29 is partitioned into the disjoint subsets of iterations 27 and 25. The subset of iteration 27 contains subjects 2, 8, 3, 10, 13, 5, 30, 17, 22, 23, 20, 11, 31, and 12 while the subset of iteration 25 has subjects 6, 16, 7, 18, 14, 28, 24, and 29 as its members.

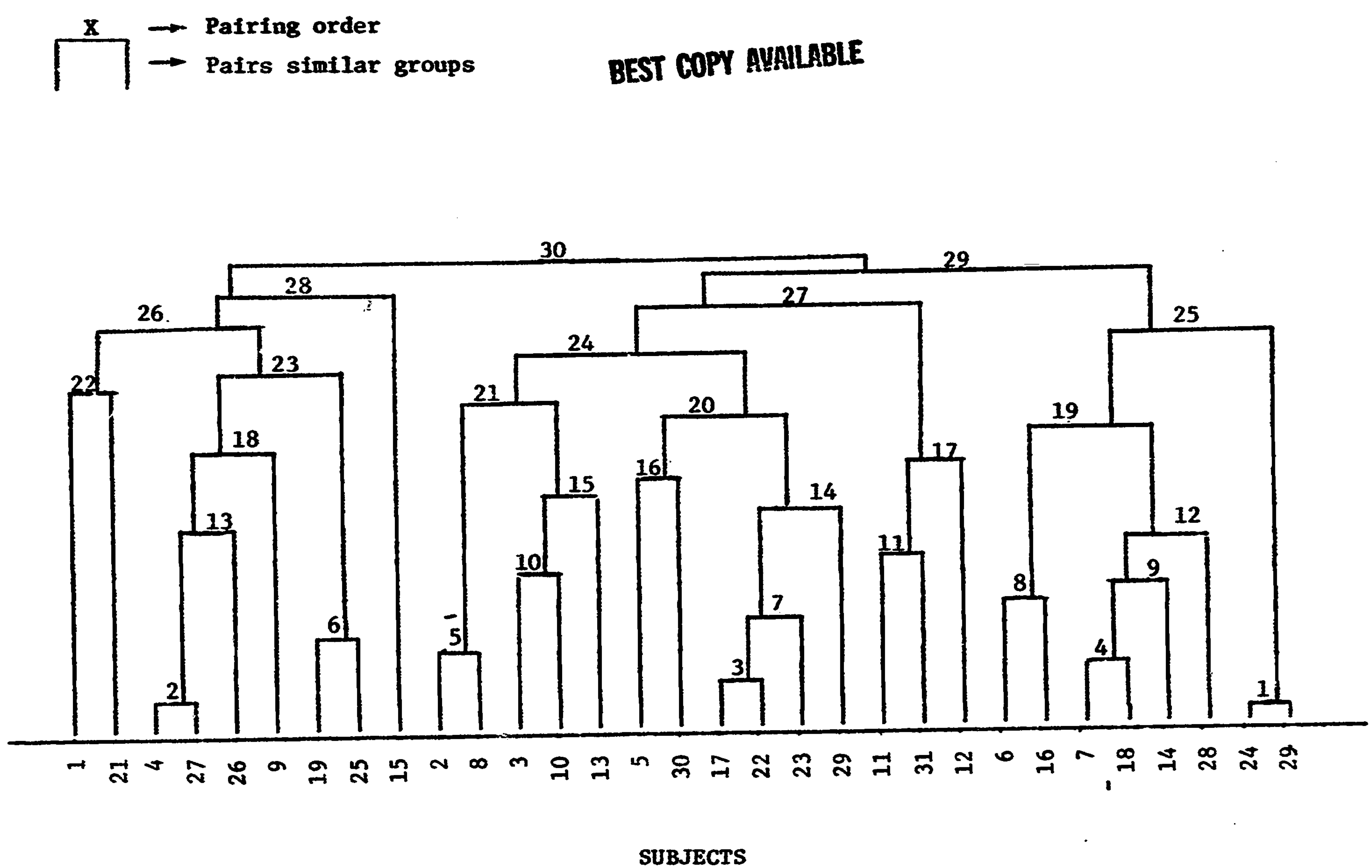


Figure 6.3. Clustering Algorithm Graph of the Sequence of Pairing Similar Groups.

Inspection of the three subsets of iteration 29, 28, 27, and 25 revealed an identifiable pattern which was strongly related to the ranking scheme developed from one dimensional scaling. The subsets of iterations 28 and 29 corresponded to the first nine subjects (15 through 21) and the last twenty-two subjects (17 through 29) as ranked in Table 6.14. Further observation of the table indicated that partitions 28, 27, and 25 divided the subjects into three disjoint groups which corresponded to the first nine (15-21), the next fourteen (12-13), and the final eight (7-29) respectively ranked subjects.

Iteration 28 can be traced backward through the sequential separations of subject 15 (rank 1) and subjects 1 and 21 (ranks 8 and 9) before the clustering loses consistency with the scaling seriation. When subjects 19 (rank 5) and 25 (rank 7) are separated from the remaining six subjects (26, 9, 27, 19, 4, and 25 respectively), the clustering configuration skips subject 4 which has rank 6.

Dimensions two, three, and four of the scaling procedure were difficult to interpret and were inconsistent with the clustering results. For example, in two dimensions, the exploratory procedures displayed agreement on the horizontal axis (vector 1) as the scaling resembled the seriation of one dimension. However, the vertical dimension (vector 2) produced a wide separation between subjects 31 and 11, the students who were paired at iteration 11 in the clustering algorithm. Since subject 31 had subscores (4, 3, 2) and subject 11 had similar subscores (4, 3, 1), and no other evidence could account for the discrepancy, no further relationships or interpretations were sought beyond one dimension.

The results of the exploratory analyses were considered encouraging for future problem solving research. The similarity measure D was different from the measure used to produce Ranking A; however, the underlying structure found by multidimensional scaling was similar to the ranking structure imposed by total scores. Furthermore, the clustering procedure reaffirmed the results of the scaling procedure by producing partitions which were highly consistent with the one dimensional ranking scheme.

Summary of Chapter VI

The written tests were completed without time being a factor and the students did not have difficulties following the test format. However, the reliability measures of the written tests were not sufficiently high for a correlation of .71 between tests to be obtained. Though the written and interview tests failed to attain the minimum correlation coefficient established as a feasibility criterion, the .68 correlation of the WT-WT2 combined score with the IT total score and the high agreement between written and interview test ranks were encouraging.

The revised coding scheme and Lucas' scoring system were applied to the protocols with good intercoder agreement and three logical ranking schemes were developed from the results. The IT scores did not produce the desired correlation coefficient with the written test scores.

Complications arose during the interviews. Nervousness which could be attributed to the experimental setting was not unexpected, but the inability of subjects to think aloud raised questions about the validity and reliability of the thinking aloud procedure. Video taped protocols held two advantages over audio taped records: they recorded important silent problem solving behaviors and they took about 22 percent less time to code.

The conclusions and implications which were made from the data are discussed in Chapter VII.

Chapter VII

CONCLUSION

Introduction

After giving a summary of the study, this chapter presents a discussion of the limitations and conclusions. The implications for problem solving evaluation and recommendations for future research conclude the chapter.

Summary

The main purpose of this study was to explore the feasibility of using a written test to assess and rank seventh graders mathematical problem solving achievement. The feasibility of the written test was to be judged on its physical dimensions, its statistical characteristics, and its agreement with the results of the complex thinking aloud procedure.

Thirty-one subjects were asked to think aloud during mathematical problem solving interviews which were taped. The recorded protocols were coded and scored to provide a valid assessment of the subjects' achievement. Three rankings were developed from the scores and compared to the ranking determined by the number correct on a 20 item written test. The length, format, and reliability criteria of the written test were met, but the correlation coefficients between the written and interviews test scores did not reach .71. However, one coefficient approached the expected value and the order of the rankings had high statistical agreement.

The effectiveness of the thinking aloud procedure for capturing mathematical problem solving was evaluated and serious doubt was cast on its reliability and validity for use with seventh graders. A revised coding scheme described the problem solving behaviors well and was applied with high intercoder agreement, but the subjects' thinking aloud abilities and reactions suggested that the procedure was not capturing their genuine mathematical problem solving tactics.

Secondary questions about recording and coding procedures arose during a pilot study and were included in this investigation. It was found that video taping was advantageous for recording subjects' unspoken behaviors and that less time was needed to code video tape than to code audio tape.

Multidimensional scaling produced an IT subject ranking which agreed closely with the one developed from total scores. The clustering procedure illustrated the grouping of subjects and reinforced the agreement between the other two rankings.

Limitations

Though care was taken to exercise as much control and to permit as much generalization as possible, each part of this exploratory study contained factors which limited the interpretations. The limitations and possible corrective measures are discussed here.

The meanings of "mathematical problem" and "mathematical problem solving" were similar in definition to Lucas' and were similar in spirit to Kilpatrick's. Yet, the definitions used in this study must be considered unique, thus limiting the generalizability of the results.

The school selected for this study was a parochial school, but the results of the WT2 on a larger population indicated that the subjects were fairly representative in achievement. However, precautions must be taken in generalizing beyond the school's population because the interview and statistical results were derived from a select subset of the school's seventh-graders. A random choice of students and schools in a larger population would have permitted a corresponding increase in generalizability.

The latitude of the interpretation also depended upon the reliability and validity of the instruments and procedures. Though most measures were acceptable, the arbitrary criterion levels and inconsistency of coder agreement measures could make coder reliability suspect. A larger number of coders and observations would establish more stable agreement measures.

The results of the thinking aloud procedure were assumed to be valid representations of a subject's problem solving achievement. However, observations made during the interviews indicated that the subjects had difficulties thinking aloud in addition to the usual reactions to an experimental setting. The combination of these observations raised serious questions about the thinking aloud procedure and only further research can determine the effects of the observed behaviors.

The exploratory clustering and multidimensional scaling procedures were subject to personal interpretations, so the results of the analyses must be treated accordingly. When the procedures and interpretations are defined more clearly, the reliability of the resulting information and conclusions will increase.

Conclusions

This section discusses the conclusions of the study with references to the main and secondary questions which were to be answered. The data and observations presented in Chapter VI were used to make the judgments and decisions discussed below.

The physical and statistical qualities of the written tests, the WT and the WT2, indicated that the instruments were suitable for administering to seventh graders in the classroom. Groups A and B in School 1 averaged less than 27 minutes for completion times on the WT2 and it was assumed that no great deviation would occur with other forms of a written test or with other groups of seventh graders. According to the results on the written tests, the directions were clear and easy to follow although the items were difficult to answer. The students filled in the proper spaces with their answers and did not hesitate to omit items which they did not understand or could not solve. The average reliability of both written tests across all groups was an acceptable .79. The small solution time average indicated that a longer written test could be administered in an hour without making the test a speed test. Assuming progress at the same rate, a 25 item written test should take about 34 minutes to solve and, according to the general Spearman-Brown formula (cf. Ebel, 1972, p. 413), it should have a reliability of .83.

The feasibility of the written test was chiefly determined by its ability to predict seventh-graders' problem solving achievement scores and ranks as measured by the IT. The product-moment correlation coefficient was .61 for the IT and WT scores and .64 for the IT and WT2 scores. Though both values were highly significant ($p<.001$) against $Ho:\rho_{xy} = 0$, neither written instrument attained the minimum correlation of .71 which was necessary to account for at least 50% of the variance between written and interview test scores. The IT subscores produced similar results when correlated with the written tests. Thus, the written test must presently be declared not feasible for the purpose of predicting mathematics achievement as measured by the thinking aloud procedure and coding scheme.

The second main question of the study was, "Is it possible to assess, separate, and rank seventh graders according to their problem solving protocols?" The answer appears to be positive. A variation of Lucas' coding system was applied with a high degree of agreement (.83 across the variables, see Table 6.7) and reliability (.80). The variables S, DX, se, and S_{20} produced low reliability measures, but the disagreements which caused the low values did not seriously affect the IT scores. Rankings A, B, and C were logically derived from the scores awarded by Lucas' point system and provided high rank order agreement measures. The scaling and clustering analyses verified that the order imposed by Ranking A was consistent with the similarities and patterns which were detected among the subjects.

Probably the most important outcome of this study resulted as the answer to the first question was sought. The question was, "How well does the thinking aloud procedure and related coding scheme capture and classify the mathematical problem solving behaviors of seventh graders?" and the answer appears to be "not very well." As indicated in the previous paragraph, the coding scheme was applied with acceptable agreement and resulted in logical ranking schemes; however, the behaviors of the students during the thinking aloud interviews raised critical questions about the reliability and validity of the information recorded in the protocols. The seven subjects (Table 6.6) who displayed obvious nervous habits were not likely to have performed as normally as those who were not nervous. Seven out of 31 is already a high ratio and if half of the subjects who gave subtle nervous indicators were indeed nervous, then almost one-third of the subjects were not performing normally. The eight subjects who were rated either "Fair" or "Poor" at thinking aloud add to the suspicion that the procedure did not adequately capture the problem solving behaviors of some subjects and that it may not be a highly valid or reliable method to use with seventh graders.

The differences in audio and video taping have indicated a distinct advantage for the latter because of its ability to detect silent rereading indicators, diagrams and alterations, and written computations. Future investigators need to decide if the extra information is worth the additional expense of video taping.

Subjects in the video taping situation did not react much differently than students who were audio taped. The occurrences of comments, retrospections, nervous subjects, and fair or poor verbalizers were approximately equal in each procedure. The audio taped subjects produced more silent pauses, but the video taped subjects took significantly more time ($p<.10$) to attempt the IT. The scores of each group were nearly identical and produced no significant difference. It appears that although video taping requires extra equipment which could be distracting, the subjects' behaviors, performance times, and achievement scores were not affected any differently than if the students had been audio taped. However, it must be remembered that both procedures may have altered the subjects' behaviors and performances equally.

Implications for Mathematical Problem Solving Assessment

The main purpose of this study was to explore the feasibility of designing a written test to predict mathematical problem solving achievement of seventh graders as measured by the Interview Test. The exploration raised other questions which were included in the study. Possible answers are presented with the recommendations which resulted from the observations and data.

The chief feasibility criterion for the written test was not met although the correlation coefficients were statistically significant.

Assuming the thinking aloud procedure produces a valid assessment of students problem solving achievement, a higher correlation is necessary before the written test scores can be used as a substitute or a predictor; however, the highly significant correlation coefficients and the extremely low probability of Kendall's tau values occurring by chance indicated that the written tests could be used to make scoring and ranking predictions with some confidence. For example, given that student A ranked above student B on a written test, the chances are about 45% greater that student A ranked above student B on the IT than that student A ranked below student B on the IT.

The sum of the WT and the WT2 scores resulted in a correlation coefficient of .68 with the IT score. Since this value indicates that over 46 percent of the variance can be accounted for by knowing one test score, it appears than an appropriately constructed written test with at least 40 items might produce the .71 minimum correlation coefficient. The lengthened test would likely require more than one hour to complete and would probably need to be given in two parts to avoid student fatique, but it would remain quicker and easier for teachers to administer and score than are the complex thinking aloud and coding procedures.

The critical observations of the thinking aloud procedure are not unique. Kilpatrick (1967) was aware of possible interference or interaction of speech and thinking when he had his eighth grade subjects think aloud, but he did not indicate that any of his subjects had difficulty verbalizing while they worked. Menchinskaya (cf. SMSG, 1969) observed that ninth graders and adults with a secondary school mathematics education were able to think aloud easily and that external speech did not hinder them in solving a problem. However, she found that first, fourth, and fifth graders had difficulty verbalizing as they solved arithmetic problems and they commented on the interference it caused in their thinking. She felt that reasoning processes changed and performance deteriorated when these students were required to think aloud. Pereira (1973) made similar observations after he had 11-12 year old girls verbalize while trying to discover the rules of a mathematical structure. He found that subjects who worked in silence during a physical mathematical learning activity (pressing buttons on a machine) performed better and retained more than subjects who verbalized overtly while learning. The evidence from the above investigations and from this study strongly suggests that the thinking aloud produce may not cause much interference with adults and youths who have attained mental maturity, but that the interference of overt speech with thinking increases as the mental maturity of the subjects decreases.

The exploratory analyses tried in this study have some potential for problem solving research. Clustering and multidimensional scaling produced graphic data which made groupings visibly apparent and detected structural patterns which were not apparent. In this study, the one dimensional scaling results and the clustered groups reinforced the structure imposed by ranking A. Future analysis may relate other dimensions to patterns among the subjects' problem solving processes.

The final implication is an outcome of the many plans, changes, observations, and facts which resulted during this investigation. Mathematical problem solving, being the complex behavior that it is, will not be easy to measure or assess with a single instrument. It appears that a written test may be feasible for predicting a subjects interview test score and ranks, but that further investigation by the thinking aloud procedure may be necessary to evaluate individual processes and strategies. assuming that the subject is able to verbalize while thinking. In situations where it is applicable, the thinking aloud procedure sometimes provides an incomplete record. Lucas (1972) suggested that retrospection be used to procure additional information about the missing behaviors although care would have to be taken not to give the subject any training or heuristic hints if such procedures were used. For the subjects who cannot verbalize well or who find that excessive interference occurs, some other procedure will have to be used to identify and record their mathematical problem solving processes.

Recommendations for Future Research

Like most exploratory studies, this investigation raised more questions than it answered. Future research could extend the efforts of this study or could investigate the new issues which were raised. Suggestions are included as the recommendations are discussed below.

The written test scores did not achieve a .71 correlation coefficient with the interview test scores, but the results were close enough to recommend that additional efforts be made to reach the desired coefficient level. The initial step is to increase test reliabilities and there are five procedures which could be tried:

1) Replicate the study with a large population.
2) Use a longer form of the written test. A two part test with a total of 40 or more items should be tried.
3) Use more mathematical problems on the interview test. Since the seventh graders took approximately 15 minutes to attempt the six IT problems, two or three more items could be included without tiring the subjects.
4) Use a revised scoring system. Lucas' system resulted in numerous ties in subjects total scores and sub-scores. Scoring which attaches large weights to Approach, Plan, and Result would better differentiate among subjects and might improve the correlation between written and interview test scores. For example, a subject might be awarded 0-2 points for Approach, 0-3 points for Plan, and 0-2 points for Result.
5) Screen the WT items and IT problems to remove those which have a poor correlation with test totals.

The interview test rankings developed in this investigation shared a strong rank order agreement with the written test rankings.

However, if a higher level of confidence is desired, new rankings might be developed. Subjects' performance on individual IT items and item difficulty could be considered in the development of new ranking schemes.

The thinking aloud procedure needs to be thoroughly examined before it is used for recording and assessing subjects' mathematical problem solving behaviors. Systematic application beginning with first graders and continuing through adults should detect general differences in ability to think aloud as the age or mental maturity of the subjects increases. A systematic approach might also uncover clues to explain why two subjects of the same age can vary greatly in their ability to verbalize. Future investigations must consider the effects of age level and individual differences before deciding to use the thinking aloud procedures.

The audio and video taping differences in recorded data were apparent. However, the differences in solution times and the differences in coding time ratios were based upon seventh graders protocols which were short and which contained relatively simple behaviors. Loomer's college students' solution times were much longer and the complex behaviors were more difficult to code. These observations raised suspicion that the differences in coding time ratios for the college students' protocols may not be consistent with the results of this study. Future studies might compare audio and video taping at different age levels to verify the solution and coding time differences.

Finally, future research should further examine the relationship of the multidimensional scaling and clustering procedures to mathematical problem solving assessment. In particular, the second and third dimensions of the scaling procedure need to be studied in order to see if problem solving behaviors, patterns, or factors can be related to them.

Comments

A simple instrument is needed to give educators a preliminary assessment of students' mathematical problem solving achievement. The written instrument which was devised for the purpose did not achieve the desired correlation coefficients, but the results came sufficiently close to make the investigator confident that the goal can be reached. Further research should complete the development of the written test and search for improved methods of assessing students' mathematical problem solving achievement.

REFERENCES **BEST COPY AVAILABLE**

Aiken, L.R. Language factors in learning mathematics. Mathematics education report. Columbus, Ohio: ERIC Information Analysis Center for Science, Mathematics, and Environmental Education, 1972.

Bourne, L.E., & Battig, W.F. Complex processes.In J. Sidowski (Ed.) Experimental Methods and Instrumentation in Psychology. New York: McGraw-Hill, 1966.

Broder, L.J., & Bloom, B.S. Problem solving processes of college students. Supplementary Educational Monographs, No. 73. Chicago: University of Chicago Press, 1950.

Claparède, E. La psychologie de l'intelligence. Scientia, 1917, 22, 353-368.

College Entrance Examination Board Commission on Mathematics. Program for college preparatory mathematics. Report. New York: College Entrance Examination Board, 1954.

Cronbach, L. Essentials of psychological testing. (3rd ed.) New York: Harper & Row, 1970.

Duncker, K. On problem solving. Washington: American Psychological Assoc., 1945, (Translated by Lynne P. Lees) Published in Berlin in 1935.

Durost, W.N., Bixler, H.H., Wrightstone, J.W., Prescott, G.A., & Balow, I.H. Metropolitan achievement tests: Advanced form F. New York: Harcourt Brace Jovanovich, 1970.

Durost, W.N., Bixler, H.H., Wrightstone, J.W., Prescott, G.A., & Balow, I.H. Metropolitan achievement tests: Teacher's handbook. New York: Harcourt Brace Jovanovich, 1971.

Dye, D. Report of the state mathematics assessment committee. St. Paul, Minn.: State Department of Education, 1972.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice Hall, 1972.

Fehr, H.F. The secondary school mathematics curriculum improvement study: A unified mathematics program. The Mathematics Teacher, 1974, 67, 25-33.

Goodman, L.A. & Kruskal, W.H. Measures of association for cross-classifications. Journal of the American Statistical Association, 1954, 49, 732-764.

Gray, G.F. An experimental study of introductory multiplication. (Doctoral Dissertation, Berkeley) Ann Arbor, Mich.: University Microfilms, 1964. No. 65-2996.

Green, B.F. Current trends in problem solving. In B.J. Kleinmuntz (Ed.) Problem solving: Research, method, and theory. New York: Wiley, 1966.

Hays, W.L. Statistics. New York: Holt, Rhinehart, and Winston, 1963.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, XII, 153-160.

Hubert, L.J. Monotone invariant clustering procedures. Psychometrika, 1973, 38, 49-55.

The Instructional Objectives Exchange. Instructional objectives exchange--Mathematics K-6. Los Angeles: Author, 1970.

Jerman, M. Instruction in problem solving and an analysis of structural variables that contribute to problem solving difficulty. Technical Report No. 180. Stanford: Institute for Mathematical Studies in Social Sciences, Stanford University, 1971.

Johnson, S.C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

Kendall, M.G. Rank correlation methods. London: Charles Griffin and Co., Ltd., 1955.

Kilpatrick, J. Analyzing the solution of work problems in mathematics: An exploratory study: (Doctoral dissertation, Stanford University) Ann Arbor, Mich.: University Microfilms, 1969. No. 68-6442.

Kilpatrick, J. Problem solving in Mathematics. Review of Educational Research, 1969, 39, 523-534.

Kilpatrick, J. Review of research in problem solving in mathematics. Paper presented at the meeting of the National Council of Teachers of Mathematics, Minneapolis, Minnesota, 1969.

Kleinmuntz, B.J. Problem solving: Research, method, and theory. New York: Wiley, 1966.

Kuder, D.J. & Richardson, M.W. The theory of the estimation of test reliability. Psychometrika, 1937, II, 151-160.

Lazerte, M.E. The development of problem-solving ability in arithmetic. Toronto: Clark-Irwin, 1933.

Lindquist, E.F., & Hieronymus, A.N. Iowa tests of basic skills. Boston: Houghton Mifflin, 1964.

Lingoes, J.C. The Guttman-Lingoes non-metric program series. Ann Arbor, Mich.: Maththesis Press, 1973.

Lord, F.M. and Novick, M.R. Statistical theory of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Lucas, J.F. An exploratory study in the diagnostic teaching of heuristic problem solving strategies in calculus. Unpublished doctoral dissertation, University of Wisconsin, 1972.

Luchins, A.S., & Luchins, E.H. New experimental attempts at preventing mechanization in problem solving. Journal of General Psychology, 1950, 42, 279-297.

Maier, N.R. Problem solving and creativity. Belmont, CA: Brooks/Cole Pub. Co., 1931.

Martin, M.D. Reading, comprehension, abstract verbal reasoning, and computation as factors in arithmetic problem solving. (Doctoral dissertation, University of Iowa, 1963) Dissertation Abstracts, 1964, 24, 4547-4548. No.11.

Menchinskaya, N.A. Intellectual activity in solving arithmetic problems. Proceedings (Izvestiya) of the Academy of Pedagogical Sciences of the RSFSR, 1946,3,99-134. Cited by J. Kilpatrick and I. Wirszup, Soviet studies in the psychology of learning and teaching mathematics. Stanford: School Mathematics Study Group, 1969.

Newell, A., Shaw, J.C., & Simon, H.A. Report on a general problem solving program. (Paper P. 1584) Skokie, Ill.: The Rand Corp.,1959.

Pereira, W.C. The relationship between girls' overt verbalization, performance, retention, rules and strategies as they learn a mathematical structure: A study based on elements of a potential theory which relates thinking, language, and learning. Technical Report No. 293. Madison: Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, 1973.

Polya, G. How to solve it. (2nd ed) New York: Doubleday, 1957.

Polya, G. Mathematical discovery. Vol I. New York: Wiley, 1962. 2 vols.

Riedesel, A.C. Problem solving: Some suggestions from research. Arithmetic Teacher, 1969, 16, 54-58.

Ruger, H.A. The psychology of efficiency. Arch. Psychology, 1910, No. 15.

School Mathematics Study Group. Soviet studies in the psychology of learning and teaching mathematics. J. Kilpatrick & I. Wirszup (Eds.). Vol. III, Stanford: School Mathematics Study Group, 1969. 5 volumes.

Suydam, M.N. An evaluation of journal-published research reports on elementary school mathematics: 1900-1965. (Doctoral dissertation, the Pennsylvania State University) Ann Arbor, Mich.: University Microfilms, 1967. No. 68-3563. 1967.

Suydam, M.N. & Weaver, J.F. Research on mathematics education (K-12) reported in 1970. Journal for Research in Mathematics Education, 1971, 2, 257-298.

Suydam, M.N. & Weaver, J.F. Research on mathematics education (K-12) reported in 1971. Journal for Research in Mathematics Education, 1972, 3, 196-232.

Thompson, B.S. The interpretative report on the 1973 Wisconsin state mathematics assessment. (Vol. 1, Report 3) Madison, Wis.: Department of Public Instruction, 1974.

Thompson, E.N. Readability and accessory remarks: Factors in problem solving in arithmetic: (Doctoral dissertation, Stanford, 1967) Dissertation Abstracts, 1968, 28, 246 2465A. No. 7.

Thorpe, L.P., Lefever, L.W., & Naslund, R.A. SRA achievement series. Chicago: Science Research Associates, Inc., 1964.

Tiegs, E.W. & Clark, W.W. California achievement tests: Complete battery. Monterey, Cal. CTB/McGraw-Hill, 1970.

Tiegs, E.W. & Clark, W.W. California short-form test of mental maturity. Monterey, Cal.: CTB/McGraw Hill, 1963.

Torgerson, W.J. Theory and methods of scaling. New York: Wiley, 1958.

Van Engen, H., Hartung, M. L., Trimble, H. C., Berger, E. J., Cleveland, R. W., Gibbs, E. G., Stochl, J. E., & Walch, R. First course-Mathematics, concepts, applications. Glenview, Ill.: Scott Foresman, 1969.

Appendix A

KILPATRICK'S CODING FORM FOR PROBLEM-SOLVING PROTOCOLS

Subject No. ________ Coder ________ Tape Readings ________

Problem No. ________ Date ________ Time ________

Score ________

PREPARATION

___ Draws figure

___ Changes condition (spec./gen.)

___ Performs exploratory manipulation

RECALL

___ Recalls same or related problem

___ Uses related problem in solution

___ Says he has forgotten procedure

PRODUCTION

___ Uses successive approximation

___ Misinterprets problem

___ Selects solution on irrelevant basis

EVALUATION

___ Checks solution by subst. in equation

___ Checks that solution satisfies condition

___ Checks solution by retracing steps

___ Checks solution is reasonable/realistic

___ Derives solution by another method

COMMENTS ABOUT SOLUTION

___ Questions existence of solution

___ Questions uniqueness of solution

___ Questions necessity/relevance of information

___ Expresses uncertainty about final solution.

___ Says he doesn't know how to solve problems

REQUESTS

___ Requests assistance, more information

___ Requests verification

COMMENTS

___ Expresses enjoyment, liking for problems

___ Expresses distaste, dislike for problems

___ Admits confusion

___ Shows concern for performance

___ Says procedure unorthodox

___ Says he can't explain result

EXECUTIVE ERRORS	Tallies	Total
Count/arith.oper.	____	____
Alg. Manipulation	____	____
Other slips	____	____

PROCESS SEQUENCE:

Appendix A (cont'd)

Process Symbols

PREPARATION

R = Reading and trying to understand problem

PRODUCTION

D = Deduction from condition

E = Setting up equation

T = Trial and error

EVALUATION

C = Checking solution

OUTCOMES OF PRODUCTION (used in conjunction with D, E, and T)

1 = Incomplete

2 = Impasse

3 = Intermediate result

4 = Incorrect result

5 = correct result

MODIFIERS

Bar over symbol = Structural error in process (used only with symbols for production)

Underlined symbol = Difficulty (hesitation, repetition) in process

PUNCTUATION MARKS

, Inserted between successive processes

/ Work stopped without solution

. Work stopped with solution

Appendix B

LUCAS' PROCESS-SEQUENCE CODES

Process Symbols

- R = reads the problem
- S = separates/summarizes data
- M_f = introduces model by means of a diagram
- $M_{f'}$ = modifies existing diagram
- M_{f_c} = introduces diagram with coordinate system imposed
- DS = deduction by synthesis
- DA = deduction by analysis
- T = trial and error: successive approximation
- An = reasoning by analogy
- Me = model introduced by means of equation, expression, or other relationship
- Alg = algorithmic process
- N = not classifiable
- C = checks the result
- V_s = varies the process (condenses/outlines; tries different method)
- V_m = varies the problem (by analogy; by changing conditions)

Outcomes of DS, DA, T Processes

- 1 = abandons process
- 2 = impasse
- 3 = incorrect final result
- 4 = correct final result
- 5 = intermediate result (correct or incorrect)

Appendix B (cont'd)

LUCAS' CODING FORM

Subject No. ____
Problem No. ____
Coder ____
Date ____

Tape No. ____
Tape Readings ____

Time: exc. looking back ____
looking back ____
total ____

Score: approach ____
plan ____
result ____
total ____

Approach
- restates problem in own words ____
- mnemonic notation ____
- representative diagram yes ____
- no ____
- auxiliary line(s) ____
- enlarges focal points ____

Production
- recalls related problem ____
- uses method of related problem ____
- uses result of related problem ____
- inductive reasoning (pattern search) ____

Looking Back
- routine check of manipulations ____
- is result reasonable? ____
- all information used? ____
- checking test for symmetry ____
- test of dimensions ____
- specialization (extreme cases) ____
- comparison with gen. known result ____

V_s
- condenses/outlines process ____
- tries to derive differently ____

V_m
- variation by analogy ____
- variation by changing conditions ____

Executive errors
- algebraic manulations ____
- numerical computation ____

Interviewer Comments
- Subject asks question about problem ____
- Subject comments about problem ____

PROCESS SEQUENCE

Punctuation Marks

- \- (dash) hesitation of approximately 2 units (30 seconds)
- () scope of DS, DA, or T process
- , inserted between successive processes
- / stops without solution
- . stops with solution (correct or incorrect)

Errors

- ↧ over process symbol = structural error in process
- ↓ over process symbol = executive error in process
- *↧ *↓ (asterisk over error symbol) = previous error of type indicated was corrected

INTERVIEW TEST ITEM POOL WITH ANSWERS

1. A farmer has a total of 39 chickens and cows in his barn. If you counted all the legs of these animals, you would get 100 legs. How many chickens does he have?

(28)

2. The average weight of Billy, Willy, and Ted is 125 pounds. Billy weighs 110 pounds and Willy weighs 120 pounds. How much does Ted weigh?

(145 pounds)

3. Mr. Director had trouble arranging his band. When he put 2 people in each row, there was one person extra. When he put 3 people in a row, there were two extra. With 4 people in a row, there were three extra. Finally, he put 5 people in a row, but then there were four extra members. How many people could there have been in his band?

(any answer of the form 59 + 60n, n=D,1,2,...)

4. If you could buy oranges at a price of 4 for 25 onets and sell them at 3 for 25 cents, how many oranges would you have to buy and sell in order to make a profit of one dollar?

(48)

5. One hundred students were divided into three groups, Group A had as many people as Group B and Group C had together. Group B had six more students than Group C had. How many students were in Group C?

(22)

6. A frustrated frog fell to the bottom of a thirty foot deep well. Every day he managed to climb up four feet but every night he slipped back three feet. How many days did it take the frog to reach the top of the well?

(27)

7. A ship leaves New York for London at noon each day, and each day at noon a ship starts from London to New York. The trip across the ocean takes exactly three days. If you left on a ship from New York at noon on Monday, how many ships from London would you see by the end of your trip on Thursday noon?

(7)

8. Mr. Carpenter makes only three-legged stools and four-legged tables. He used 60 legs to make twice as many stools as tables. How many stools did he make?
(12)

9. On Monday, John bought a mototbike for $60. On Wednesday, he sold it to his friend Paul for $70. On Friday, John bought the bike back from Paul for $80. and sold it to his brother Craig for $90. How much money did John make or lose for all his work, or did he come out even?
(Made $20.)

10. In a television survey concerning two programs, 350 people said that they enjoyed program X, 400 said that they enjoyed program Y, and 200 said they enjoyed both programs. What is the least number of people that could have been interviewed in this survey?
(550)

11. On one television station, they show one minute of ads and then five minutes of the program. At this rate, how many minutes of commercials do they show in three hours?
(30)

12. Midge was planning to join a hike to raise money for charity. Midge's mother promised to pay her ten cents for each mile she walked and her brother Jim promised to pay a certain amount for each mile too. If Midge marched 25 miles and collected a total of four dollars from her brother and mother together, how much did Jim pay her for each mile?
(6 cents)

13. Mr. Stout weighed 300 pounds, so he went on a diet. The first week he lost ten pounds, but then became careless and gained back five pounds the next week. The third week he lost ten pounds again, but the fourth week gained back five pounds. If he kept this strange diet, after how many weeks would he first weigh 250 pounds?
(9)

14. Joe's sister Susan is nine years older than he is. In three years, Susan will be twice as old as Joe will be. How old is Joe now?
(6 years)

15. An ostrich egg weighs about 3 pounds. A hen's egg weighs about 2 ounces. It would take 400 hummingbird eggs to weigh as much as a hen's egg. How many hummingbird eggs would it take to weigh as much as one ostrich egg?

(9600)

16. Jack has six coins. One third of his coins are dimes, but they are worth one fourth of the total value of the coins. What coins does Jack have?

(2 dimes, 2 quarters, 2 nickels)

17. Janet had 69 cents. Shelly asked her for change for a half dollar. Janet tried to make the change, but found that she didn't have have the right coins to do it. What coins did she have if each coin was less than a half dollar?

(4 dimes, 4pennies, and 1 quarter)

18. A dozen cookies and two loaves of bread costs $1.20. Two dozen cookies and a loaf of bread costs $1.26. How much does one loaf of bread cost?

(38 cents)

19. Pete the Pirate buried ½ of his sack of gold coins and spent 1/3 of his sack of gold coins. Then he had 300 coins left. How many gold coins did Pete have before he buried or spent any?

(1800)

20. Two adult tickets and one child's ticket for a movie cost $6.25. Two adult tickets and three chilren's cost $8.75. What is the cost of one adult ticket?

($2.50)

21. Suppose you could fill an old bucket with water in 40 seconds. Then it springs a leak and all the water drains out in 120 seconds. How many seconds will it take you to refill the bucket now that it has the leak?

(60)

22. Mr. Ketchum wants to cut a 70 yard long piece of fish line into three parts. The second piece should be twice as long as the first piece, and the third piece should be twice as long as the second piece. How many feet long should the third piece be?

(40)

23. A candy producer puts a blue ticket good for one free bar in every 80th candy bar he produces and a red ticket good for two gree bars in every 180th bar. Which candy bar was the first one with both a red and a blue ticket in it?

(720th)

24. The Restful Hotel receives its glasses in full cartons of 40 glasses each and the Towers Hotel gets its glasses in full cartons of 24 glasses each. One time, they both ordered the same number of glasses and both got all full cartons to fill the order. What is the smallest number of glasses they could order for this to happen?
(120)

25. The P.T.A. raised $40. at a bake sale. Cakes were $1.50 each and pies were $1.00 each. Twice as many cakes as pies were sold. How many cakes were sold?
(20)

26. A six-pack of eight ounce bottles of pop costs 60 cents. At this rate, how much should an eight-pack of sixteen ounce bottles cost? (Don't count the deposit for bottles)
($1.60)

27. There once was a country where a chicken was worth 1/10 as much as a pig and a pig was worth 1/10 as much as a cow. A farmer who owned 8 hens, 7 pigs, and 2 cows decided to trade his pigs and cows in for hens. How many hens did he have after the trade?
(278)

28. It takes 96 square inches of paper to wrap without overlapping a box shaped like a cube. How many cubic inches of space are inside the box?

29. The Girl Scouts wanted to sell 600 boxes of cookies. The number of boxes each troup had to sell depended on the number of members it has. How many boxes of cookies should Troop 3 sell to do its share?
(150)

Troop 1	20 scouts
Troop 2	35 scouts
Troop 3	25 scouts
Troop 4	20 scouts

30. The junior high school band marched in rows with the same number in each row and there were three marchers left over. When eight more marchers joined the band in marching with the same size rows as before, there were two marchers left over. How many marchers were in each row?
(9)

31. On Tuesday, the phy ed teacher divided the class into eight teams to get the same number on each team. On Thursday, three more students came. Then he made seven teams in order for there to be an equal number of students on each team. How many students could have been in class on Tuesday?
(any answer of the form 32 + 56n, n=0,1,2,..)

32. Mr. Shopper goes to the store once every two days and his neighbor Mr. Buyer goes to the same store once every five days. On Friday, the two men meet at the store. On what day of the week will both men meet at the store again?

(Monday)

33. There are 35 girls and 28 boys at the seventh grade field day. They join into teams so that there are both boys and girls on each team. To keep the teams even, there has to be the same number of boys on each team and the same number of girls on each team. How many boy-girl teams should there be so that everyone gets to be on a team?

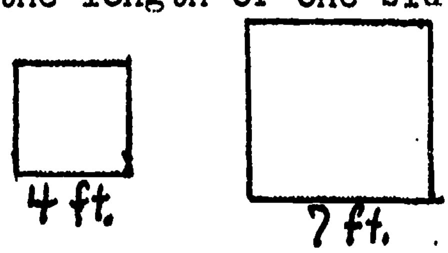
(7)

34. Sixty wooden cubes measuring one inch on a side are glued together to form one big solid block. When the big block is painted, six of the little blocks don't get any paint on them because they have blocks glued to all sides of them. How many inches long, wide, and high is the big block?

(5" x 4" x 3")

35. A large square has an area equal to the sum of the areas of the two smaller squares. To the nearest foot, what is the length of one side of the large square?

(8 ft.)



36. A fireman stood on the middle rung of a ladder, directing water into a burning building. As the smoke lessened, he stepped up three rungs. A sudden flare-up forced him to go down five rungs. Later he climbed up seven rungs and worked there until the fire was out. Then he climbed the remaining six rungs to the top of the ladder and entered the building. How many rungs did the whole ladder have?

(23)

37. On a balance scale (like a teeter-totter), a brick on one side balances evenly with one third of a brick and a one pound weight on the other side. What is the weight of one brick?

($1\frac{1}{2}$ pounds)

38. A barrel full of oil weighs 50 pounds. The same barrel filled with gasoline weighs 35 pounds. If oil is twice as heavy as gasoline, how much does the barrel weigh if it is empty?

(20 pounds)

39. The egg man sent a bill for 24 dozen eggs, but the first and last digits were missing. If eggs cost less than one dollar a dozen, how much should the bill be?

Bill for eggs 24 dozen $_2.4_

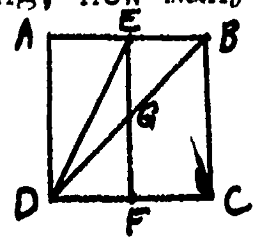
($12.48)

40. Two pirates found a bag of gold coins and agreed to split it in the morning. After they went to bed, the first pirate got up and took one third of the coins. Later, the second pirate got up and took one half of the coins that were left. In the morning, there were still 200 coins left. How many coins were there before either pirate sneaked any out?

(600)

41. ABCD is a square with E halfway between A & B and F halfway between D and C. If each side of the square is ten inches long, how many square inches are in triangle DEG?

($12\frac{1}{2}$)



42. The Yum-Yum ice cream man has vanilla, chocolate, and strawberry ice cream. He has marshmallow, fudge, coconut, and peanut toppings. If he uses two scoops of ice cream and one kind of topping for each sundae, how many different kinds of sundaes can he make?

(24)

43. Tom spent one dollar for his lunch. He spent 20 cents more for french fries than he did for pop, and he spent 15 cents more for a hamberger than he did for the french fries. How much did the hamberger cost him?

(50cents)

44. Mr. Butcher mixes two pounds of fat with eight pounds of lean meat when making hamburger. The lean meat is worth $1.20 a pound, but Mr. Butcher only charges $1.10 a pound for the hamberger and he still makes ten cents profit on each pound. How much is each pound of fat worth?

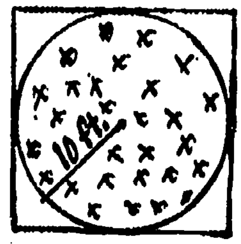
(20 cents)

45. Mr. Hasty forgot his brief case when he left town. An hour later, his son jumped on a motor cycle to catch him. If Mr. Hasty drives 50 miles per hour and his son drives 60 miles per hour, how long will it take the son to catch up with him?
(5 hours)

46. Car A gets twenty miles to a gallon and car B gets sixteen miles to a gallon. Both cars are taking a trip of the same distance and it is found that both cars used a whole number of gallons of gasoline. How many miles long could the trip have been?

(Any answer of the form 80n, n=1,2,3,...)

47. A new round rug was put on a square floor. The radius (distance from center to edge) of the rug was 10 feet and the material covered about 314 square feet of the floor. About how many square feet were not covered by the rug?



(86)

48. Hot dogs cost ten cents each and buns cost five cents each. How much should the art club sell a hot dog in a bun for if they want to make twenty dollars profit on five hundred sandwiches?

(19 cents)

49. A long freight train was moving 15 miles an hour on the tracks parallel to a highway. It took an auto 4 minutes from the time it was even with the caboose to the time it passed the engine. If the auto was going 30 miles an hour, how long was the train?

(1 mile)

50. Wilma is running 6 yards a second and is 120 yards from the finish line. Dorla is 40 yards behind Wilma. How many yards a second will Dorla have to run to tie Wilma?

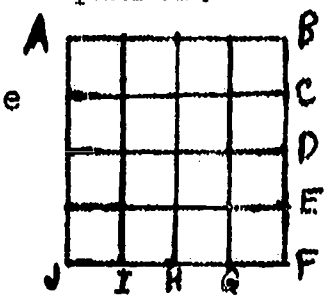
(8)

Appendix D

WRITTEN TEST ITEMS WITH ANSWERS

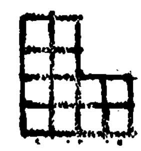
1. ABFJ is a square divided into equal smaller squares. Draw a segment from point A to one of the other named points so that the area on one side of the segment will be three times the area on the other side of the segment.

   (to D or to H)

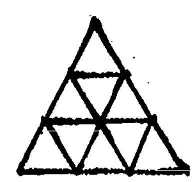


2. When you buy stamps at the post office, their edges are usually attached to each other. In how many different ways can three stamps be attached to each other?
   (6)

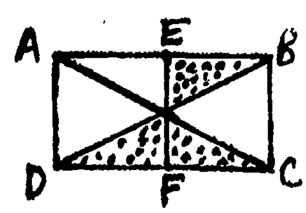
3. How many squares are there in the diagram at the right? Include those which overlap.
   (17)



4. How many triangles are there in the diagram at the right? Include those which overlap.
   (13)



5. If E is the midpoint of AB and F is the midpoint of DC, what fractional part of the rectangle ABCD is spotted?
   (3/8)



6. What number comes next in 1,2,4,7,11,__ ?
   (16)

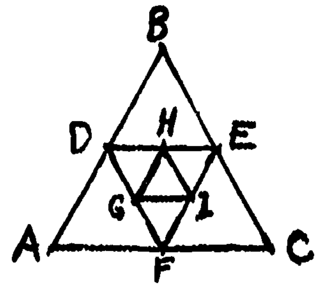
7. A class of 30 students was divided into two groups. One group had eight more students than the other. How many students were in the larger group?
   (19)

8. Using pennies, nickels, dimes, or a combination of the coins, how many different ways could a person make change for a quarter?
   (12)

9. The perimeter (distance around) of a swimming pool in the shape of a rectangle is 148 feet. If the length of the pool in 50 feet, how many square feet of surface does the pool have?
   (1200)

10. A mouse wants to get to its house, but it has to go through two walls to get there. If the first wall has four holes and the second wall has three holes, how many different paths can it take to get to its house?
(12)

11. Triangle ABC has all sides equal. If the area of the little triangle HGI is 5 square inches, what is the area of ABC? (D, E, F, G, H, and I are all midpoints.)
(80 square inches)



12. How many ounces are in one gallon?

1 cup	= 8 ounces
2 cups	= 1 pint
2 pints	= 1 quart
4 quarts	= 1 gallon

(128)

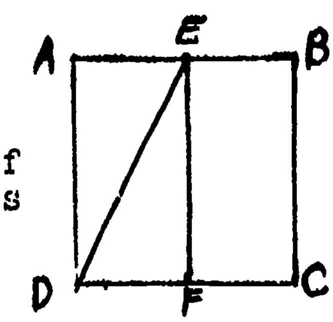
13. A race horse runs about 30 miles per hour. How many feet does it run in one minute? (5,280 feet in 1 mile.)
(2640)

14. Dr. Curem charges ten dollars for the first visit and five dollars for each visit after that. If Mr. Ailings' bill was one hundred dollars, how many visits did he make?
(19)

15. A pen costs a dollar more than an eraser. Together they cost $1.10. How much does the eraser cost?
(5 cents)

16. What whole number for "a" will make $a \cdot b + a \cdot c = 56$ if b is 3 and c is 4?
(8)

17. ABCD is a square with E halfway between A and B and F halfway between D and C. If each side of the square is ten inches, what is the area of triangle DEF?
(25 square inches)



18. There was half of Mom's apple pie left. Then Nate ate one half of the half and Kate ate one half of what Nate left. What part of the pie was left after Kate ate?
(1/8)

19. Fran gave Jan half of her cookies and another cookie besides. Fran had seven cookies laft. How many cookies did Fran give to Jan?
(9)

BEST COPY AVAILABLE

20. Mr. Baker's recipe for cookies needs $\frac{1}{2}$ cup of sugar and two eggs. He is making a bigger batch of cookies, so he used $2\frac{1}{4}$ cups of sugar. How many eggs should he use?
(9)

21. In one school, there are five girls to every four boys. If there are 1 hundred boys in school, how many girls are there in the school?
(125)

22. If 76 cookies fill five boxes with six cookies left over, how many of the same sized boxes will 100 cookies fill?
(7)

23. It takes thirty chocolate chip cookies to fill two thirds of a box. How many chocolate chip cookies would be needed to fill the whole box with them?
(45)

24. There were 18 brown eyed students on the bus and 12 students had brown hair. If there was a total of 26 students on the bus, what is the smallest possible number of students that had both brown eyes and brown hair?
(4)

25. Jean has four different sweatshirts and five different pants. How many different outfits with one sweatshirt and one pair of pants each could she make?
(20)

26. The Yum-Yum ice cream man has vanilla, chocolate, and strawberry ice cream. He has marshmallow, fudge, peanut, and coconut toppings. How many different kinds of sundaes can he make if he only uses one kind of ice cream and one kind of topping for each sundae?
(12)

27. One small country has very few cars in it, so they use only a one digit number followed by one letter of the alphabet for their license plates. How many different license plates can they make?
(260)

28. Two test car drivers departed from the car company at the same time, but they drove away in opposite directions. The driver of car C averaged 60 miles per hour and the driver of car F averaged 40 miles per hour. How many hours was it before they were 600 miles apart?
(6)

29. On a travel tour, the Tripp family drove eight hours the first day, five hours the second day, and seven hours the third day. Their average speed was the same each day and they traveled a total of 1000 miles. How far did they

travel the second day?
(250 miles)

30. Who is the shortest player of the team?

Players' heights

Lee is 5 feet.
Jerry is 63 inches.
Wilt is 2 yards.
Cazzie is 1yard, 2 feet, and 3 inches.
Lou is 3 feet and 30 inches.

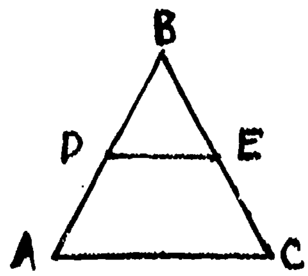
(Lee)

31. Four people are going to sit by a square table, one at each side. How many different seating arrangements are possible?
(24)

32. Sandy has a red book, a blue, a yellow one, and a green one. She wants to place them in an empty shelf of a bookcase. In how many different orders could she arrange the books?
(24)

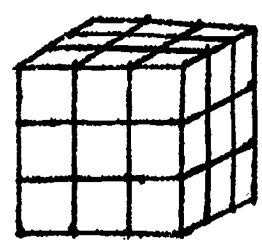
33. Three pounds and 8 ounces of hamburger costs $2.80. How much does one pound of hamburger cost?
(80 cents)

34. Triangle ABC has all sides equal. Point D is the midpoint of AB and E is the midpoint of BC. If the area of triangle ABC is 48 square inches, what is the area of figure ADEC?
(36 square inches)



35. The perimeter (distance around) of a rectangular flower garden is 60 feet. There is a 2½ foot wide sidewalk around the garden. What is the perimeter of the outer edge of the sidewalk?
(80 feet)

36. The large cube was painted red on all sides and then cut up into 27 smaller cubes. How many of the smaller cubes have exactly two red sides?
(12)



37. Nancy spent two fifths of her money for a sweatshirt. If the shirt cost four dollars, how many dollars did Nancy have after she bought the shirt?
(6)

38. Mixing four gallons of alcohol with twelve gallons of water makes a solution which is one fourth alcohol. If four more gallons of alcohol were added to the solution, then what fractional part wold be alcohol?
(two fifths)

39. Mr. Driver gets 20 miles to the gallon with his compact car. If he drives three hours at 60 miles per hour, how many gallons of gas does he use?
(9)

40. A bell is made of a special metal which has 3 parts of copper for each part of tin. How many pounds of tin are in a bell that weighs 3000 pounds?
(750)

41. Roberto gets about two hits for every six times he gets to bat. How many times would he have to bat in order to get 150 hits?
(450)

42. Here is a line segment AB. A__________B If you put two more points C and D on the segment so that no points are the same, how many segments will there be?
(6)

43. What is the greatest number of angles less than 180 degrees that is possible when three lines cross at the same point?
(12)

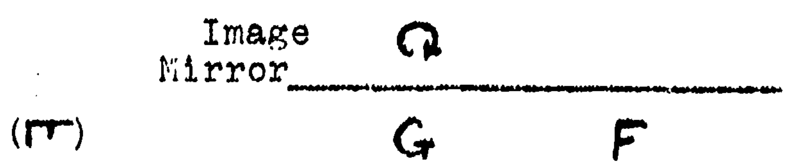
44. The population of Boom Town has doubled every five years for the last 20 years. It had 400 people in 1970. By what year will the population reach 12,800 if it continues growing at the same rate?
(1995)

45. One cell divides into two cells every five seconds. If you started with 5 cells, how long would it take to have over 1,000 cells?
(40 sec.)

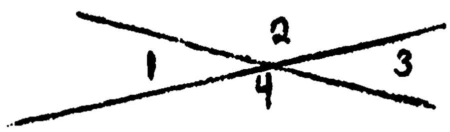
46. Two lines can cross at only one point, but three lines can cross at three points. What is the most points at which five lines can cross?
(10)

47. For every two dollars Jenny earned towards a new bicycle, her father gave her one dollar more. How much money would Jenny's father end up giving her if she wanted to work until she had enough for a bicycle that costs $45?
(15)

48. This is a funny mirror. Look what it does to the letter G. Draw in the image of the letter F.



49. Two lines divide a plane into four separate (non-overlapping) areas. What is the largest number of separate areas that four lines can divide a plane into?
(11)



50. Using the edges of a cube as lines, how many pairs of parallel lines are there?
(18)

51. Lance, Larry, and Lena agreed to split the money they earned for doing errands. Lance earned \$1.75 and Lena earned \$2.75, but after the split each person got \$2.00. How much money did Larry earn before they divided up the money?
(\$1.50)

52. If six bushels of wheat will plant four acres, how many bushels of wheat are needed to plant 30 acres?
(45)

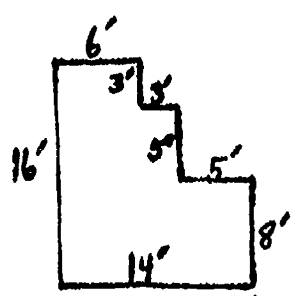
53. If 24 chocolates fill 3/4 of a box, how many will it take to fill the whole box? (32)

54. If 1*2=3, 1*3=4, 2*3=7, and 3*4=13, how much is 4*5?
(21)

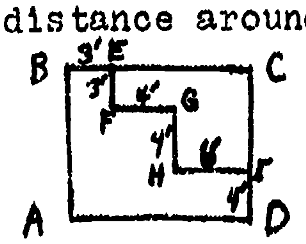
55. The perimeter (distance around) of a square is 40 inches. What is its area?
(100 sq. in.)

56. The perimeter (distance around) of a rectangle is 30 inches. If the width is six inches, what is its area?
(54 sq. in.)

57. What is the area of this figure?
(175 sq. ft.)



58. What is the perimeter (distance around) of the rectangle ABCD?
(48 ft.)



59. The formula for finding the area of a circle is $A=\pi r^2$ where r is the radius of the circle. How many times larger does the area of a circle become of you make its radius twice as long?
(4)

60. The minute hand on a clock makes one complete turn (360 degrees) in one hour. How many degrees does the hour hand turn in one hour?
(30)

61. Rachel went to a sale where bicycles were selling for 1/3 off the regular price. She paid $40. for a new 3-speed bike. How much was the bike before the sale?
(60.)

62. Forty seventh graders were divided into two groups so that the larger group had six more students than the smaller group. How many students were in the larger group?
(23)

63. Two numbers c and b have a sum of 90. If c is twice as large as b, what number is c?
(60)

64. The band director had the members march with three in each row, then with four in each row, and finally with five people in each row. In each case, there were no extra people left over. What is the smallest number of members this band could have?
(60)

65. If you mix eight pounds of meat worh one dollar a pound with two pounds of soybeans worth 25 cents a pound, how much a pound should you charge for the mixture?
(85¢)

66. Alex walks to school. After walking 2/3 of the way, he still has 1/4 of a mile to go. How far is his school from home?
(3/4 mile)

67. On a map, three and one half inches represents 70 miles. How many miles does six inches represent on this map?
(120)

68. If you painted all the sides of a certain sized cube, you would paint 600 square inches of surface. How long is one side of the cube?
(10 in.)

69. Scrooge has n nickels and 3n dimes. How many cents is the total value of the dimes and the nickels together?
(35n)

70. When Vincent answered 60 questions correctly on a test, he had 4/5 of the answers right. How many questions were on the test?
(75)

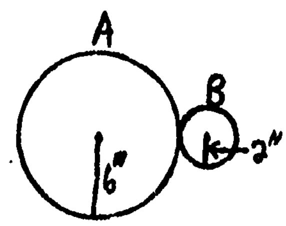
71. Wausau is 150 miles from Madison. A truck traveling at 40 miles per hour leaves Wausau towards Madison at the same time a car averaging 60 miles per hour leaves Madison to Wausau. How many miles will the truck travel before it meets the car if they travel on the same road?
(60)

72. Lora got scores of 63, 72, and 65 on her first three tests. What score must she get on her fourth test in order to end up with an average of 70 for the four tests?
(80)

73. Mr. Racer drives two hours at 50 miles per hour and three hours at 60 miles per hour. What is his average speed for the five hours?
(56 mph)

74. A tree has a 24 foot long shadow while a 12 inch ruler standing next to the tree has a four inch long shadow. How tall is the tree?
(72 ft.)

75. Gear A has a radius of six inches and gear B has a radius of two inches. If gear A makes 5 turns, how many turns will gear B make?
(15)



76. What whole number must m be in order for $\frac{100}{m}$ be in order for $\frac{100}{m}$ to be fractional number between 12 and 13?
(8)

77. Polly Hiker takes five steps to walk over three squares of cement in the sidewalk. How many squares could she cover if she took 150 steps?
(90)

73. A box holds three pounds of mint candy. If we made the box twice as long, twice as wide, and twice as deep, how many pounds of mint candy could it hold?
(24)

79. There are 25 students in third hour science class and 35 students in fifth hour English class. When the two classes are put together, there are 52 students. How many students from the science class are also in the English class?
(8)

80. Two numbers m and n have a sum of 80. If m is four times as large as n, what number is m?
(64)

81. What is the largest number that can divide into both 80 and 144 without leaving any remainders except zero?
(16)

82. Six girls belong to the basketball team but only five can play at a time. How many different groups of five players could be formed by the six girls?
(6)

83. Here are four sections of chain. It costs 15 cents to cut a link open and 25 cents to weld a link shut. What is the least it would cost to make a bracelet using all of these sections?
($1.20)

84. Four chickens lay six eggs in two days. At this rate, how many eggs could eight chickens lay in four days?
(24)

85. The number ab4 divided by 13 gives an answer of cd and a remainder of zero. What digit does d have to be for this to happen? (The letters a, b, c, and d all represent digits.)
(8)

$$13\overline{)ab4}\ \ \text{cd}$$

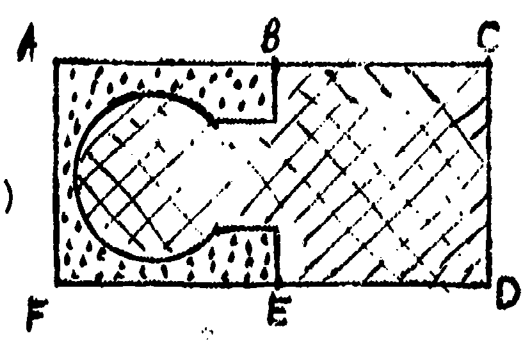
86. Five students are running for class president and vice president. The one with most votes is president and the student with the second most votes is vice-president. How many different combinations of president and vice-president are possible?
(20)

87. Each of John's five marbles is a different color. He chooses two marbles to play a game. How many different pairs of marbles are possible to be chosen?
(10)

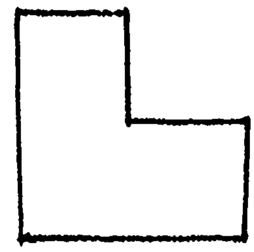
88. When numbering the pages of a book, a printer uses the digits (0, 1, 2, ---9) together to form larger numbers like 94 or 617. If a printer used 51 of the digits for a small book, how many pages did it have?
(30)

89. If 1 * 1 = 3, 1 * 2 = 4, 2 * 3 = 6, and 3 * 4 = 8, what does 4 * 5 equal?
(10)

90. Squares ABEF and BCDE are the same size. The perimeter (distance around) of the spotted area is 50 ft. while the distance from G to J (through H & I) is 15 ft. How many feet is the perimeter of the shaded area?
(50)

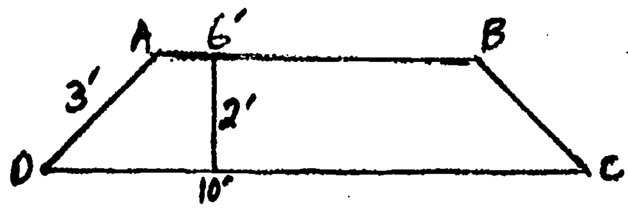


91. A man owned three connected squares of land and wanted to divide it among his four children. Draw lines to show how he could divide up the land so each child gets an equal share.
(many solutions)



92. Twenty-five marbles are in a sack. Eight marbles are blue, ten are green, and the rest are red. If I take out two marbles without looking and they are two different colors, what two colors are they most likely to be?
(blue and green)

93. The figure ABCD has side AB parallel to side CD. What is the area of the figure?
(16 sq. ft.)

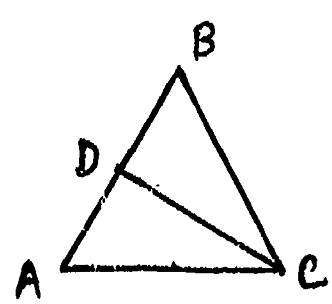


94. Diane's bus left Wausau at 1:40 and arrived in Madison at 4:15. How many minutes long was her bus ride?
(155)

95. Jim left Racine at 3:20 and took one hour and fifty minutes to drive to Madison. What time did he arrive in Madison?
(5:10)

96. Julie painted the entire surface of a board three feet long, ten inches wide, and one inch thick. How many square inches of surface did she paint?
(812)

97. Towns A, B, and C are all ten miles apart. Town D is half-way between A and B and is about eight and a half miles from C. If you lives in town D and wanted to visit all three other towns, one day, what is the smallest number of miles you would need to travel?
(30)



98. About two thirds of a fish can be eaten, the rest is waste. How many pounds of fish must Mr. Angler catch in order to have 12 pounds to eat:
(18)

99. Jeremy paid $10 for 100 hot dogs and $5 for 100 buns. If he wants to make five dollars profit when selling sandwiches. How much should he charge for each hot dog in a bun?
(20¢)

100. The pottery Club sold 20 dozen cookies for 60 cents a dozen. If it cost 35 cents a dozen to make the cookies, how much total profit did they gain?
($5.00)

101. According to the tax table, how much would you end up paying for a taxable item priced at $3.59?
($3.73)

102. For a new blanket Oscar paid $5.69 including tax. What was the price of the blanket before tax?
($5.47)

Tax table

Cost	Tax
0-12¢	0
13-37¢	1¢
38-62¢	2¢
63-87¢	3¢
88-99¢	4¢
For each dollar	4¢

(For items 101 & 102)

103. In a class of thirty students, 20 students wore shoes and 10 wore sandals. If half of the class is boys, what is the least possible number of boys wearing shoes?
(5)

104. Jack gave half of his money to Jill. Then Jill gave half of the money she got from Jack to Jane. After Jane spent ten cents of the money from Jill, she had a quarter left. How much money did Jack have before he gave any away?
($1.40)

105. Candy bars cost ten cents each if you buy them separately or three for a quarter if you buy them in groups of three. How much would you save on two dozen candy bars if you bought them in groups of three instead of separately?
($0.40)

106. The Mathematics Club has four committees of two people each. Members may belong to more than one committee, but no two committees have the same people working together. What is the smallest number of people that could belong to the Mathematics Club?
(4)

107. N is a number on the number line half way between 1/2 and 3/4. What number is N?
(5/8)

108. Paul has 60 different baseball cards and Jim has 50 different baseball cards. Twenty of Paul's cards have the same players that Jim has. How many different players do Paul and Jim have together?
(90)

109. Mr. Grocer has 5 pounds and 2 ounces of sunflower seeds to put into 2 ounce bags. How many bags of sunflower seeds can he make?
(41)

110. D is the decimal number half way between .5 and .6. What number is D?
(.55)

111. Which player has the best record when you consider both shots attempted and shots made?
(Art)

Name	Attempted	Made
Ray	12	7
Art	8	5
Luke	10	5
Rod	8	4
Eli	15	7

(For item 111.)

112. Scott threw 60 passes and completed 25 of them. Jim has thrown only 36 passes, but has completed the same percentage of them as Scott has. How many passes has Jim completed?
(15)

113. Joe has completed 25 passes in 60 attempts while Jerry has completed 9 passes in 20 attempts, and Rudy has completed 11 passes in 25 attempts. Which passer has the best record?
(Jerry)

114. Rent-a-car charges $7.00 a day plus ten cents a mile. If Mr. Salesman's bill for 6 days was $79.80, how many miles did he travel?
(378)

115. One Tuesday, the temperature reached 25 degrees above zero at noon and dropped to 19 degrees below zero at night. The next day, the temperature at noon was half way between Tuesday's warmest and coldest readings. What was the temperature at noon on Wednesday?
(3° above)

116. Jess weighs 175 pounds and Marsha weighs 113 pounds. If Neil's weight is half way between the two weights, How much does he weigh?
(144 lbs.)

117. Lucy had five yards of ribbon. Snoopy bit off sixteen inches of it, Peanuts took two feet of it, and Charlie took two yards of it. How much ribbon did Lucy have left?
(1 8/9 yds., OR 1 yd. 2 ft. 8 in. OR 68 in.)

118. Maud earns $2.10 an hour. How much money does she earn in ten minutes?
($0.35)

119. Wes earned $2.00 for working one hour and fifteen minutes. How much would he earn in an hour?
($1.60)

120. Material costs $1.80 a yard. How much would five feet and three inches of material cost?
($3.15)

121. Scott cut a five yard and two foot pole into halves. How long was each piece?
(8½ ft.)

122. Marsha works three hours and forty-five minutes on her part time job after school each day. How many hours does she work each school week of five days?
(18 3/4)

123. The area of a rectangle is 180 square inches. If its width is one foot, what is the perimeter (distance around) of the rectangle?
(54 inches)

124. The area of a rectangle is 5 1/4 square feet. If its width is six inches, how many feet is the length of the rectangle?
(10 1/2)

125. A party mix needs 3 ounces of Rice Chex, four ounces of corn Chex, and five ounces of peanuts. If you wanted to make two pounds of mix, how many ounces of Rice Chex would you need?
(8)

126. A 6 gallon bucket has a hole that leaks out one quart of water in a minute. If a faucet can pour in one gallon in a minute, how long will it take to fill the bucket? (4 quarts makes 1 gallon)
(8 min.)

127. Using only nickels or quarters or a combination of them, how many ways are there to make change for a dollar?
(5)

128. Terra ate five pancakes in twelve minutes, Sam ate 3 pancakes in eight minutes, and Gail ate 4 pancakes in 10 minutes. Who ate the fastest?
(Sam)

129. If a car is traveling at 40 miles per hour, how far will it travel in 75 minutes?
(50 miles)

130. A freight train had stopped on the tracks and Tony jogged along side of it from the caboose to the engine in five minutes. If Tony jogs at 6 miles per hour, how long is the train?
(1/2 mile or 2,640 ft.)

131. Archie runs four feet per second faster than Bob. It takes Archie 5 seconds to run the 40 yard dash. How long does it take Bob to run 40 yards?
(6 sec.)

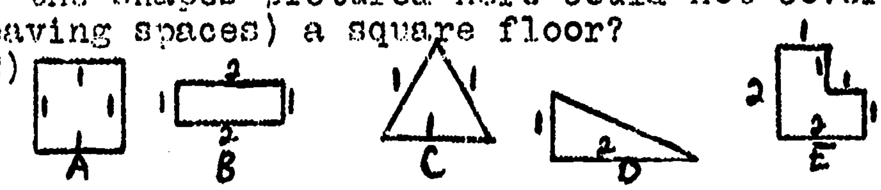
132. The bakery put its fresh batch of cookies into 6 size M boxes with ten cookies left over. The next batch was twice as big, and fit evenly into 13 size M boxes. How many cookies were in each box?
(20)

133. Cindy borrowed $3000 to buy a car. She agreed to pay $100.00 a month for 3 years to repay the loan plus interest. How much interest did she pay in the 3 years?
($600.)

134. Jan put $15.50 in a bank where they pay six cents interest for each dollar you leave in for one year. How much money would she have in the bank after one year?
($16.42)

135. Jan put money in a bank where they pay six cents interest for each dollar you leave in for one year. A year later, her money plus the interest totaled $53.00. How much had she put in the bank?
($50.00)

136. Tiles for floors come in different shapes. Which one of the shapes pictured here could not cover (without leaving spaces) a square floor?
(C)



137. One plane cuts space into two parts and two planes can cut space into at most four parts. What is the largest number of parts that three planes can cut space into?
(8)

138. Mrs. Cord has a 35 foot rope, a 49 foot rope, and a 56 foot rope. He wants to cut all three ropes into smaller pieces so that all the pieces are the same length. He wants these equal pieces to be as long as possible without wasting any rope. How long should each piece be?
(7 ft.)

BEST COPY AVAILABLE

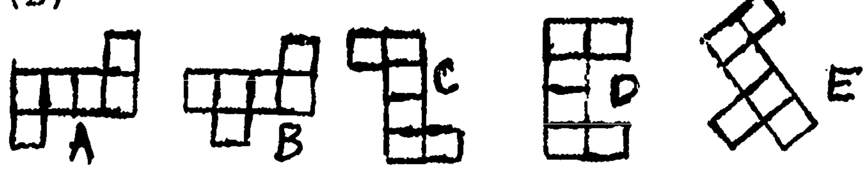
139. Boontown is 50 miles from Clinton and Clinton is 30 miles from Adams. What is the closest possible distance from Adams to Boontown?
(20 miles)

140. What is the smallest number that can be divided by 8, 10, and 12 without leaving any remainder except zero?
(120)

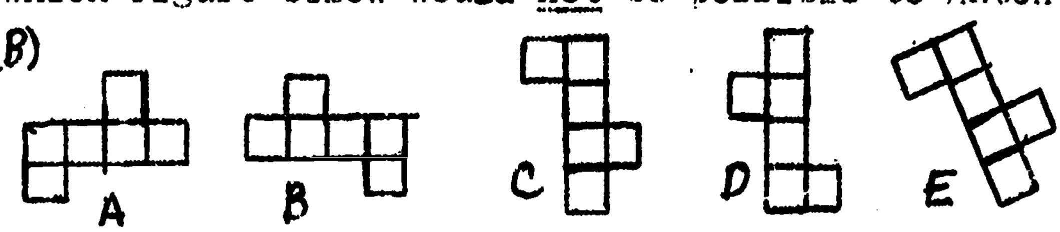
141. A group of boys are standing in the lunch line so that there are two boys in front of a boy, there are two boys behind a boy, and there is a boy between two boys. What is the smallest possible number of boys in the lunch line?
(3)

142. Four students are standing in the lunch line. How many different ways could these four students be lines up?
(24)

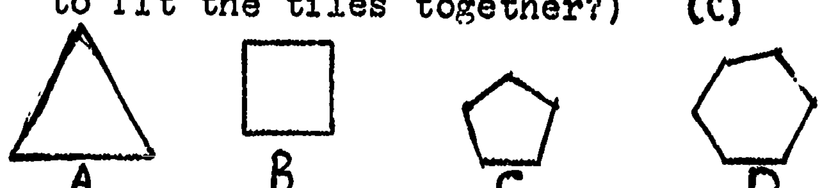
143. Here are shapes made up of six attached squares. ..ch shape could not be folded into the shape of a cube?
(D)



144. Here is a figure made up of six squares. If are allowed to slide and turn, but not flip this figur which figure below would not be possible to match?
(B)



145. Each of the figures below has all sides and angles equal. Which figure could not be used as a tile on a floor (because they would leave spaces of you tried to fit the tiles together?) (C)



146. Mr. Driver fills his gas tank whenever he gets down to one-fourth of a tank. During a trip, he started with a full tank filled up twice along the way, and had half a tank left when he returned home. If his tank holds twenty gallons, how many gallons of gasoline did he use on the trip?
(40)

147. Bret was the season's leading scorer with 63 points, but Jeff was only six points behind. During the next game, Bret scored nine points and Jeff scored 17 points. In the next game after that, each player scored 15 points. How many points has Jeff scored this season? (89)

148. Mrs. King has won three times as many tennis matches as she has lost. If she has played 120 matches, how many has she won? (90)

149. Mr. Riggs has five wins for every two losses in his tennis matches. If he has won 150 matches, how many losses does he have? (60)

150. Willa spent 50 cents on 10¢ pencils and 5¢ erasers. If she bought at least one pencil and one eraser, how many different combinations of pencils and erasers could she buy? (5)

151. Fred has to put 175 marbles into sacks so there is the same number in each sack. If he can't put all the marbles into one sack, what is the smallest number of sacks he will need? **(5)**

152. Janis started her trip with a full tank of gasoline. After driving 1 1/2 hours, she had 2/3 of a tank of gasoline left. How many hours can she drive on a whole tank of gasoline under similar conditions? (4 1/2)

153. John said that he paid about $240.00 for his television set. If he had rounded off to the nearest ten dollars, what is the least he could have paid for his set? ($235.)

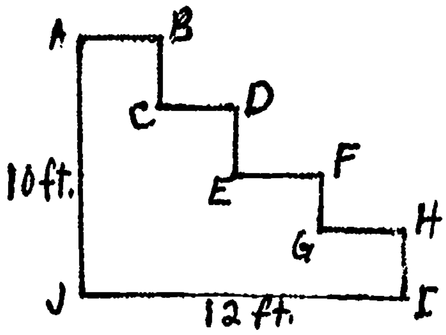
154. One taxi driver gets 35¢ for each dollar clocked on the taxi meter. He also gets tips. If he made a total of $25.00 one day for clocking $60.00 on the meter, how much money in tips did he get? ($4.00)

155. A person working in a restaurant gets paid by the hour plus tips. If the tips average half of the hourly wage and the total of the two is $2.40 an hour, how much an hour does the person get in tips? ($0.80)

156. A three inch rubber band can be stretched to seven inches without breaking it. If a five inch rubber band were made out of the same batch of rubber, how long should it be possible to stretch it without breaking? (11 2/3 in.)

157. John buys pencils at three for ten cents and sells them at a nickel each. How much profit would he earn on a dozen pencils? (20¢)

158. In figure ABCDEFGHIJ, all the horizontal parts of the steps are equal lengths and all the vertical parts of the steps are equal lengths. What is the area of the figure? (72 sq. ft.)



159. If apples cost five pounds for 99¢ and there are about five apples to a pound, approximately how much would twenty apples cost (to the nearest cent)? (80¢)

160. Mr. Roofer charges $200. to reshingle a rectangular roof that is 40 feet by 60 feet. His next job is on a rectangular roof twice as long and twice as wide. How much should he charge for the bigger roof? ($800.)

161. After Mrs. Merchant reduced a $5.00 shirt by a certain fraction of the price, the new price was 4.00. Later she reduced the $4.00 price by the same fraction as before. What is the price of the shirt after the second reduction? ($3.20)

162. A box of candy was passed around the class. Each student in turn took one piece and passed the box on until all 100 pieces were gone. Joe got four pieces including the first piece and the last piece. How many students were in the class? (33)

163. In the last two months, gasoline has increased from thirty five cents a gallon to forty cents a gallon. If it keeps increasing at the same rate, how many months will it be before gasoline will cost one dollar a gallon? (24)

164. By slowing down from sixty miles an hour to fifty miles an hour, Don gets three more miles per gallon of gasoline with his car. He gets eighteen miles per gallon at the slower speed. How many gallons would he save on a 180 mile trip if he traveled at 50 instead of 60 miles per hour?
(2)

165. The rent for an indoor ice rink is $40. per hour. If 25 people skate for 45 minutes and share the cost equally, how much will each have to pay?
($1.20)

Appendix E **BEST COPY AVAILABLE**

INSTRUCTIONS FOR THE INTERVIEW TEST

The purpose of this interview is to obtain some information on the ways in which people like you solve mathematical problems. This is not a test and you don't have to worry about passing it or getting a grade on it. Try to do your best though.

You will be asked to work on a small set of problems and to think aloud as you work on each problem. This means that you should say out loud all the things you are thinking while you try to solve the problem. I will record what you say so that I can remember how you solved the problem and so that I can talk to you.

There are only four rules to follow while you work on the problems.

1. Read each complete problem out loud before you start to work on it. Talk in your usual tone of voice and try to be clear enough for me to understand what you are saying.
2. Write down anything that you want. There is more paper if the problem sheet isn't enough. Don't erase anything: just draw a line through it if you decide not to use it. Keep talking even when you are writing.
3. If you have tried hard to solve a problem and can't get the answer, then just tell me and we can go to the next one.
4. Tell me when you have finished one problem and are ready to start the next one.

Some of your friends might be helping me do this study, so please do not talk about the problems or the interview with them. It may only cause them to get confused and mix up the results of this study. Thank you for helping me.

Name____________________________________ Date________________

Time________________

Appendix F

SUMMARY OF LUCAS' SCORING SYSTEM

Maximum of 5 points broken into 3 subscores:

1) Approach score

A maximum of one point was awarded if it was clear that the subject understood the data, conditions, and objective of the problem. This was indicated by the nullification or correction of all structural errors. No points were awarded if confusion on any of the three parts of the problem prevented the subject from establishing a direction which could lead to a correct solution.

2) Plan Score

A maximum of two points was awarded when the subject had derived enough relationships to solve the problem or had produced a sequence of approximations which had focused on the correct solution. Structural errors had to be corrected or nullified. Executive errors were permitted if they did not obscure the solution path.

One point was awarded if the rationale for a key step in the solution was lacking or an important relationship or step prevented the subject from achieving a completed solution. An uncorrected structural error would also be a source of an incomplete or unclear solution path. No points would be awarded for haphazard, unclear, or undirected procedures or plans.

3) Result Score

A maximum of two points was awarded when the subject established a correct form of the solution. All structural or executive errors had to be corrected or nullified to score two points.

One point was awarded for a correct numerical result but with incorrect units, or if the result was a close approximation of the solution, or if the subject failed to provide all the required unknowns.

4) Total Score

The total score for a single problem was the sum of the approach, plan, and result scores. Thus, an integral score ranging from 0 to 5 inclusive was possible.

Appendix G

PROCESS-SEQUENCE CODES

<u>Process Symbols</u>

R	=	reads the problem
Rr	=	rereads the problem or parts of it
Rs	=	restates the problem in his own words
S	=	separates or summarizes data
M_f	=	introduces model by means of a diagram
$M_{f'}$	=	modifies existing diagram
Me	=	model introduced by means of equation, expression or other relationship
Alg	=	algorithmic process
DX	=	exploratory work with data (direction not apparent)
DS	=	deduction by synthesis (direction apparent)
DA	=	deduction by analysis
TR	=	random trial and error (no pattern apparent)
TS	=	systematic trial and error (pattern apparent)
An	=	reasoning by analogy
N	=	not classifiable
C	=	checks the result

<u>Outcomes of DX, DS, DA, TR, TS, N Processes</u>

1	=	abandons process
2	=	impasse
3	=	incorrect final result
4	=	correct final result
5	=	intermediate result

Appendix G

Punctuation Marks

-	(dash) hestitation of approximately 15 seconds
( )	scope of DX, DS, DA, TR, TS or N process
,	inserted between successive processes
.	stops with solution (correct or incorrect)
/	stops without solution

Errors

se	above process symbols = structural error in process
ee	above process symbols = executive error in process
sec	above process symbols = structural error corrected
eec	above process symbols = executive error corrected

CODING FORM

Subject No. ______

Problem No. ______

Coder ______

Date ______

Tape No. ______

Tape Readings ______

Time: exc. looking back ______

looking back ______

Total ______

Approach

- rereads entire problem ______
- rereads data only ______
- rereads question only ______
- restates problem in own words ______
- mnemonic notation ______
- representative diagram yes ______
- no ______
  - auxiliary lines ______
  - enlarges focal points ______

Production

- recalls related formula ______
- recalls related problem ______
- inductive reasoning (pattern search) ______

Looking Back

- routine check of manipulation ______
- is result reasonable? ______
- does result satisfy data? ______

Executive Errors

- algebraic manipulations ______
- arithmetic computation ______
- incorrect use of formula or algorithm ______
- other ______

Structural Errors

- misinterprets data ______
- misinterprets question ______
- other ______

Subject Reactions

- asks question about problem ______
- comments about problem ______
- comments on ability to do problem ______

Interviewer Comments

PROCESS SEQUENCE

Appendix H

PILOT STUDY WT RESULTS

8 Subjects 16 items

Subject	Problem 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
1	I	I	I		I		I	I	I	I		I		I	I	I	12
2		I			I		I	I	I	I				I	I		8
3	I				I		I	I				I			I		6
4				I	I	I		I	I	I		I				I	8
5		I			I		I	I	I	I				I	I		8
6		I	I	I			I	I	I	I					I	I	9
7		I			I	I	I	I	I	I		I		I	I		10
8	I	I		I			I		I	I		I			I	I	9

A "I" indicates that the subject got the correct solution.
A blank indicates that the subject got an incorrect answer.

Kuder-Richardson Formula 20:

$$KR = \frac{n}{n - 1}\left[1 - \frac{\Sigma p_i \, q_i}{S_x^2}\right]$$

p_i = proportion of subjects who got item i correct

q_i = proportion of subjects who got item i wrong

S_x^2 = variance of the total scores of the subjects

Appendix I

SOLUTION AND CODING TIMES OF SUBJECTS' PROTOCOLS

Subject Number	Solution Time (Minutes)	Coding Time (Minutes)
1	22	44
2	10	20
3	*	*
4	11	58
5	14	48
6	12	34
7	16	27
8	18	25
9	28	56
10	18	46
11	25	87
12	22	56
13	20	28
14	8	28
15	18	50
16	9	28
17	10	32
18	11	38
19	*	*
20	14	82
21	10	33
22	11	36
23	17	45
24	8	24
25	14	36
26	13	54
27	17	58
28	17	55
29	8	18
30	12	36
31	20	50

*Due to technical problems, the time was not recorded.

Appendix J

AGREEMENT ON CODING AND SCORING VARIABLES

BEST COPY AVAILABLE

Variable	Index of Reliability	Coders 1 and 2 Frequency of Agreement	Frequency of Disagreement	No. of Positive Observations	Agreement Ratio
Rr	.89	24	9	29	.73
S	.40	12	5	7	.71
DS	.94	25	9	29	.74
DX	.68	15	2	5	.88
DA	.81	19	10	19	.66
TS	.97	17	2	7	.89
TR	.59	14	2	4	.88
Me	.97	32	20	45	.62
ee	.98	20	4	16	.83
se	.71	13	4	9	.76
M_f(d)	1.00	16	0	3	1.00
Alg	.86	42	21	61	.67
C	.86	16	6	12	.73
X_6(d)	1.00	16	0	5	1.00
X_{16}	.99	19	1	6	.95
X_{17}	.87	15	1	6	.94
X_{20}	.30	13	3	4	.81
X_{21}	.68	15	1	2	.94
X_{26}(d)	.88	13	3	16	.81
X_{27}	.58	11	5	16	.69
X_{28}	.91	12	4	16	.75

d = dichotomous variable

Appendix J (Continued)

AGREEMENT ON CODING AND SCORING VARIABLES

Coders 2 and 3

Variable	Index of Reliability	Frequency of Agreement	Frequency of Disagreement	No. of Positive Observations	Agreement Ratio
Rr	.91	22	7	23	$\frac{22}{29}$ = .76
S	.29	15	1	3	$\frac{15}{29}$ = .94
DS	.91	23	9	26	$\frac{23}{32}$ = .72
DX	.33	14	2	4	$\frac{14}{16}$ = .88
DA	.89	19	4	11	$\frac{19}{23}$ = .83
TS	.82	16	0	4	$\frac{16}{16}$ =1.00
TR	.59	14	2	3	$\frac{14}{16}$ = .88
Me	.79	31	21	45	$\frac{31}{52}$ = .60
ee	.95	18	4	15	$\frac{18}{22}$ = .82
se	.36	11	6	9	$\frac{11}{17}$ = .65
M_f (d)	1.00	16	0	2	$\frac{16}{16}$ =1.00
Alg	.84	42	17	57	$\frac{42}{59}$ = .71
C	.76	17	4	11	$\frac{17}{21}$ = .81
X_6 (d)	1.00	16	0	5	$\frac{16}{16}$ =1.00
X_{16}	.75	16	0	3	$\frac{19}{22}$ = .86
X_{17}	.87	16	0	3	$\frac{16}{16}$ =1.00
X_{20}	.30	14	2	4	$\frac{14}{16}$ = .88
X_{21}	.68	15	1	2	$\frac{15}{16}$ = .94
X_{26} (d)	1.00	15	1	16	$\frac{15}{16}$ = .94
X_{27}	.72	11	5	16	$\frac{11}{16}$ = .69
X_{28}	.89	12	4	16	$\frac{12}{16}$ = .75

d = dichotomous variable

Appendix J (Continued)

AGREEMENT ON CODING AND SCORING VARIABLES

Coders 1 and 3

BEST COPY AVAILABLE

Variable	Index of Reliability	Frequency of Agreement	Frequency of Disagreement	No. of Posivite Observations	Agreement Ratio
Rr	.93	24	5	23	$\frac{24}{29}$ = .83
S	.74	14	3	6	$\frac{14}{17}$ = .82
DS	.97	31	6	29	$\frac{31}{37}$ = .84
DX	.68	15	1	4	$\frac{15}{16}$ = .94
DA	.84	19	9	17	$\frac{19}{28}$ = .69
TS	.97	17	2	7	$\frac{17}{19}$ = .89
TR	1.00	16	0	3	$\frac{16}{16}$ =1.00
Me	.89	38	24	53	$\frac{38}{62}$ = .61
ee	.94	17	7	18	$\frac{17}{28}$ = .71
se	.66	15	1	4	$\frac{15}{16}$ = .94
M_f(d)	1.00	18	0	5	$\frac{18}{18}$ =1.00
Alg	.95	48	16	62	$\frac{48}{64}$ = .75
C	.81	16	7	13	$\frac{16}{23}$ = .70
X_6(d)	1.00	16	0	5	$\frac{16}{16}$ =1.00
X_{16}	.76	18	3	8	$\frac{18}{21}$ = .86
X_{17}	.71	16	0	4	$\frac{16}{16}$ =1.00
X_{20}	.43	15	1	3	$\frac{15}{16}$ = .94
X_{21}	1.00	16	0	1	$\frac{16}{16}$ =1.00
X_{26}(d)	1.00	14	2	16	$\frac{14}{16}$ = .88
X_{27}	.70	12	4	16	$\frac{12}{16}$ = .75
X_{28}	.96	15	1	16	$\frac{15}{16}$ = .94

d = dichotomous variable

Appendix K

SUBJECT SCORES ON THE INTERVIEW TEST

Subject	P1	P2	P3	P4	P5	P6	Sub-totals	Total
	0	1	1	1	1	1	5	
1	0	1	0	2	0	2	5	14
	0	0	0	2	0	2	4	
	0	0	1	1	0	0	2	
2	0	0	1	2	0	0	3	9
	0	0	2	2	0	0	4	
	1	0	0	1	0	0	2	
3	1	0	0	2	0	0	3	8
	2	0	0	1	0	0	3	
	0	1	1	1	1	1	5	
4	0	2	2	2	0	1	7	18
	0	2	2	1	0	1	6	
	0	0	1	1	1	0	3	
5	0	0	1	2	1	0	4	8
	0	0	0	1	0	0	1	
	0	0	0	1	0	0	1	
6	0	0	0	1	0	0	1	3
	0	0	0	1	0	0	1	
	0	0	0	1	1	0	2	
7	0	0	0	1	1	0	2	5
	0	0	0	1	0	0	1	
	0	1	1	0	0	0	2	
8	0	2	1	0	0	0	3	9
	0	2	2	0	0	0	4	
	1	1	1	1	1	1	6	
9	1	2	2	2	0	0	7	19
	0	2	2	2	0	0	6	
	0	0	1	1	0	0	2	
10	0	0	1	1	0	0	2	7
	0	0	2	1	0	0	3	
	0	1	1	1	1	0	4	
11	0	1	0	1	1	0	3	8
	0	0	0	1	0	0	1	
	0	1	1	1	1	1	5	
12	0	1	1	1	0	0	3	9
	0	0	1	0	0	0	1	

Appendix K (Continued)

SUBJECT SCORES ON THE INTERVIEW TEST

Subject	P1	P2	P3	P4	P5	P6	Sub-totals	Total
	1	1	1	0	1	0	4	
25	1	2	2	0	1	0	6	17
	2	2	2	0	1	0	7	
	1	1	1	1	1	0	5	
26	1	2	2	2	1	0	8	20
	1	2	2	1	1	0	7	
	0	1	1	1	1	1	5	
27	0	2	2	1	0	2	7	18
	0	2	2	0	0	2	6	
	0	1	0	1	0	0	2	
28	0	0	0	1	0	0	1	3
	0	0	0	0	0	0	0	
	0	0	0	0	0	0	0	
29	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	
	0	1	0	1	0	0	2	
30	0	2	0	2	0	0	4	8
	0	1	0	1	0	0	2	
	1	1	1	1	0	0	4	
31	1	0	1	1	0	0	3	9
	0	0	2	0	0	0	2	

Appendix K (Continued)

SUBJECT SCORES ON THE INTERVIEW TEST

Subject	P1	P2	P3	P4	P5	P6	Sub-totals	Total
	0	0	1	1	0	0	2	
13	0	0	1	1	0	0	2	6
	0	0	2	0	0	0	2	
	0	1	0	1	0	0	2	
14	0	0	0	1	0	0	1	4
	0	0	0	1	0	0	1	
	1	1	1	1	1	1	6	
15	2	2	2	2	0	2	10	24
	2	2	2	1	0	1	8	
	0	0	0	1	0	0	1	
16	0	0	0	2	0	0	2	4
	0	0	0	1	0	0	1	
	0	1	1	1	0	0	3	
17	0	0	2	2	0	0	4	10
	0	0	2	1	0	0	3	
	0	0	1	1	0	0	2	
18	0	0	0	2	0	0	2	5
	0	0	0	1	0	0	1	
	0	1	1	1	0	1	4	
19	0	2	2	1	0	2	7	18
	0	2	2	1	0	2	7	
	0	1	1	1	0	0	3	
20	0	0	2	1	0	0	3	8
	0	0	2	0	0	0	2	
	0	1	1	1	0	0	3	
21	0	2	2	2	0	0	6	13
	0	1	2	1	0	0	4	
	0	1	0	1	1	0	3	
22	0	2	0	2	0	0	4	10
	0	2	0	1	0	0	3	
	0	1	1	1	0	0	3	
23	0	0	2	1	0	0	3	9
	0	0	2	1	0	0	3	
	0	0	0	0	0	0	0	
24	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	

BEST COPY AVAILABLE

Appendix L

SIMILARITY MEASURES*

Subject Number

	15	26	9	27	19	4	25	1	21	12	17,22**	31	23	2,8**	11	20
15	0	127	227	267	331	267	451	765	929	1787	1434	1662	1656	1769	1903	1860
26	127	0	76	36	57	36	108	320	398	1094	731	930	884	950	1134	1049
9	227	76	0	40	178	40	195	182	448	778	677	731	797	982	896	929
27	267	36	40	0	59	0	76	142	250	737	479	612	599	704	777	731
19	331	57	178	59	0	59	17	275	224	981	492	737	612	599	941	777
4	267	36	40	0	59	0	76	142	250	737	479	612	599	704	777	731
25	451	108	195	76	17	76	0	224	207	861	407	617	492	479	821	657
1	765	320	182	142	275	142	224	0	176	235	195	182	246	425	275	301
21	929	398	448	250	224	250	207	176	0	479	87	267	172	193	360	227
12	1787	1094	778	737	981	737	861	235	479	0	250	59	233	524	40	178
17,22	1434	731	677	479	492	479	407	195	87	250	0	76	17	76	131	36
31	1662	930	731	612	737	612	617	182	267	59	76	0	59	233	19	40
23	1656	884	797	599	612	599	492	246	172	233	17	59	0	59	114	19
2,8	1769	950	982	704	599	704	479	425	193	524	76	233	59	0	326	114
11	1903	1134	896	777	941	777	821	275	360	40	131	19	114	326	0	59
20	1860	1049	929	731	777	731	657	301	227	178	36	40	19	114	59	0

* All measures have been multiplied by 100.

** Both subjects had identical subscores.

Appendix L (continued)

SIMILARITY MEASURES*

	Subject Number											
	5	3	30	10	13	7,18**	14	16	28	6	24	29
15	1839	1934	1916	2189	2393	2634	2924	2991	3202	3281	4320	4320
26	1100	1082	1094	1269	1434	1638	1860	1916	2101	2138	2991	2991
9	974	1075	1087	1228	1360	1525	1712	1882	1916	2069	2932	2932
27	776	797	809	797	1082	1247	1434	1525	1638	1712	2495	2495
19	861	731	776	884	1049	1253	1440	1451	1681	1638	2379	2379
4	776	797	809	797	1082	1247	1434	1525	1638	1712	2495	2495
25	776	611	691	731	896	1100	1253	1298	1498	1448	2157	2157
1	343	444	448	529	584	677	797	955	929	1075	1717	1717
21	235	212	182	332	387	480	633	599	765	752	1269	1269
12	176	431	393	448	393	374	425	652	444	703	1164	1164
17,22	74	57	59	108	127	182	267	301	360	386	797	797
31	76	59	176	195	176	195	246	393	301	444	862	862
23	91	40	76	57	76	131	182	250	275	301	677	677
2,8	224	19	91	36	91	184	235	224	367	275	611	611
11	57	233	195	250	19	176	227	374	246	425	921	921
20	36	59	57	76	57	76	127	195	182	246	584	584

* All measures have been multiplied by 100.

** Both subjects had identical subscores.

Appendix L (continued)

SIMILARITY MEASURES*

	Subject Number															
	15	26	9	27	19	4	25	1	21	12	**17,22	31	23	**2,8	11	20
5	1839	1100	974	776	861	776	776	343	235	176	74	76	91	224	57	36
3	1934	1082	1075	797	731	797	611	444	212	431	57	59	40	19	233	59
30	1916	1094	1087	809	776	809	691	448	182	393	59	176	76	91	195	57
10	2189	1269	1228	797	884	797	731	529	332	448	108	195	57	36	250	76
13	2393	1434	1360	1082	1049	1082	896	584	387	393	127	176	76	91	19	57
7,18	2634	1638	1525	1247	1253	1247	1100	677	480	374	182	195	131	184	176	76
14	2924	1860	1712	1434	1440	1434	1253	797	633	425	267	246	182	235	227	127
16	2991	1916	1882	1525	1451	1525	1298	955	599	652	301	393	250	244	374	195
28	3202	2101	1916	1638	1681	1638	1498	929	765	444	360	301	275	367	246	182
6	3281	2138	2069	1712	1638	1712	1448	1075	752	703	386	444	301	275	425	246
24	4320	2991	2932	2495	2379	2495	2157	1717	1269	1164	797	862	677	611	921	584
29	4320	2991	2932	2495	2379	2495	2157	1717	1269	1164	797	862	677	611	921	584

* All measures have been multiplied by 100.

** Both subjects had identical subscores.

Appendix L (continued)

SIMILARITY MEASURES*

	Subject Number											
	5	3	30	10	13	7,18**	14	16	28	6	24	29
5	0	131	59	182	127	108	193	227	212	312	649	649
3	131	0	36	17	36	91	142	131	235	182	479	479
30	59	36	0	87	68	87	172	127	227	127	506	506
10	182	17	87	0	19	74	91	114	184	131	394	731***
13	127	36	68	19	0	19	36	59	91	76	301	301
7,18	108	91	87	74	19	0	17	40	36	57	246	246
14	193	142	172	91	36	17	0	57	19	40	195	195
16	227	131	127	114	59	40	57	0	76	17	127	127
28	212	235	227	184	91	36	19	76	0	59	176	176
6	312	182	127	131	76	57	40	17	59	0	76	76
24	649	479	506	394	301	246	195	127	176	76	0	0
29	649	479	506	394	301	246	195	127	176	76	0	0

* All measures have been multiplied by 100.

** Both subjects had identical subscores.

*** Error: Subjects 24 and 29 should have identical values. Both should be 394.

Appendix M

MULTIDIMENSIONAL SCALING RESULTS

FOR 2, 3, AND 4 DIMENSIONS **BEST COPY AVAILABLE**

Kruskal-Guttman-Lingoes-Roskam Smallest Space Coordinates for M = 2 and M = 3 (Weak monotonicity)

Variable (Subject)	Dimension* 1	2	Dimension** 1	2	3
1	-21.393	-71.282	-20.912	-57.460	-86.621
2	27.379	-74.084	25.113	-97.414	-94.084
3	33.240	-65.017	33.171	-86.088	-90.849
4	-50.811	-60.816	-50.695	-77.364	87.606
5	28.655	-42.829	30.577	-60.625	-67.418
6	72.142	-61.083	73.213	-85.313	-84.304
7	53.957	-55.576	56.596	-72.477	-84.451
8	27.390	-74.121	25.113	-97.417	-94.077
9	-61.852	-72.602	-61.299	-61.977	-90.878
10	39.599	-68.416	38.289	-86.138	-100.000
11	24.459	-33.732	27.143	-43.454	-80.710
12	13.176	-100.000	16.945	-29.435	-84.099
13	47.337	-59.902	48.207	-79.245	-91.553
14	60.832	-54.958	63.836	-70.576	-88.902
15	-100.000	-63.828	-100.000	-82.651	-83.109
16	66.193	-64.123	66.751	-83.409	-80.840
17	14.534	-56.334	15.307	-76.831	-81.454
18	53.980	-55.710	56.596	-72.479	-84.441
19	-49.825	-43.806	-47.854	-98.040	-76.593
20	28.022	-54.670	30.833	-66.993	-82.649
21	- 8.782	-50.000	-6.437	-86.377	-67.633
22	14.534	-56.301	15.308	-76.833	-81.448
23	21.690	-59.488	22.400	-74.375	-90.991
24	98.835	-63.762	99.074	-94.341	-83.519
25	-42.629	-45.080	-40.152	-100.000	-82.019
26	-66.565	-54.666	-66.238	-85.147	-79.799
27	-50.810	-61.023	-50.698	-77.345	-87.523
28	66.934	-48.035	71.371	-63.088	-83.121
29	100.000	-61.070	100.000	-91.891	-74.758
30	34.721	-51.848	35.035	-85.087	-71.493
31	15.702	-68.303	20.641	-53.233	-90.783

* Kruskal's stress = .07619 in 7 iterations.
Guttman-Lingoes' coefficient of alienation = .00000

** Kruskal's stress = .01274 in 89 iterations.
Guttman-Lingoes coefficient of alienation = .00000

Appendix M (continued)

MULTIDIMENSIONAL SCALING RESULTS

FOR 2, 3, AND 4 DIMENSIONS

Kruskal-Guttman-Lingoes-Roskam Smallest Space Coordinates for M = 4 (Weak monotonicity)

Variable	Dimension			
(Subject)	1	2	3	4
1	-20.687	-56.701	-80.473	-95.642
2	26.156	-99.989	-89.823	-93.869
3	34.056	-87.886	-91.438	-89.940
4	-51.374	-77.728	-81.850	-93.626
5	31.523	-60.236	-76.348	-71.010
6	75.098	-86.680	-79.439	-88.800
7	58.531	-73.115	-85.723	-86.009
8	26.155	-100.000	-89.926	-93.744
9	-61.537	-62.700	-73.722	-100.000
10	39.790	-88.931	-100.000	-91.950
11	27.890	-42.843	-88.532	-81.811
12	17.548	-28.036	-84.703	-91.738
13	49.877	-81.734	-87.708	-94.021
14	64.843	-71.061	-85.658	-96.004
15	-100.000	-85.255	-64.448	-93.155
16	69.148	-88.944	-79.444	-82.251
17	15.947	-77.590	-79.953	-85.395
18	58.530	-73.119	-85.697	-86.005
19	-49.872	-96.979	-89.753	-81.769
20	31.481	-67.382	-82.112	-86.358
21	-7.008	-88.841	-73.829	-72.443
22	15.944	-77.590	-80.053	-85.326
23	23.002	-75.776	-86.584	-96.073
24	100.000	-93.268	-70.650	-99.151
25	-41.900	-97.215	-97.775	-87.731
26	-68.483	-83.395	-83.550	-85.235
27	-51.415	-77.580	-82.100	-93.410
28	73.804	-63.524	-84.693	-88.976
29	99.610	-89.709	-61.741	-96.777
30	35.933	-85.893	-78.342	-73.376
31	21.162	-53.518	-90.792	-93.215

Kruskal's stress = .01151 in 40 iterations.
Guttman-Lingoes' coefficient of alienation = .00000

Appendix N

GAMMA VALUES FOR CLUSTERING

Iteration	Gamma
1	.79212
2	1.00000
3	1.00000
4	1.00000
5	1.00000
6	1.00000
7	1.00000
8	1.00000
9	1.00000
10	1.00000
11	1.00000
12	.99664
13	.99400
14	.99353
15	.99446
16	.98954
17	.98429
18	.97703
19	.96978
20	.95885
21	.96379
22	.95678
23	.94025
24	.93508
25	.90695
26	.86401
27	.84028
28	.81560
29	.87275

☆ U.S. GOVERNMENT PRINTING OFFICE: 1974— 651-298/35

National Evaluation Committee

Helen Bain
Past President
National Education Association

Lyle E. Bourne, Jr.
Institute for the Study of Intellectual Behavior
University of Colorado

Sue Buel
Dissemination and Installation Services
Northwest Regional Educational Laboratory

Francis S. Chase
Professor Emeritus
University of Chicago

George E. Dickson
College of Education
University of Toledo

Chester W. Harris
Graduate School of Education
University of California

Hugh J. Scott
Consultant
National Evaluation Committee

H. Craig Sipe
Department of Instruction
State University of New York

G. Wesley Sowards
Dean of Education
Florida International University

Joanna Williams
Professor of Psychology and Education
Columbia University

Executive Committee

William R. Bush
Director, Program Planning and Management
Deputy Director, R & D Center

M. Vere DeVault
Professor
School of Education

Herbert J. Klausmeier
Principal Investogator
R & D Center

Joel R. Levin
Principal Investigator
R & D Center

Donald N. McIsaac
Associate Dean, School of Education
University of Wisconsin

Richard A. Rossmiller, Committee Chairman
Director
R & D Center

Len VanEss
Associate Vice Chancellor
University of Wisconsin–Madison

Dan Woolpert
Director, Management Systems
R & D Center

Faculty of Principal Investigators

Vernon L. Allen
Professor
Psychology

B. Dean Bowles
Associate Professor
Educational Administration

Frank H. Farley
Associate Professor
Educational Psychology

Marvin J. Fruth
Associate Professor
Educational Administration

John G. Harvey
Associate Professor
Mathematics

Frank H. Hooper
Associate Professor
Child Development

Herbert J. Klausmeier
V. A. C. Henmon Professor
Educational Psychology

Gisela Labouvie
Assistant Professor
Educational Psychology

Joel R. Levin
Associate Professor
Educational Psychology

L. Joseph Lins
Professor
Institutional Studies

James Lipham
Professor
Educational Administration

Wayne Otto
Professor
Curriculum and Instruction

Robert Petzold
Professor
Curriculum and Instruction

Thomas A. Romberg
Associate Professor
Curriculum and Instruction

Dennis W. Spuck
Assistant Professor
Educational Administration

Richard L. Venezky
Associate Professor
Computer Science

Larry M. Wilder
Assistant Professor
Communication Arts